Research data management in Computational and Experimental Molecular Sciences

A Research Data Management (RDM) "Green Shoots" Pilots Project Report by Henry S. Rzepa, Matt Harvey, Andrew Mclean and Nick Mason Imperial College London

This project was funded as part of Imperial College's RDM "Green Shoots" Programme. In 2014, the Vice-Provost, Research, approved an allocation of £100K for academically-driven projects to identify and generate exemplars of best practice in RDM, specifically frameworks and prototypes that would comply with key funder RDM policies and the College position. The call for projects outlined that frameworks could be based either on original ideas or integrating existing solutions into the research process, improving its efficacy or the breadth of its usage. There was an expectation that solutions would support open access for data; solutions that supported Open Innovation were strongly encouraged.

Six projects were funded, covering different disciplines, faculties and research areas. The projects ran for six months, finishing at the end of 2014. Project reports were made available in 2015.

For more information on the programme and projects please visit:

http://www3.imperial.ac.uk/researchsupport/rdm/policy/greenshoots

Imperial College London

Research data management in Computational and Experimental Molecular Sciences

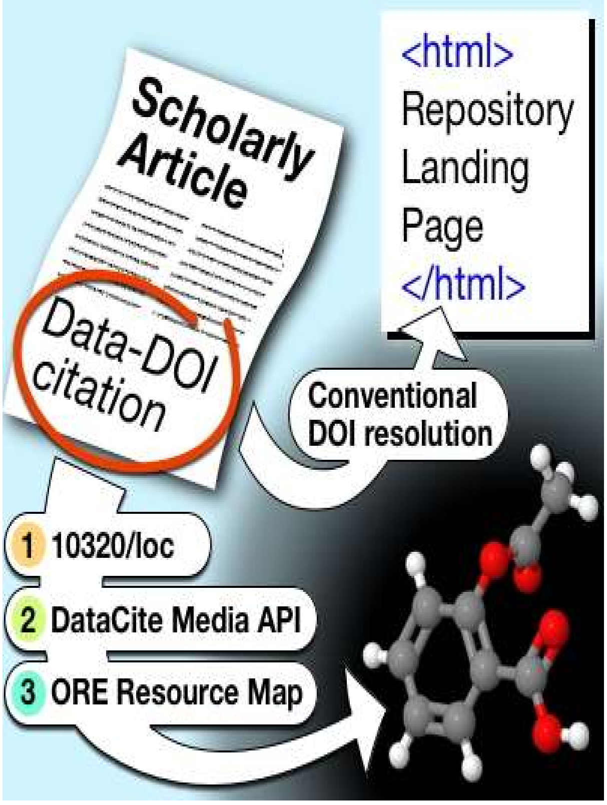
Henry S. Rzepa (ORCID: 0000-0002-8635-8390), Matt Harvey (ORCID: 0000-0003-1797-3186), Andrew Mclean, Nick Mason (ORCID: 0000-0001-9475-0328); Imperial College London

Context of the project

This project is based around an existing data repository (SPECTRa), its font end (uportal) and the enhancement of the system by enabling standards-based access to the data held on the repository, thereby allowing creation of visual representations of the data for use in tables and figures published in scholarly journals and incorporation into conference presentations and teaching material. An important element of this project has been the development and improvement of an overall workflow which provides a practical (and proven) method of data gathering, both new and already existing, together with the addition of the essential data and DOIs that actually facilitate subsequent discovery and re-use.

Issues

1. The first issue addressed was to package up a workflow based on the software systems constituting the uportal front end into a package that could be openly distributed for installation and use by other research groups, departments or institutions.
2. The data repository itself is based on DSpace, and we wished to enhance its functionality by adding a retrospective DataCite persistent identifier to each item in the repository, and to investigate the appropriate metadata for optimising the interaction between SPECTRa and various DataCite services.
3. In preliminary work prior to this project, we had implemented an extension to the DSpace Handle records to allow machine access to the individual data files based on specifying their persistent identifier together with a specification of the type of datafile required. This system was not directly compatible with the DataCite mechanisms providing this functionality, an issue that we addressed in the present project.
4. We wished to address the issue of data discovery and datametrics, based on metadata records and the facilities provided by DataCite.



5. The issue of content retrieval by authors wishing to make use of these features in journal articles.
6. The issue of facilitating data curation of existing data collections which did not conform to modern repository standards.
7. The issue of disseminating best repository and RDM practice and the experiences we had gained as a result of our own project.

Implementations and deliverables

These are discussed in more detail on our demonstrations page. A brief summary of the main points is below.

8. A professional programmer was employed to refactor the uportal front-end code and create an installer package that others can use to build their own repository system based on the DSpace open code.
9. The metadata held on SPECTRa has been added to and updated to make it fully compliant with the latest DataCite specifications. This involved adding cross-walks between the internal DSpace metadata schemas and those used by DataCite. The metadata includes full incorporation of the ORCID identifier.
10. Examples of data discoverability using the DataCite search resources have been collected.
11. We have added two further mechanisms to the machine actionable procedures available for automatic retrieval, visual presentation and re-use of data from the SPECTRa repository. These have been incorporated into a talk and demonstration (DOI2Data) intended for presentation at the FORCE2015 conference on Research communications and e-scholarship in January 2015.
12. The curation of a ten year old set from Cambridge University is now available as a SWORD-endpoint. The project is adding around 140,000 newly curated datasets and illustrating the use of critical discovery metadata.
13. A template based on the DOI2Data demonstrator has been deployed in three published high-profile journal articles. Two further articles are being prepared on the results of this project for peer-reviewed publication to encourage such activity by other authors. The template is also now incorporated into enhanced teaching notes for undergraduate lectures.
14. A presentation on these themes was given at the ODIN-2014 meeting in Amsterdam in September 2014.
15. Our project places Imperial 2nd of UK universities and the 7th highest datacentre globally in terms of external exposure, visibility and discoverability of open research data.

Lessons learnt

16. The importance of SEO or search engine optimisation using standards-based metadata deployment to enable data-discovery.
17. The unique local expertise in the procedures gained from the project will allow Imperial College to maintain its rank as a leading global site in RDM and data discovery and re-use.

Recommendations

18. With the SPECTRa data repository now upgraded to conform to the important DataCite metadata standards, we recommend that these features be considered for incorporation into both the institutional and any data repositories that may be deployed in the future.

19. External repositories such as Figshare/Arkivum either currently do not implement the features we have introduced, or do so only partially. We recommend that if such external repository service providers are considered for future use within College, these aspects be considered for inclusion in the operational requirement drafted as part of any procurement.
20. We encourage new RDM resources to be considered for incorporation into both undergraduate and postgraduate teaching as the means for introducing best practice in RDM at an early stage.

It is important to establish a culture of data curation and sharing in College. The provision of workflows of this type for other disciplines will help to overcome some of the impediments to sharing envisaged by even those researchers keen to share.