Imperial College London

Green Shoots: RDM Pilot at Imperial College London

Ian McArdle
Head of Research Systems & Information
i.mcardle@imperial.ac.uk

Torsten Reimer
Project Manager: Open Access & RDM
t.reimer@imperial.ac.uk

Presenting projects by: M. Bearpark & C. Fare; G. Thomas, S. Butcher & C. Tomlinson; M. Mueller; H. S. Rzepa, M. J. Harvey, N. Mason & A. Mclean; G. Gorman, C. T. Jacobs & A. Avdis; N. Jones

www.imperial.ac.uk/researchsupport/rdm/policy/greenshoots
IDCC15, 10th February 2015

Imperial College London

Imperial College London

- Seven London campuses
- Four Faculties: Engineering, Medicine, Natural Sciences and Business School
- Ranked 2nd in world
  (QS University Ranking)
- Net income (2014): £855m, incl. £351m research grants and contracts
- ~15,000 students, ~7,200 staff, incl. ~3,700 academic & research staff
- Staff publish ~10,000 scholarly articles per year

http://www.imperial.ac.uk/

Imperial College London

College Position Statement

"Imperial College London is committed to promoting the highest standards of academic research, including excellence in research data management. This includes a robust digital curation infrastructure that supports open data access and protects confidential data.

Imperial acknowledges legal, ethical and commercial constraints on data sharing and the need to preserve the academic entitlement to publication."

- Approved by Provost Board, publicised via Staff Briefing

Imperial College London

Investing in RDM

College acknowledges that excellence in RDM will require significant investment and academic engagement

Considering large research income and reputation, Imperial cannot afford to “get it wrong”

So where, specifically, should the College invest?

“Green Shoots” scheme is born

Imperial College London

"Green Shoots" Funding - £100K Investment

What did we want?

- Academically-driven projects to demonstrate best practice in RDM
- Specifically frameworks / prototypes that would comply with funder policies and College position
- Frameworks could be based either on original ideas or integrating existing solutions into the research process
- Projects that supported Open Innovation and open access for data

What did we hope to achieve?

- Encourage a "bottoms up" approach to maximise use of local early adopters and innovators
- Generate solutions that could be grown to support RDM more widely
- Demonstrate that innovative, academically-driven, beneficial RDM is possible and to stimulate this further
- Advice concerning how Imperial should proceed in supporting RDM

Imperial College London

FUNDING OPPORTUNITY:
Research Data Management

Funding is available for academically-driven projects to identify and generate exemplars of best practice in Research Data Management (RDM), specifically frameworks and prototypes that comply with key funder RDM policies and the College position.

There is an expectation that solutions will support open access for data and solutions that support Open Innovation are strongly encouraged.

More Information: http://www.imperial.ac.uk/researchstrategy/funding

Contact: Ian McArdle i.mcardle@imperial.ac.uk

Submission Deadline: Friday 28th March 2014

Imperial College London

Funded Projects

- Haystack – A Computational Molecular Data Notebook
  - M. Bearpark & C. Fare
- The Imperial College Tissue Bank: A Searchable Catalogue for Tissues, Research Projects and Data Outcomes
  - G. Thomas, S. Butcher & C. Tomlinson
- Integrated Rule-Based Data Management System for Genome Sequencing Data
  - M. Mueller
- Research Data Management in Computational and Experimental Molecular Science
  - H. S. Rzepa, M. J. Harvey, N. Mason & A. Mclean
- Research Data Management: Where Software Meets Data
  - G. Gorman, C. T. Jacobs & A. Avdis
- Research Data Management: Placing [Time Series] Data in its Context
  - N. Jones

Imperial College
London

Haystack – A Computational Molecular Data Notebook

M. Bearpark & C. Fare

Idea

- Extend a working prototype of a computational chemical IPython notebook making it available for all on github

