DCC

The Digital Curation Centre

Michael Day
Digital Curation Centre
UKOLN, University of Bath
http://www.dcc.ac.uk/

**Society of Archivists EAD/Data Exchange Group meeting,
London, 8 December 2005**

UKOLN

Funded by: JISC e-Science dti

Presentation outline

- Definitions:
  - Digital curation and preservation
- The Digital Curation Centre:
  - Aims and objectives
  - Main task areas:
    - Research, Development, Services, Outreach
  - Standards
  - Collaboration with others

Definitions

Curation and preservation (1)

- Digital curation:
  - New(ish) term, from science data world (e.g. bioinformatics)
  - Reflects those extra things that need to be done to facilitate access and reuse
  - "... managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse" - Philip Lord, *et al*. (2004)
  - "Maintaining and adding value to a trusted body of information for current and future use" -- DCC presentation at CNI (2005)

Curation and preservation (2)

- Digital curation (continued):
  - Active management of data over life-cycle of scholarly and scientific interest
    - Reproducibility of results
    - Reuse and adding value
    - Managing digital information from point of creation
    - Ensuring long-term accessibility and preservation
    - Ensuring authenticity and integrity

Curation and preservation (3)

- Digital preservation:
  - Dealing with the potential technical problems that impede continued access to all types of digital resource
  - No longer possible to place physical artefact on a shelf and ignore for 100+ years
  - Sometimes seen as focused on the maintenance of specific object over time (e.g., a facet of curation)
  - But older definitions emphasise that it is not just a technical problem:
    - "... The planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable" - Margaret Hedstrom (1998)

Specific problems (1)

- An increasing flood of 'born-digital' data
  - The World Wide Web
    - Comprises billions of pages + "deep Web"
    - Internet Archive = >1 petabyte, and growing @ 20 Tb. per month (http://www.archive.org/)
  - Data deluge in science and engineering
    - Petabytes generated by high throughput instruments, streamed from sensors and satellites, etc.
    - Data-driven science, e-science, cyberinfrastructure, ...
  - 5 exabytes of *new* information created in 2002:
    - http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Specific problems (2)

- Need for (open) access to this data
  - Results in added scientific value
  - New analytic techniques
  - 2004 - OECD member states endorsed the principle that publicly funded research data should be openly available to the maximum extent possible
- Interoperability
  - Technical and cultural

The Digital Curation Centre (DCC)

DCC history (1)

- Background:
  - JISC Continuing Access and Digital Preservation Strategy
  - Lord and Macdonald report on e-science curation (2003): http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf
- JISC Circular 6/03 called for bids for a Digital Curation Centre (2003)
  - JISC and EPSRC funding:
    - For development, services and outreach in digital curation
    - For a research programme

DCC history (2)

- Main drivers:
  - The 'data deluge' resulting from e-science
  - An increasing awareness that:
    - Digital assets can be reused
      - Much science is now based on the reuse and recombination of data
    - Continuing access is vital to ensure that scholarship is reproducible and verifiable
    - Digital materials are inherently fragile

DCC purpose

- Supporting and promoting continuing improvement in the quality of data curation and digital preservation activity …
- Specifically ...
  - To promote preservation of digital information to support scholarship
  - To help enable scholarly communication and e-Learning

DCC objectives

- From proposal:
  - Lead a vibrant international research programme
  - Create an active, innovative and collaborative network of associates
  - Deliver effective, efficient and high demand services.
  - Evaluate tools, methods, standards and policies
  - Establish registries of tools and technical information

