The Digital Curation Centre (DCC) Information Day
University College London (UCL)
16th of January 2007

Introduction

The 11th DCC Information Day aimed to kick-start institutional thinking on digital curation at University College London (UCL). Additionally this event was an opportunity for participants from UCL to discuss their digital curation issues with each other as well as several delegates from the DCC. The twenty-nine participants included representatives from Physics, Astronomy, Biochemistry, the Centre for Health and Informatics and Multiprofessional Education (CHIME), the Education and Information Support Division (EISD), the School of Library, Archive and Information Studies (SLAIS) and Library Services. With a mixture of both e-Science and Library Science backgrounds, the group represented a range of disciplines and levels of understanding of digital curation.

Presentations

The event commenced with an introductory presentation delivered by Hugh Corley from the DCC. This presentation offered an overview of digital curation and stressed that the active management of digital resources over their entire life-cycle is essential for ensuring their long-term accessibility and reusability. Corley then went on to explain how the DCC can assist the UK research community to better curate their digital information. Delegates were encouraged to exploit this opportunity to establish contacts, to share resources, experiences, and strategies, and to develop reciprocal links for local curation and preservation support.

Professor Jon Butterworth of the Department of Physics and Astronomy provided the first of a series of presentations from UCL staff with a presentation on curation needs of data originating from high energy physics (HEP) research. He emphasised the immense scale of the HEP experiments, with large international teams of scientists working on dedicated facilities such as the ATLAS store that accommodates massive amounts of data from both experiments and simulations. The sheer cost of recreating data produced on this scale underlines the importance of data curation in this discipline, as it cannot easily or inexpensively be reproduced. Furthermore, as the analysis software used in experiments is often specific to a given experiment, the software – and associated usability - must be preserved alongside the data itself. Experiments tend to last for around ten years, with further data being added and different analyses carried out on an ongoing basis. Research teams often attempt to construct clean and definitive data sets towards the end of an experiment life-cycle, but this can be compromised as researchers' priorities move on to begin new projects and experiments. The result is a frequent under use of old HEP data, although Butterworth noted that there had been some re-analysis of data older then ten years, such as that from DESY (Deutsches Elektronen-Synchroton) and CERN's LEP (Large Electron and Positron collider). Butterworth noted that activities to curate data over

the course of an experiment's project life could be improved. He proposed that one solution would be to fund a specialist team to curate the data at specific intervals over the course of the project. The additional expense of these curation activities could be justified by the sheer cost and time invested in generating the data in the first instance. Through projects like CEDAR HepForge, Professor Butterworth and his colleagues have started to build an open source development environment that encourages documentation of physics software that can be reused for small to medium HEP experiments. The environment facilitates knowledge transfer beyond the specific research teams and experiments. Software in this environment includes tools for tuning and validating analytical models (HZTool and Rivet (Robust Independent Validation of Experiment and Theory)) as well as XML-based formats for the exchange of derived data (HepML).

Dr Jo Milan from UCL's Centre for Health Informatics and Multiprofessional Education (CHIME) followed, delivering a presentation on health informatics. The management of complex patient generated records requires data to be safely stored for future services and research. It was noted that although the general public believe that data are shared between different parts of the health system, this is not the case. Efforts are however underway to achieve such integration, for example through the NHS Connecting for Health programme and the National Institute for Health Research (NIHR). However, data integration is complicated by patient confidentiality requirements and research needs. The curation of health informatics data, like particle physics, tends to be detached from the data collection and research processes.

Drs Elizabeth Shephard, Melissa Terras, and Clare Warwick from the School of Library, Archive, and Information Studies (SLAIS) delivered the final presentation of the morning session that introduced various projects researching curation in the humanities that are currently underway at SLAIS:

- LEADERS (Linking EAD to Electronically Retrievable Sources) - an AHRC-funded project focused on the use of XML-based technologies (Encoded Archival Description (EAD), Encoded Archival Context (EAC) and the Text Encoding Initiative (TEI) standards) for facilitating improved access to and use of archival collections in digital form
- ReACH (Researching e-Science Analysis of Census Holdings) – an AHRC-funded project looking at the use of e-Science technologies in the humanities, starting with re-use of historical census data
- LAIRAH (Log Analysis of Internet Resources in the Arts and Humanities) - an AHRC-funded project investigating the re-use of online archived digital resources in the humanities, e.g., a log analysis of AHDS and institutional search portals revealed that a relatively large proportion of resources remained unused. The project also looked at the types of researchers using these resources and attempted to identify any common characteristics of well-used collections
- UCIS (User-Centred Interactive Search with Digital Libraries) - an EPSRC-funded project devoted to development of a better understanding of the use of digital libraries, including their integration into wider work patterns.

It is clear that the SLAIS can play an essential role in the development of an institutional records management and curation system. The value of cross-departmental collaboration in UCL on curation and preservation issues was widely

recognised in subsequent discussions, as well as collaboration with external and specialist centres such as the DCC.

