Facilitating the deposit of experimental chemistry data in institutional repositories: Project SPECTRa

(**S**ubmission, **P**reservation, and **E**xposure of **C**hemistry **T**eaching and **R**esearch Dat**a**)

Peter Morgan
Project Director, SPECTRa
Cambridge University Library

Outline

- **Research data and Open Access**

- **Institutional context**

- **SPECTRa project**
  - the problem
  - methodology
  - investigations
  - software

- **Conclusions**

Research data & Open Access

Research data & Open Access

- Open Access declarations (Budapest, OECD, etc.) include access to data
- 855 repositories - only 6% contain data
- machine-understandable data is needed for:
  - e-Science
  - Semantic Web
  - re-analysis through informatics

Research data & Open Access

- Open Data is not the same as Open Access
- OA licences often don't address reuse of data
- OA papers associated with restricted data
- Open Data can co-exist with non-OA papers
- many publishers are unconvinced about OA
- therefore many researchers are also unconvinced about OA
- lack of exemplars to illustrate benefits
- lack of practical tools to overcome obstacles

Institutional context

University of Cambridge

- DSpace@Cambridge, 2003-
  - Cambridge University Library + Computing Service
- accepts all file formats, all types of content
- contains open- and closed-access collections
- few OA research papers
- large collection (>175,000 files) of chemistry data files
- Chemistry Dept (Peter Murray-Rust) keen to explore potential of repository

Imperial College London

- member of LEAP (London Eprints consortium)
- DSpace planned as institutional repository
  - to accept texts and supporting datasets
- Computational Chemistry (Henry Rzepa) in close collaboration with Cambridge colleagues
- potential links with High Performance Computing

SPECTRa

(**S**ubmission, **P**reservation, & **E**xposure of
Chemistry **T**eaching and **R**esearch dat**a**)

SPECTRa project

- **18-month project partnership between**

  Cambridge University Library (lead site)

  Cambridge University Chemistry Dept

  Imperial College London - Chemistry Dept

  Imperial College London - Library

- **in collaboration with eBank-UK**

- **funded by JISC** (Joint Information Systems Committee) **Digital Repositories Programme**

- **3 project staff plus librarians & chemists**

- **ended March 2007**

SPECTRa aims

"to investigate the needs of the academic chemistry research community in capturing and re-using experimental scientific data, facilitating the routine extraction of data in high volumes and their ingest into institutional repositories"

Methodology

- survey at Cambridge and Imperial of researchers' requirements in crystallography, computational chemistry, and synthetic organic chemistry

- development of customised Open Source tools as part of researchers' workflow to enable deposit of, and access to, Open Data using DSpace institutional repositories

Investigations

- **Crystallography**
  - interviews with key research leaders
- **Computational Chemistry:**
  - interviews with key research leaders
- **Synthetic organic chemistry**
  - survey of current use of computers & Internet
  - questionnaire (28 questions)
  - follow-up interviews
  - analysis of results

Survey results

- much data not stored electronically
- many file formats (mainly proprietary)
- significant ignorance of digital repositories
- repositories must be able to restrict access to experimental data

Data embargo

- **publication of chemical structures must be embargoed until the chemist**
  - publishes work involving those structures, or
  - moves on to a different line of research
- **chemists may need to hide research data from competitive research groups or for commercial reasons**
- **researchers are unlikely to deposit their data without adequate recognition of concerns**
- **need for an embargo process to control public release of data**

The SPECTRa workflow

- capture selected data
- validate against file specifications
- add metadata inc. InChI and systematic name
- add persistent identifiers (Handles)
- add METS packaging
- deposit in closed 'embargo' repository
- manage the release of embargoed data into open repository by agreement
- data can then be searched and harvested

Repository platform

- DSpace already adopted by both libraries
- separate DSpace repositories for SPECTRa
  - departmental installations
  - embargo repository managed at departmental level
  - potential role as intermediate repositories
- institutional repository architecture to include co-ordinated network of repositories?