UKOLN

D C C

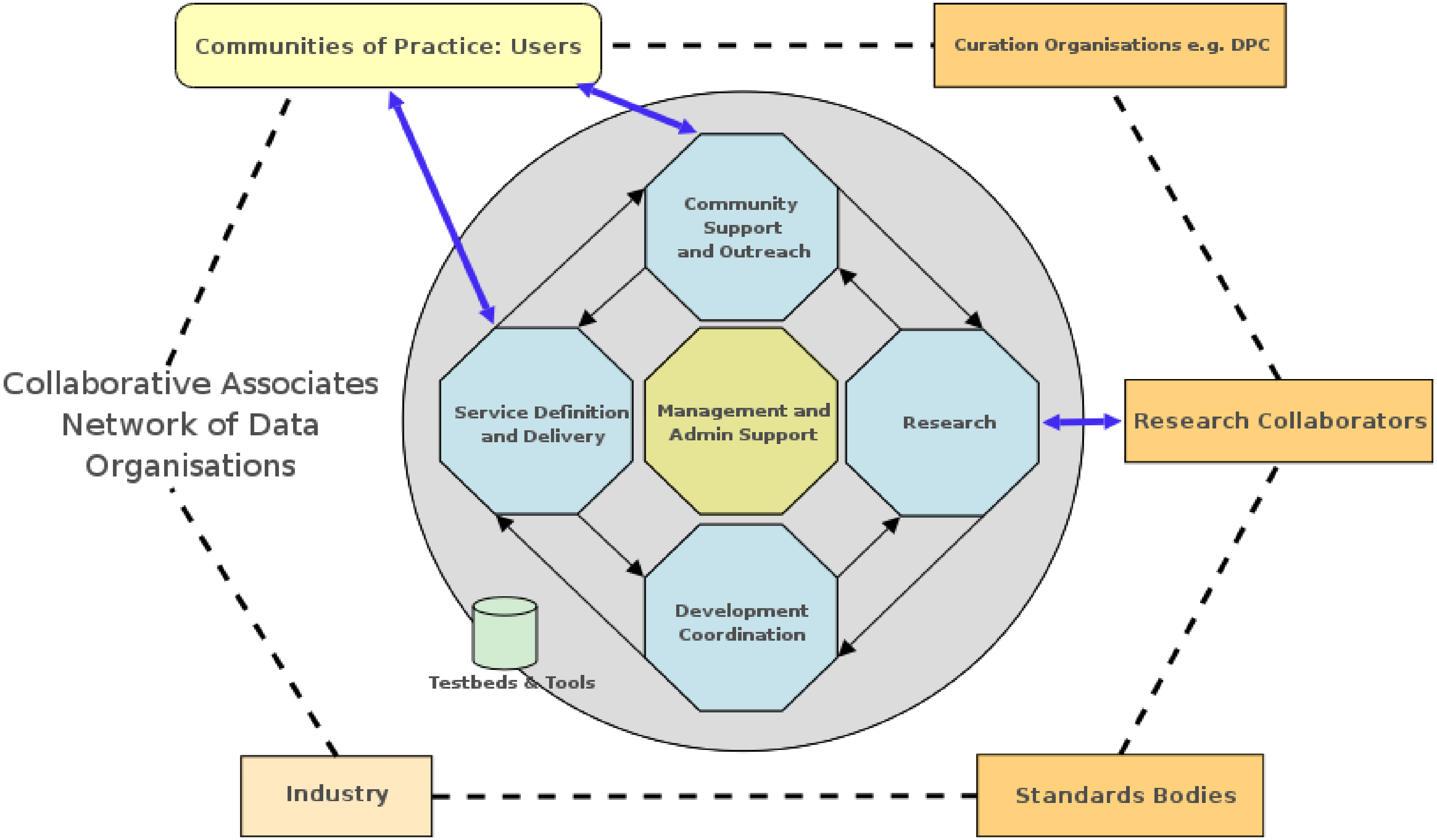
DCC partners

- University of Edinburgh (lead partner)
  - Chris Rusbridge (Director)
  - Prof. Peter Buneman (School of Informatics)
- University of Glasgow
  - Prof. Seamus Ross (Director of the Humanities Advanced Technology and Information Institute and ERPANET)
- UKOLN at University of Bath
  - Dr. Liz Lyon (Director of UKOLN)
- Council for the Central Laboratory of the Research Councils (CCLRC)
  - Dr. David Giaretta (Astronomical Software and Services)