Following some particularly engaging discussion over lunch, Michael Day from UKOLN and the DCC delivered the final presentation of the day on preservation metadata. He described the vital role that preservation metadata plays in the long-term curation of digital objects. Preservation metadata is the information necessary to ensure that a digital object is self-documenting in the long-term. This is vital for maintaining long-term access to this data through technological and social change. Day then described two key models within this context, OAIS and PREMIS. He concluded by discussing the sustainability of metadata and the role of shared registries.

Discussion

The group discussion was chaired by Professor Roland Rosner, Director of UCL's Education and Information Support Division (EISD). He began by drawing on the call to action raised in Corley's presentation, as the aim of the day was to kick-start action on digital curation issues at an institutional level and to identify the activities that are required for the management of corporate documentation and other information assets.

There was interest in finding model curation policies at other HE/FE institutions that could be followed. Most existing examples of such policies come from national archives and data centres and, it was argued, not always directly relevant. This raised the issue of what research JISC had done regarding curation policy and policy development at the HE/FE institutional level. Any new policy needs to encourage collaboration between records management approaches and digital curation. One suggestion is more regular liaising between JISC's digital curation initiatives and its records and information management activities.

It was decided that a member of SLAIS should be on UCL's Electronic Document Records Management Systems Procurement Committee and that this committee should have input in determine the policy for the creation, storage and curation of data throughout the university.

Professor Butterworth shifted discussion from the broader subject of corporate records to institutional repositories. He suggested that there needed to be some form of reward or recognition for those that create high-quality digital resources with appropriate metadata and additional representation information. Dr Elizabeth Shephard pointed out that many of the funding bodies in the arts and humanities already require grantees to deposit data with nominated data centres as a condition of funding, but that this is not the case with other funding bodies. Even then, funding bodies did not always stipulate the precise form in which data should be deposited.

The discussion then turned to the necessity of integrating institutional repository and corporate documentation (administrative records and academic publications) deposits with other services like library catalogues. Research undertaken by SLAIS demonstrated that academics, and particularly those in the humanities, place a significant amount of trust in their own libraries. In addition, other research demonstrated that if information was not easy to find then it would not be used. Paul

Ayris (Director of Library Services) pointed out that the suppliers of library systems - including Ex Libris which UCL use - are now beginning to provide systems that are compatible with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This facilitates users in finding information across different locations by making it widely accessible to metadata harvesting software. In any case, metadata - once created - can be made available for retrieval from many different places simultaneously, providing multiple routes to content.

Jeremy Yates, in one of the final points of the meeting questioned the value of keeping a printed PhD as the only record of research work since the thesis lacks the code and data that support the results of that research. This supporting data is not currently being adequately managed, with much of it stored in an unstructured form on hard drives within university departments. Ultimately, data generated and used in digital form must be curated and preserved in digital form if it is to remain quantifiable, re-usable, and validated.

This event aimed to kick-start institutional thinking on digital curation at UCL by providing a forum for UCL academics to come together and share their digital curation concerns, challenges and solutions. Motivated by their involvement in this event, Information Day participants agreed that a working group within UCL should be set-up to continue institutional discussions on digital curation. In collaboration with the DCC, the working group will address both digital curation and records management for all digital corporate outputs and assets. The DCC will follow up periodically to see how the working group is progressing and provide them with updates relating to the latest digital curation developments, especially with regard to the development of curation policies which was of particular interest to the participants. Information about the DCC's work to align recommendations on digital data curation with JISC will also be made available. The DCC looks forward to collaborating with UCL as they develop these digital curation policies.

List of Attendees

Name	Affiliation
Paul Ayris	Library Services
Jon Butterworth	Physics and Astronomy
Anna Clark	UCL Business PLC
Robert Clark	Research Computing (EISD)
Tim Cole	Institute of Child Health
Hugh Corley	DCC
Jan Cropper	Library Services
Rosamund Cummings	Estates and Facilities (Records Office)
Andrew Dawson	Research Computing (EISD)
Michael Day	DCC and UKOLN
Jane Fenoulhet	Faculty of Arts and Humanities
Nick Fox	Institute of Neurology
Mike Griffiths	Research Services
Clare Gryce	Research Computing (EISD)
David Hawkes	Centre for Medical Image Computing
Mike Hubank	Institute of Child Health
Jacob Hurst	Biochemistry

Kathryn Lewis	Education and Information Support Division (EISD)
Andrew Martin	Biochemistry
Jo Milan	Centre for Health Informatics and Multiprofessional Education (CHIME)
Martin Moyle	Library Services (Science Library)
Maureen Pennock	DCC and RSP
Vito Perrone	Computer Science
Jon Rimmer	School of Library, Archive, and Information Studies (SLAIS)
Roland Rosner	Education and Information Support Division (EISD)
Elizabeth Shepherd	School of Library, Archive, and Information Studies (SLAIS)
Melissa Terras	School of Library, Archive, and Information Studies (SLAIS)
Claire Warwick	School of Library, Archive, and Information Studies (SLAIS)
Jeremy Yates	Physics and Astronomy and Research Computing