Achievements

- Installation is now much simplified
- A tree document structure has been implemented
- Calculations using mainstream computational chemistry software can be set up
- Calculations can be submitted to run on a high-performance computing cluster
- Data from completed calculations can be retrieved and visualised

RDM Benefits

- Enables computational molecular researchers to easily share a curated subset of their results and document how those results were generated

More Information

- http://github.com/clyde-fare/cc_notebook

Imperial College London

Imperial College Tissue Bank: A Searchable Catalogue for Tissues, Research Projects and Data Outcomes

G. Thomas, S. Butcher & C. Tomlinson

Idea

- Extend the ICH tissue bank infrastructure to accept and catalogue research data alongside the collection of 60,000 physical tissues specimens and donor records

Achievements

- A tool to automatically exchange data with the National Cancer Registry was built, updating patient outcome data where known
- A pipeline to transfer summary sequencing data and metadata into the tissue bank and a UI to view this information
- Prototyped a means for tracking location of associated raw sequencing data for future development
- Began to investigate means to link publications back to associated tissue samples

RDM Benefits

- Enhances existing datasets and enables their reuse to maximise the benefits gained from each tissue sample

More Information

- http://www.imperial.ac.uk/tissuebank/

Imperial College London

Integrated Rule-Based Data Management System for Genome Sequencing Data

M. Mueller

Idea

- Set up a data management system for the DNA sequencing service that will integrate with existing central Imperial HPC infrastructure for processing, analysis and dissemination of raw data and analysis results

Achievements

- See system on following slide
- iRODS-based system was implemented that:
  - 1 – Transfers data from sequencer to HPC Service (different campus)
  - 2 – Data are reformatted and split by sample and project and a quality report generated
  - 3 – Reads are mapped to a reference genome, reformatting again, reducing file size
  - 4 – Further compression achieved via compression algorithm
  - 5 – Data transferred to a webserver and made available for download
- Overcame concerns over authentication by excluding the HPC storage from iRODS

RDM Benefits

- A robust infrastructure is now in place to effectively manage large volumes of complex sequencing data
- The data are being made publicly available for re-use of this expensive resource