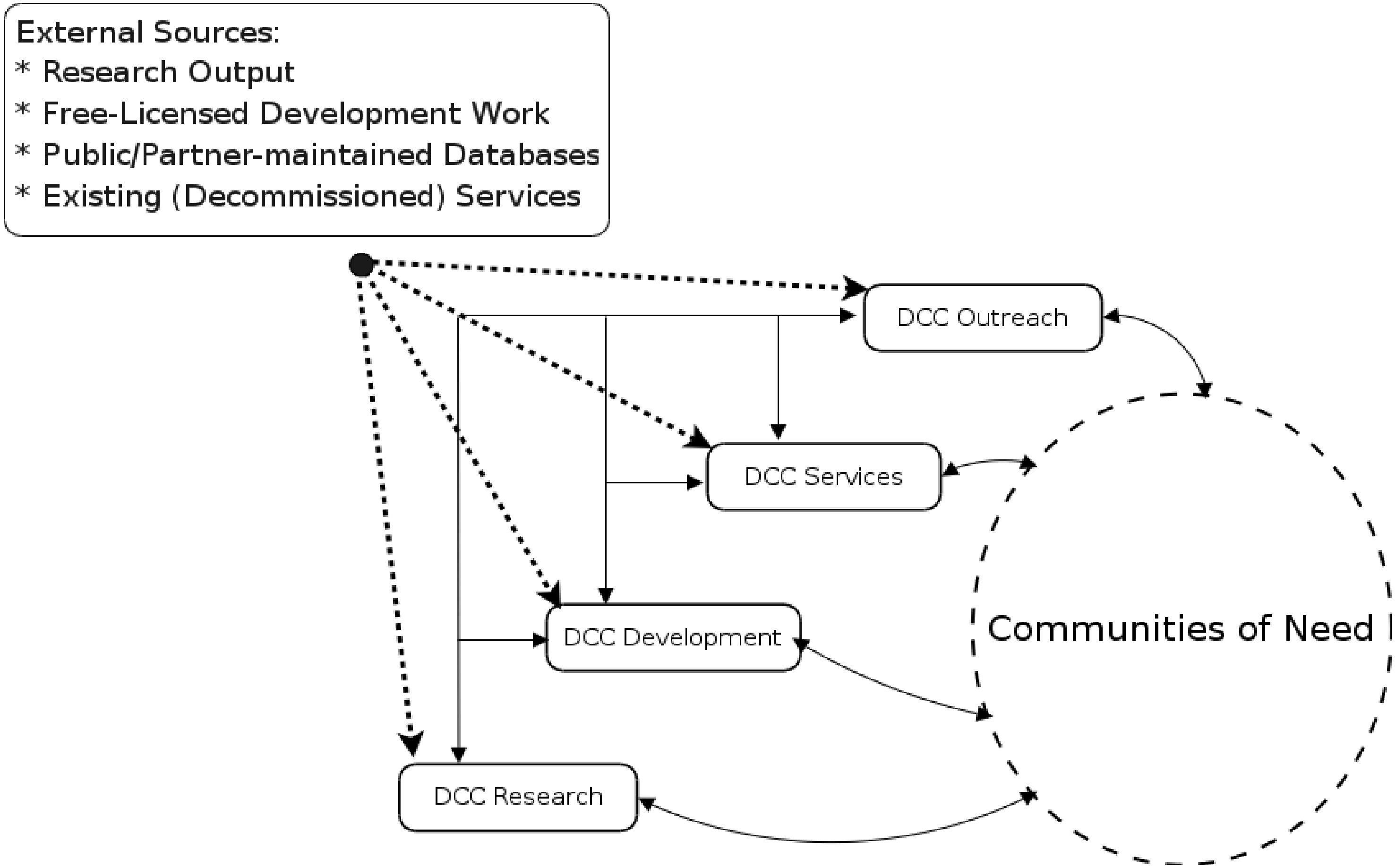
Engaging communities of practice (1)