SPECTRa Project architecture
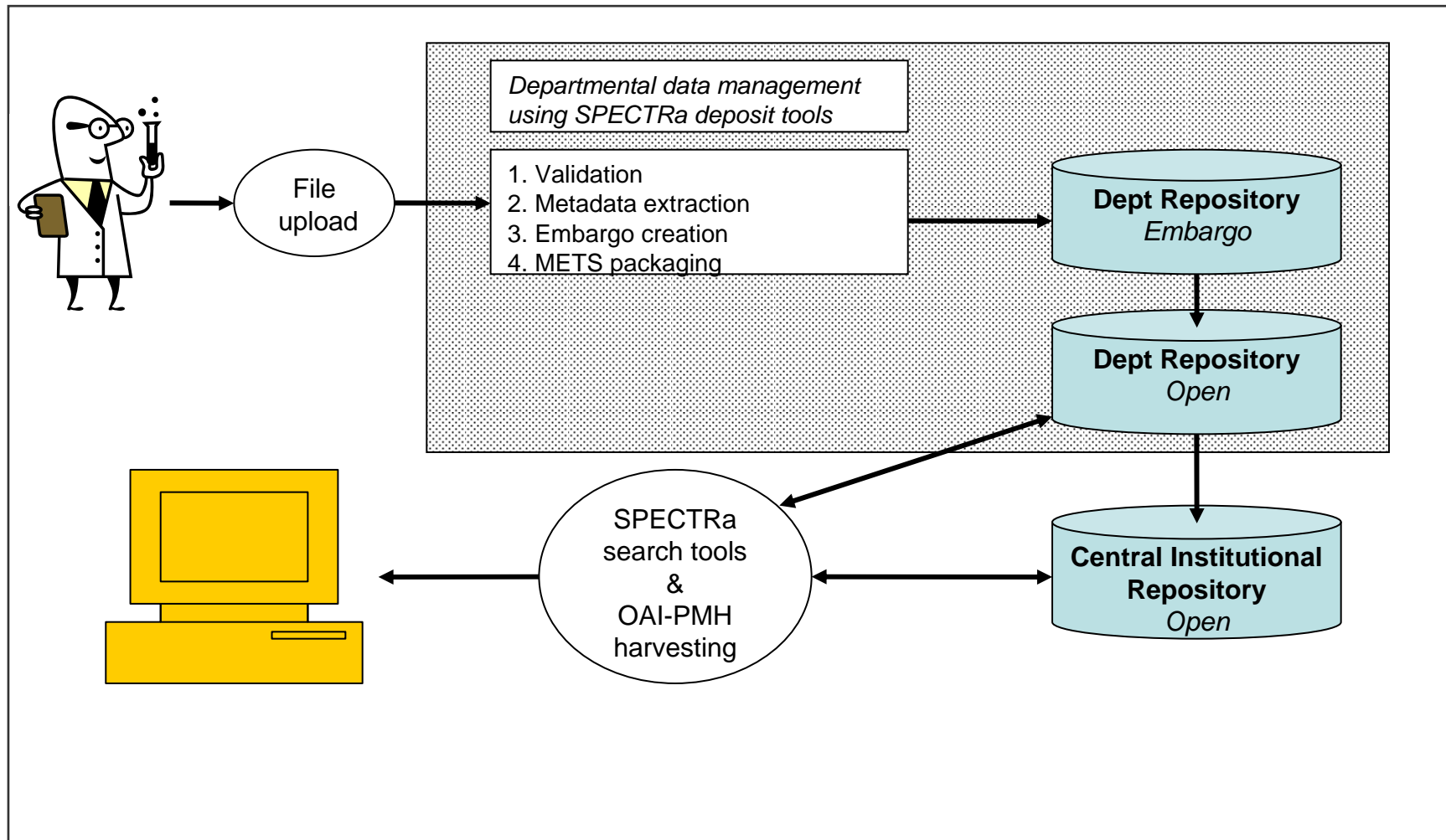
Conclusions

Conclusions (1)

- Discipline-specific repository tools
- The "Golden Moment"
- Data embargo
- Automated deposit & human editing
- DOIs have cost implications
- DSpace Handle system underdeveloped

Conclusions (2)

- Data re-use requires IPR guidance
- Scientific data = major asset value
- Local solutions are resource-intensive
- Generic solutions involve compromises
- Role of departmental repositories in an institutional repository architecture

Thank you for listening

Peter Morgan

pbm2@cam.ac.uk

SPECTRa: www.lib.cam.ac.uk/spectra/

DSpace@Cambridge: www.dspace.cam.ac.uk