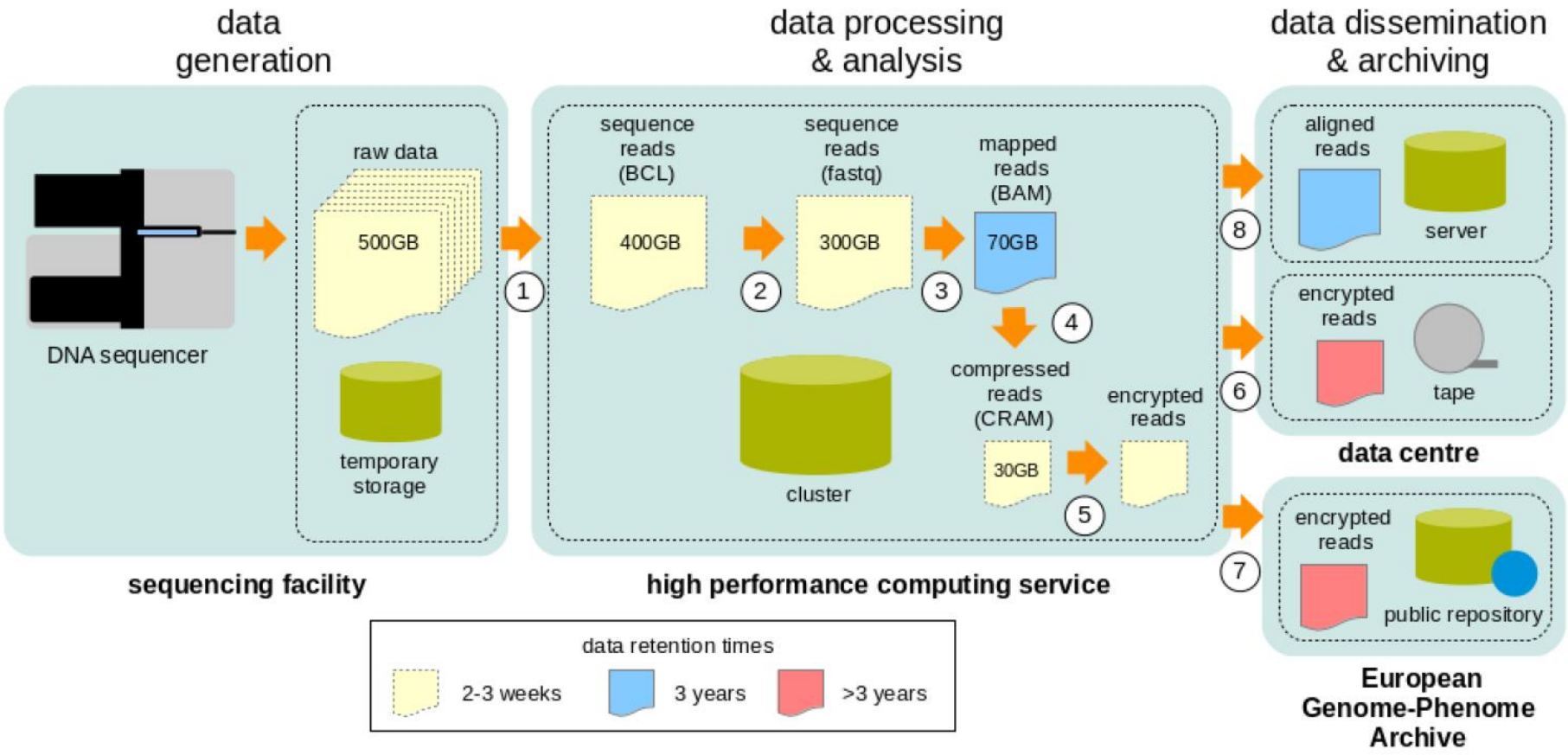
More Information

- http://www.imperial.ac.uk/genomicsfacility/informatics/

Imperial College London

Integrated Rule-Based Data Management System for Genome Sequencing Data

M. Mueller



Imperial College London

Research Data Management in Computational and Experimental Molecular Science

H. S. Rzepa, M. J. Harvey, N. Mason & A. Mclean

Idea

- Address sustainability and scalability of a hub interfacing electronic lab notebooks with HPC resources and digital data repositories

Achievements

- Produced an installer package to allow reuse of uportal DSpace front end
- Enhanced metadata in local repository to make it compliant with DataCite specifications – all repository content automatically receives a DOI
- Integrated ORCID into their solution
- Developed a procedure using DOIs for directly retrieving data from a digital repository and displaying it using Javascript components
- Curated 170,000 datasets from Cambridge to Imperial, adding standards-based metadata

RDM Benefits

- Molecular data can be referenced more robustly with persistent identifiers – step forward in data citation