- Those who have responsibility for curation
- Promoting good practice
- Engaging research in productive domains:
  - e.g. informatics, law, e-science ...
- Research and development should lead to services of relevance
  - To turn products of research and development into tools and services for use

Engaging communities of practice (2)


Communities of Practice: Users
Curation Organisations e.g. DPC
Community Support and Outreach
Collaborative Associates Network of Data Organisations
Service Definition and Delivery
Management and Admin Support
Research
Research Collaborators
Development Coordination
Testbeds & Tools
Industry
Standards Bodies

Engaging communities of practice (3)



External Sources:
* Research Output
* Free-Licensed Development Work
* Public/Partner-maintained Databases
* Existing (Decommissioned) Services

DCC Outreach

DCC Services

DCC Development

DCC Research

Communities of Need

DCC organisation

- Director:
  - Chris Rusbridge (University of Edinburgh)
- Four multi-partner teams:
  - Research (EPSRC grant) - led by Professor Peter Buneman (University of Edinburgh)
  - Development - led by David Giaretta (CCLRC)
  - Services - led by Professor Seamus Ross (University of Glasgow)
  - Outreach - led by Liz Lyon (UKOLN, University of Bath

DCC research team

- The DCC research team
  - Led by Professor Peter Buneman (School of Informatics, University of Edinburgh)
  - Concentrated in Edinburgh, but also distributed throughout all four DCC partner organisations
  - Strong links with other DCC components, through multi-team working, etc.
- Links with other research groups
  - Visitors programme

DCC research objectives

- To draw together the various functions of curation, from the traditional archival functions to the maintenance and publication of evolving knowledge as seen in scientific databases
- To conduct research in areas already identified by the partners as crucial to digital curation
- To identify through direct research collaboration, and through interaction with the service arm of DCC, the key projects in which research is needed
- To institute two-way conduits between research and service in which practical issues can be drawn to the attention of researchers and the products of research can be tested in practice

UKOLN

DCC

DCC research agenda

- Main topics:
  - Data integration and publishing
  - Annotation
  - Metadata extraction
  - Archiving and Appraisal
  - Legal issues
  - Provenance and data quality
  - Networks of trusted repositories
  - Economic cost-benefit analysis of curation

Current research priorities (1)

- Data transformation, integration and publication
  - Review of techniques
  - Schema directed XML publishing and integration
- Performance and optimisation
  - Safe data analysis environments within data centres
    - Initial testbed based on sky survey databases (in collaboration with the Wide Field Astronomy Unit and AstroGrid)

Current research priorities (2)

- Performance and optimisation (continued)
  - Automated metadata extraction and generation
    - Essential for testing the scalability of metadata-based preservation strategies
    - Review of tools, assessment of text mining techniques
  - Metadata curation
    - Dealing with changes in underlying metadata standards

Current research priorities (3)

- Annotation and provenance in databases
  - Scoping report
  - AstroDAS - Annotating sky objects over distributed astronomy catalogues
    - Builds on the concept of distributed annotation servers in bioinformatics (BioDAS)
  - Annotation Management of Scientific Databases
    - Data models for querying both data and annotations, MONDRIAN prototype to demonstrate the concept

Current research priorities (4)

- Annotation and provenance in databases (continued)
  - Formal models of provenance
    - stores provenance information about the effects of updates that modify the data, facilitating provenance queries of the form "Where did this data come from?" and "Has any part of this data been modified since it was obtained?"
  - Provenance retrieval in GIS

UKOLN

Current research priorities (5)

- Appraisal and long-term preservation
  - Appraisal techniques
    - Investigating the applicability and scalability of traditional appraisal techniques in 'data-intensive' contexts
    - Dynamic databases
    - Preservation techniques for evolving metadata and databases

