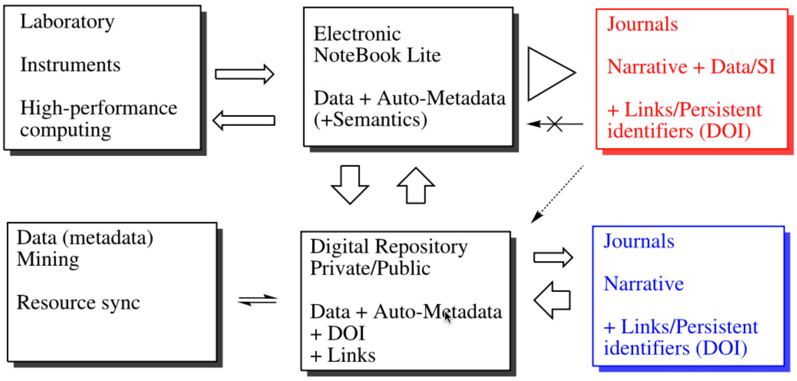
Research data management in Computational and Experimental Molecular science.

Henry S. Rzepa (PI), Matt J. Harvey and Nick J. Mason

The idea: Data for small and medium sized discrete molecules deriving from research in molecular sciences is increasingly nowadays a composite of semantically related instrumental data measurements and computationally generated models derived from high performance computing (HPC) resources. The experiments are increasingly managed *via* electronic laboratory notebook environments or data portals, prior to being published along with appropriate narratives in scientific journals. Our management concept in **phase 1** has been to create a Janus-like user portal (Uportal, Figure 1) as the hub interfacing to the HPC resources to enable automated metadata capture using community standards and open source tools where possible. The other Uportal face is to digital data repositories. These represent the public face of the data and have been optimized to interface with publisher's journals *via* the concept of the separation of narrative and data (*data emancipation*). Phase 2, this proposal, is to address the issues of sustainability and scalability of RDM using a variety of new standards introduced for this purpose.



Laboratory
Instruments
High-performance computing
Electronic NoteBook Lite
Data + Auto-Metadata (+Semantics)
Journals
Narrative + Data/SI
+ Links/Persistent identifiers (DOI)
Data (metadata) Mining
Resource sync
Digital Repository Private/Public
Data + Auto-Metadata + DOI + Links
Journals
Narrative
+ Links/Persistent identifiers (DOI)

Issues surrounding RDM. The molecular science community as an example generates far more data than can be conventionally published in journals. Only a proportion of this data (estimated at most as <20%) is captured and curated in the form of supporting information, *via* "*unfit-for-purpose*" paginated PDF files in which the data loses structure, semantics and metadata. Lost data in turn cannot be mined using powerful techniques known as TDM (text and data mining). In 2005, we led the world in providing a solution for lost data by deposition/publication into a DSpace-based digital repository SPECTRa, a UK JISC funded project.[1] Injection was *via* a locally developed **uportal** interface. Since then we have added the capability of capturing instrumental data (IR, MS, NMR, crystallography) and two further digital repositories to the system to robustify the handling of semantics (Chempound) and take advantage of externally developed support for some standards (Figshare, implementing *e.g.* **ORCID** metadata harvesting). This work has resulted in five prototype publications serving as state-of-the-art exemplars for **Phase 1** of our long-term strategy.[2] Here we propose **Phase 2** to address challenges such as how to introduce redundancy and robustness into the system, to ensure scalability for data sizes and sustainability of implementations.

Benefits to the academic Community. Our strategy is to dramatically improve the current situation of < 20% data capture and subsequent curation and archival, along with poor installed metadata and semantic structures. Our solution will directly improve this for our core discrete molecule community by automating capture of metadata and providing templates (datuments) for the interface between the digital repository and the scientific journal. The intended technologies used here (HTML5, CSS, SVG, Javascript) will also benefit the teaching and learning community by providing a data-rich framework for developing materials for mobile-computing (tablets, phones). We will promote the wider use of this technology by identifying and scoping the requirements of other high-performance-computing (HPC) applications from domains such *e.g.* bioinformatics, physics, business.

Specific Deliverable 1: UPortal Sustainability (1 month). The HPC-UPortal was initially developed in 2006 and has subsequently evolved on an *ad hoc* basis. Although it has proven durable and popular, many of the encapsulated design and implementation decisions make it increasingly unwieldy to maintain and extend with new functionality. We propose that the current UPortal will undergo maintenance and development to enhance its capabilities and sustainability. This work will result in a **public release** that can be easily deployed on third-party systems. It will involve

* Code clean-up and refactoring, to improve maintainability and fix outstanding bugs.
* Generalisation of interface to HPC system and web presentation.
* Generalisation of the interface to DSpace, Figshare and Chempound repositories.
* Movement of the codebase to a public repository (eg Github) for third-party use.
* Writing deployment and operational documentation.

Specific Deliverable 2: UPortal Development and Promotion (3 months). UPortal is currently focused on Computational Chemistry applications. Where possible, UPortal presentations of exemplars will be scoped, produced and promoted to selected other HPC-user constituencies. Where the UPortal workflow is found to be too restrictive to accommodate identified applications, a gap analysis will be performed to provide the basis for future development. To take advantage of the establishment of the Imperial DSpace repository Spiral, we will implement direct deposition to it from UPortal. By co-publishing directly to the college DSpace we inherit the high quality of service assured of a production ICT-managed system, access to true DOIs for depositions and higher impact for publications by improved web visibility (collaboration with Andrew Maclean/ICT). This work will require initial scoping-development of a SWORD/ResourceSync-compliant deposition tool to broaden the appeal of UPortal for third-party deployments (*e.g.* Edinburgh RSpace).[3]

Specific Deliverables 3: Promoting Data Emancipation (2 months). Following on from our exploratory work[2] using under-exploited **10320/loc** features of the Handle system to improve the machine-accessibility of structured data in repository depositions, we aim to promote awareness of these methods in the wider community by outreach to other existing and prospective repository developers (Figshare, RSpace at Edinburgh, RSC Chemspider, DataCite)[3] This will facilitate collaborations outlasting this initial project, and which will ultimately act to promote community consensus on best practices. To support this activity, we will perform software development on our proof-of-concept Handle server extensions to enable other Handle server operators (Rspace) to more easily experiment with and deploy such features.

How the funding requested will be used to achieve these deliverables.

1. **Specific Deliverable 1.** External consultant/Professional Software developer, 1 month. £5000
2. **Specific Deliverables 2+3**. Nick Mason, PG student developer, 5 months@£2982=£14910.
3. Travel costs for discussions with DataCite/Edinburgh. £1000
4. **Total: £20910.**

[1] J. Downing, P. Murray-Rust, A. P. Tonge, P. Morgan, H. S. Rzepa, F. Cotterill, N. Day and M. J. Harvey, "SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories", *J. Chem. Inf. Mod.*, **2008**, *48*, 1571 - 1581. DOI: 10.1021/ci7004737

[2] M. J. Harvey, N. J. Mason and H. S. Rzepa "Digital data repositories in chemistry and their integration with journals and electronic laboratory notebooks", submitted for publication.

[3] Emails of support from RSpace and Figshare have been received, and HSR is a member of an RSC working party on development of standards for research data management.