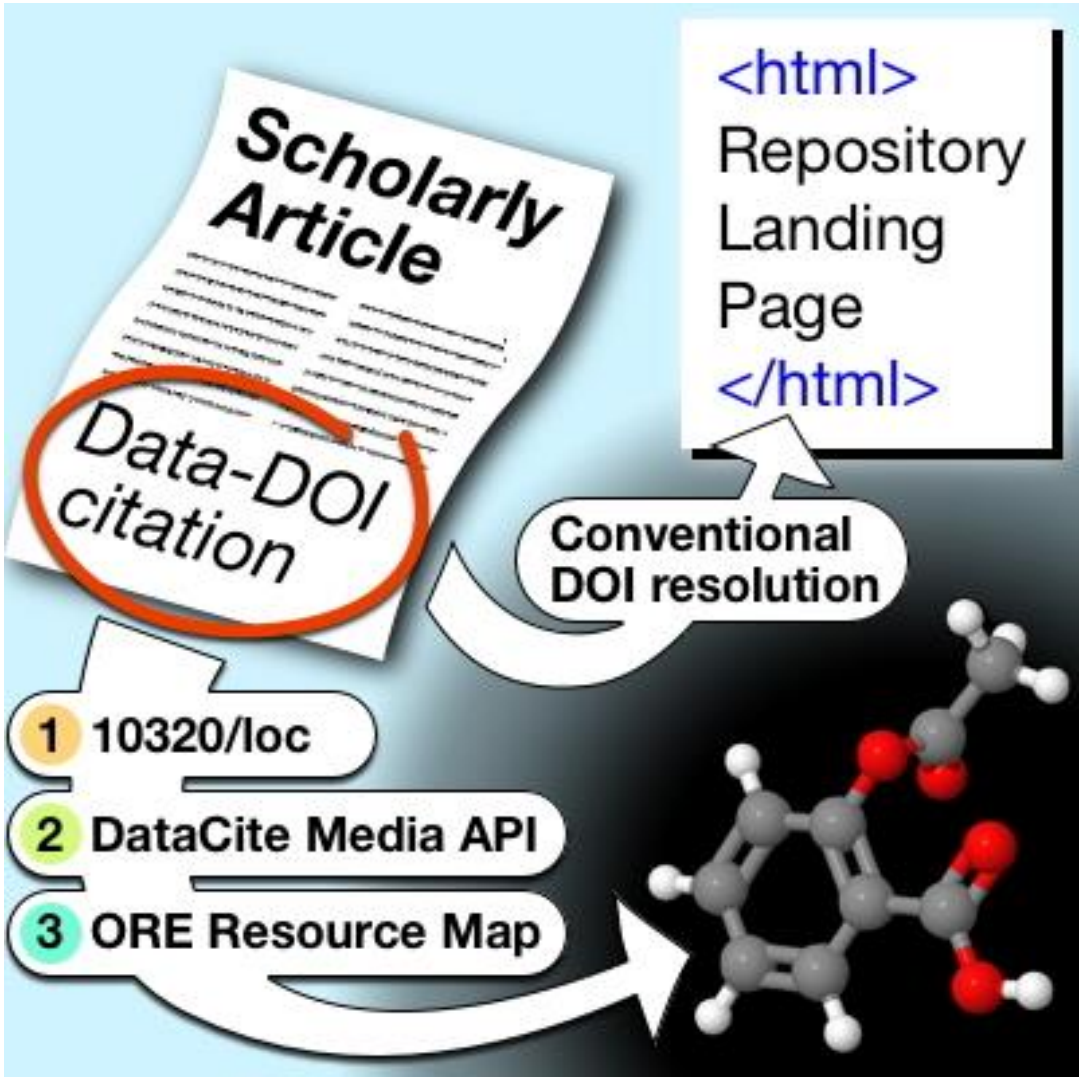
More Information

- http://doi.org/10042/a3v1w

Imperial College
London

Research Data Management in Computational and Experimental Molecular Science

H. S. Rzepa, M. J. Harvey, N. Mason & A. Mclean



Imperial College London

Research Data Management: Where Software Meets Data

G. Gorman, C. T. Jacobs & A. Avdis

Idea

- Integrating research data management into the research workflow so that data and software can be curated at the push of a button using Figshare and Git

Achievements

- Developed and released an open source software library: PyRDM
- Automatically transfers software source code (stored under Git control) and data to Figshare
- Figshare generates a DOI for that code version and the data
- Metadata including author details and cross-referencing between code and data are uploaded automatically
- Hoping for ORCID authentication via Figshare API to be added
- PyRDM was integrated into the Fluidity computational fluid dynamics code
- DOIs minted are stored in Fluidity to improve data provenance and allow a new revision of the repository to be created if the data are updated at a later stage

RDM Benefits

- Research data published in line with funder expectations
- The DOI for a specific code version enables better recomputability of data
- Automated metadata generation reduces academic burden

More Information

- http://github.com/pyrdm http://dx.doi.org/10.5334/jors.bj www.fluidity-project.org