UKOLN

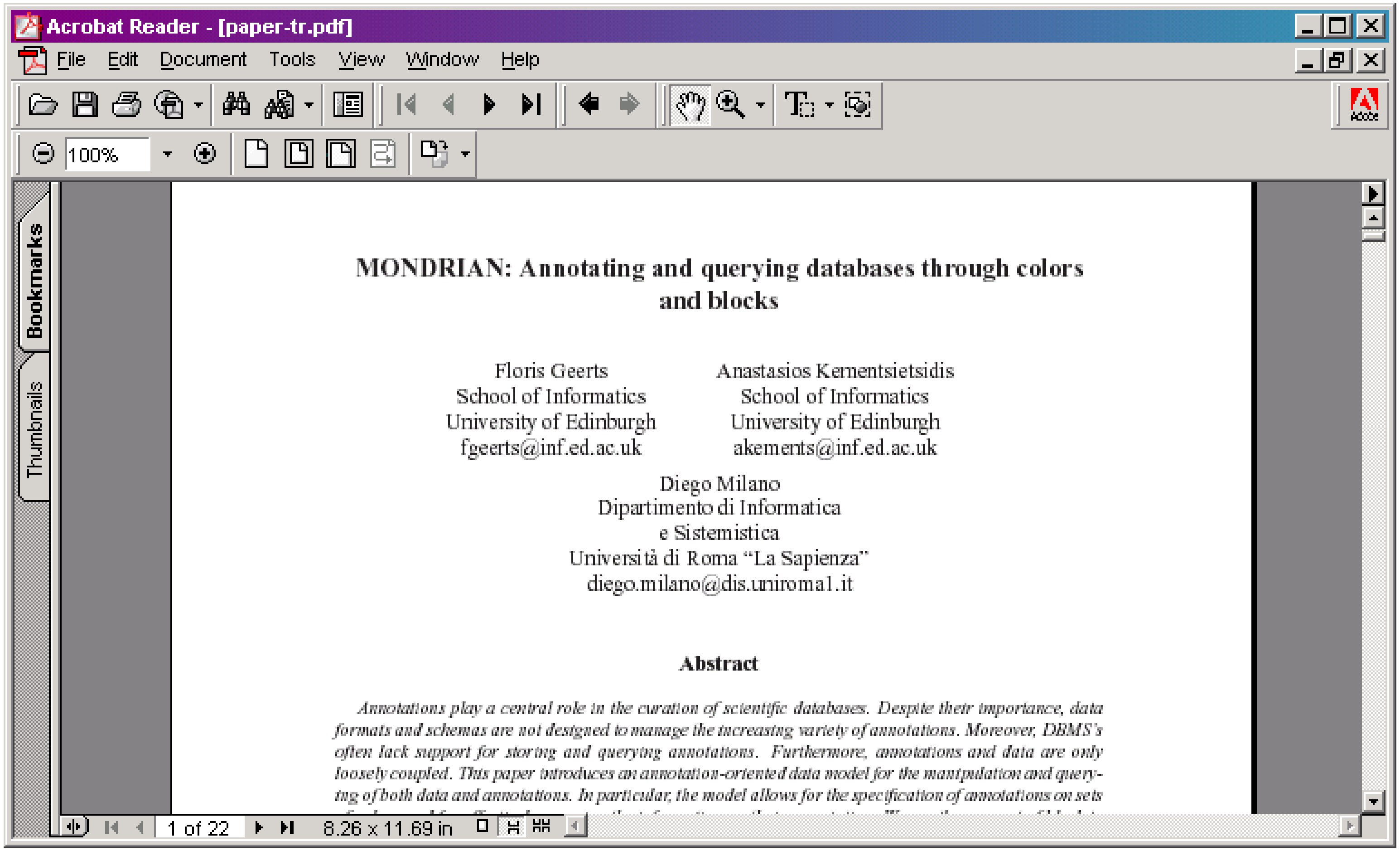
Current research priorities (6)

- Socio-economic and legal contexts
  - Networks of trusted repositories
    - Varying preservation role for repositories
    - Roles for co-operation, exchange formats, replication, etc.
  - Economic cost-benefit analysis of curation processes

Current research priorities (7)

- Socio-economic and legal contexts (continued)
  - Rights and responsibilities
    - The legal contexts of curation, e.g. impacts of the Database Directive on scientific data
    - Complexity of rights held in databases, impacts on aggregation and reuse of data

DCC research dissemination



Acrobat Reader - [paper-tr.pdf]
File Edit Document Tools View Window Help
100%
Bookmarks
Thumbnails
1 of 22
8.26 x 11.69 in

MONDRIAN: Annotating and querying databases through colors and blocks

Floris Geerts
School of Informatics
University of Edinburgh
fgeerts@inf.ed.ac.uk

Anastasios Kementsietsidis
School of Informatics
University of Edinburgh
akements@inf.ed.ac.uk

Diego Milano
Dipartimento di Informatica
e Sistemistica
Università di Roma "La Sapienza"
diego.milano@dis.uniroma1.it

Abstract

Annotations play a central role in the curation of scientific databases. Despite their importance, data formats and schemas are not designed to manage the increasing variety of annotations. Moreover, DBMS's often lack support for storing and querying annotations. Furthermore, annotations and data are only loosely coupled. This paper introduces an annotation-oriented data model for the manipulation and querying of both data and annotations. In particular, the model allows for the specification of annotations on sets

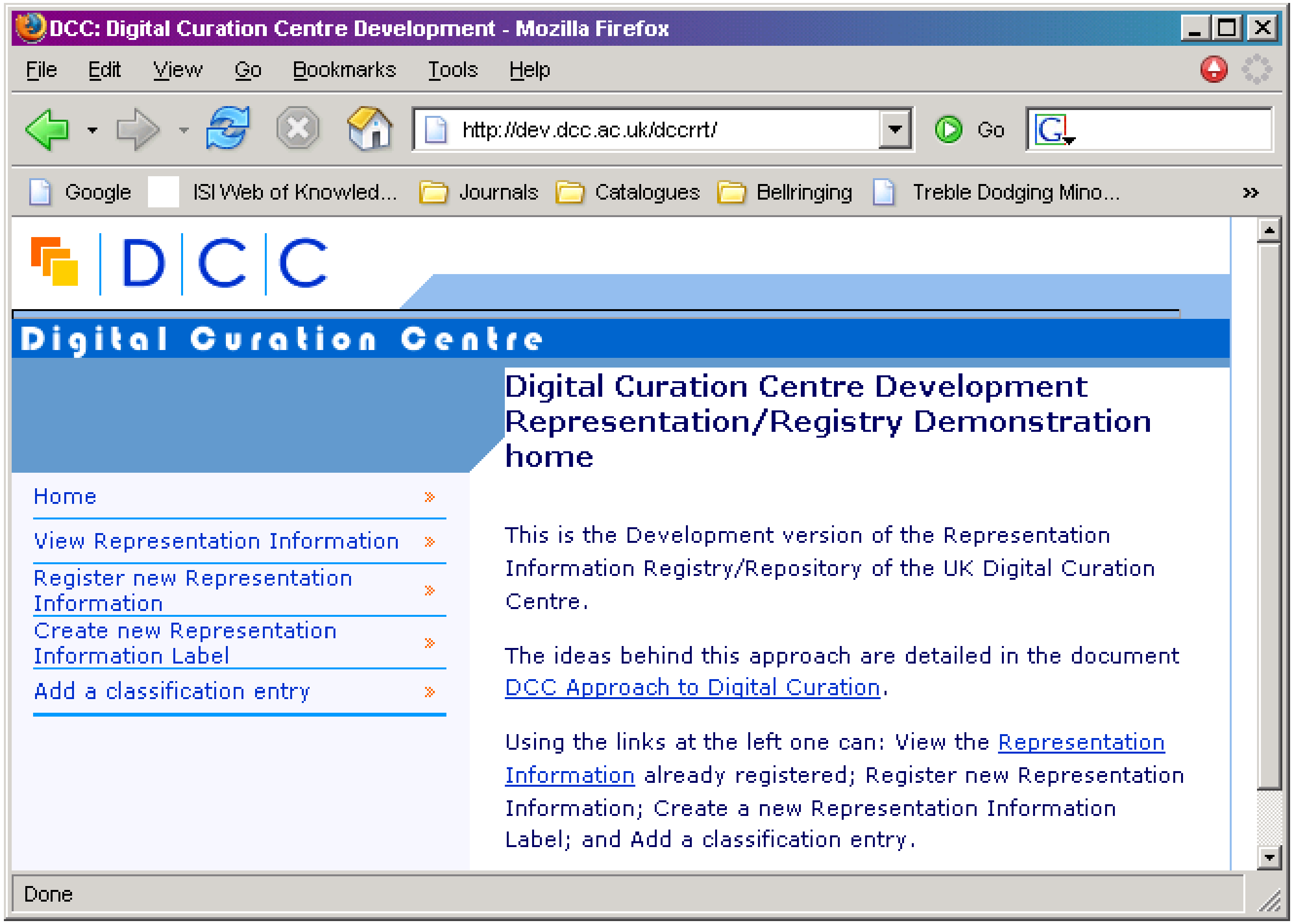
DCC development objectives

- "… to transform research-led innovation into services that enhance productivity of digital curation practice"
- Activities based on those defined by OAIS model:
  - Monitoring international standards
  - Development of a Representation Information registry/repository
  - Development of recommendations for tools and methods for generating Representation Information
  - Creating testbeds for digital curation tools
  - Developing audit and certification processes for trusted repositories

Current development priorities

- Representation Information registry/repository:
  - Work based on Representation Information concept defined by the Reference Model for an Open Archival Information System (OAIS) (ISO 14721:2003)
  - Representation Information = any additional information required (metadata, documentation, community knowledge, etc.) to render objects
  - Recognition that there is a need for trusted repositories of Representation Information
    - Information model for registry