Imperial College London

Research Data Management: Placing [Time Series] Data in its Context

N. Jones

Idea

- Provide a platform and technology which automatically connects researchers through their time-series data, models and analysis methods

Achievements

- Online interdisciplinary collection of time-series data and time-series analysis code
- Functionality to automatically profile time series
- Functionality to automatically profile time series algorithms
- Functionality to use these profiles to place a user's work in the context of others

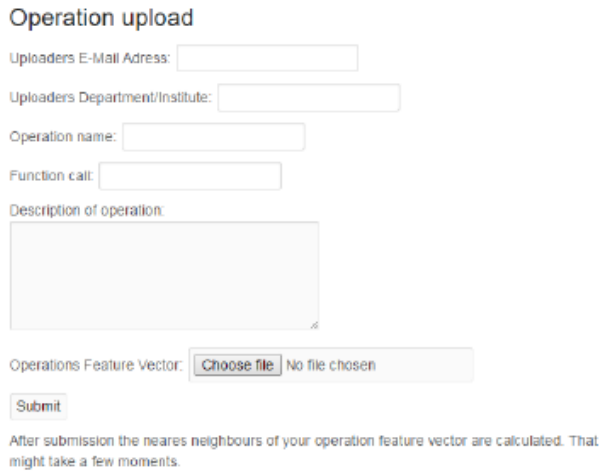
RDM Benefits

- Incentivises data sharing by allowing data comparison – increases discoverability of an academic's data plus increases likelihood of finding other relevant data
- Resource also available to general public

More Information

- http://www.comp-engine.org/timeseries/

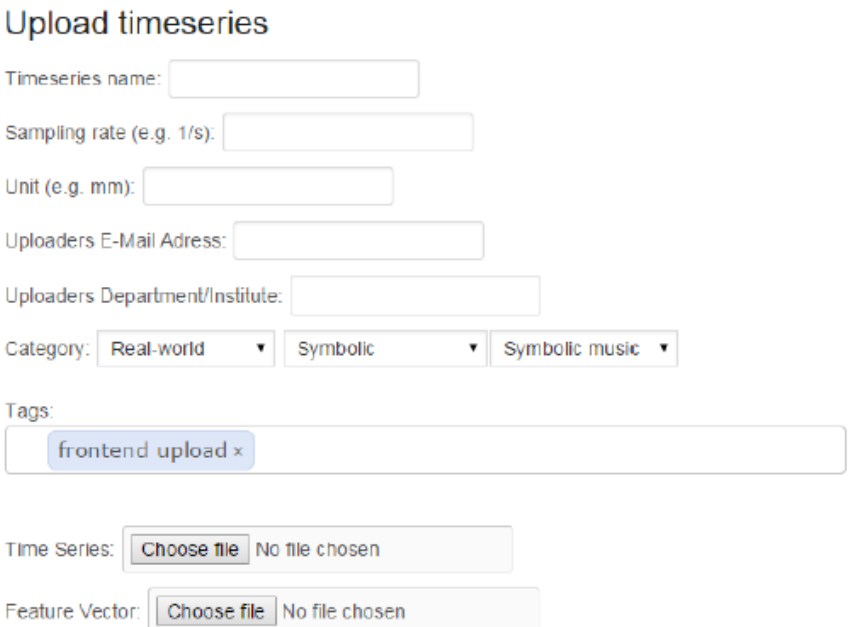
Incentivizing data sharing by allowing data comparison

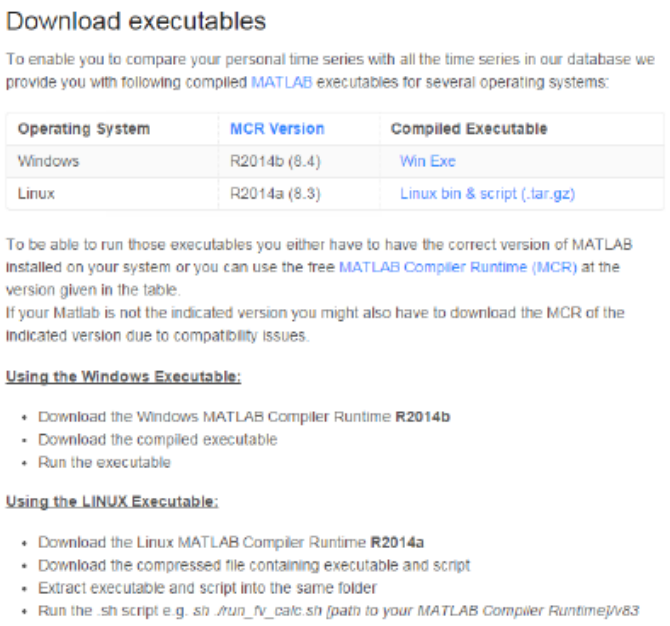


www.comp-engine.org/timeseries/

Users can now upload and compare time series by passing their data through a downloadable executable

They can also upload and compare their code

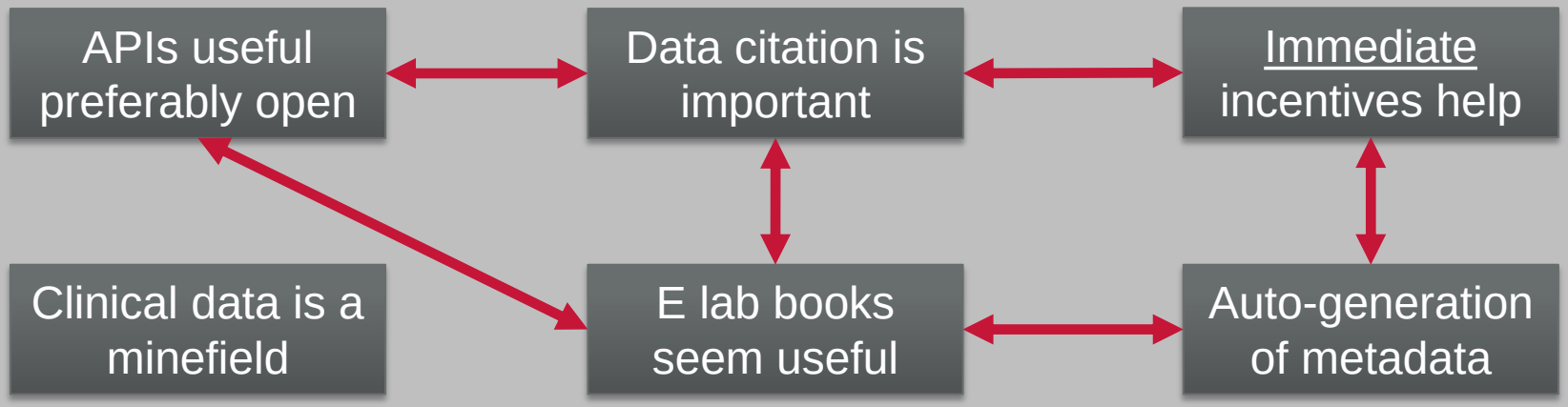




Imperial College London

Overall Conclusions

Good data curation is HARD and EXPENSIVE



Development of sustainable research software is *also* HARD and EXPENSIVE

Nucleus of an RDM community at Imperial

Ideas to consider for wider deployment for cross-College benefit

Imperial College London

Thanks and Questions

Review of applications:

- Kevin Ashley, DCC Director

Green Shoots academics:

- M. Bearpark & C. Fare
- G. Thomas, S. Butcher & C. Tomlinson
- M. Mueller
- H. S. Rzepa, M. J. Harvey, N. Mason & A. Mclean
- G. Gorman, C. T. Jacobs & A. Avdis
- N. Jones

Provision of funds:

- Imperial Vice-Provost Advisory Group: Research