    - Pilot registry (http://dev.dcc.ac.uk/dccrrt/)
    - Potential linking with file format registries like PRONOM or GDFR

DCC registry/repository (demo)



DCC: Digital Curation Centre Development - Mozilla Firefox

File
Edit
View
Go
Bookmarks
Tools
Help
http://dev.dcc.ac.uk/dccrrt/
Go
Google
ISI Web of Knowled...
Journals
Catalogues
Bellringing
Treble Dodging Mino...
DCC
Digital Curation Centre
Home
View Representation Information
Register new Representation Information
Create new Representation Information Label
Add a classification entry
Done

Digital Curation Centre Development Representation/Registry Demonstration home

This is the Development version of the Representation Information Registry/Repository of the UK Digital Curation Centre.

The ideas behind this approach are detailed in the document DCC Approach to Digital Curation.

Using the links at the left one can: View the Representation Information already registered; Register new Representation Information; Create a new Representation Information Label; and Add a classification entry.

DCC services mission statement

- *To transform research and development results into facilities and resources that improve digital curation practices and action within the scientific and research communities in the United Kingdom*

DCC information services

- Curation Manual:
  - internationally renowned and in-depth technical expertise on a range of digital curation topics
    - Open source software (already available); metadata (various topics), appraisal and selection, etc. (available soon)
    - http://www.dcc.ac.uk/resource/curation-manual/chapters/
- Briefing Papers:
  - Quick insights into a range of digital curation topics
- Tools Repository

DCC advisory services

- Help desk system:
  - Provides on-demand responses to queries - from legal to technical guidance (info@dcc.ac.uk)
- Database of FAQs
- Checklist for compliance with best practices and standards
- Case Studies
- Preservation Technology and Standards Watch
- Information Days

DCC training services

- On-line learning resources
- Training events to bring together practitioners and researchers
- Integrated educational resources
- Forthcoming Workshops/Training Events:
  - Future-proofing Institutional Websites, Wellcome Library, London, 19-20 January 2006 (http://www.dcc.ac.uk/training/fpw-2006/)

UKOLN

Audit and certification

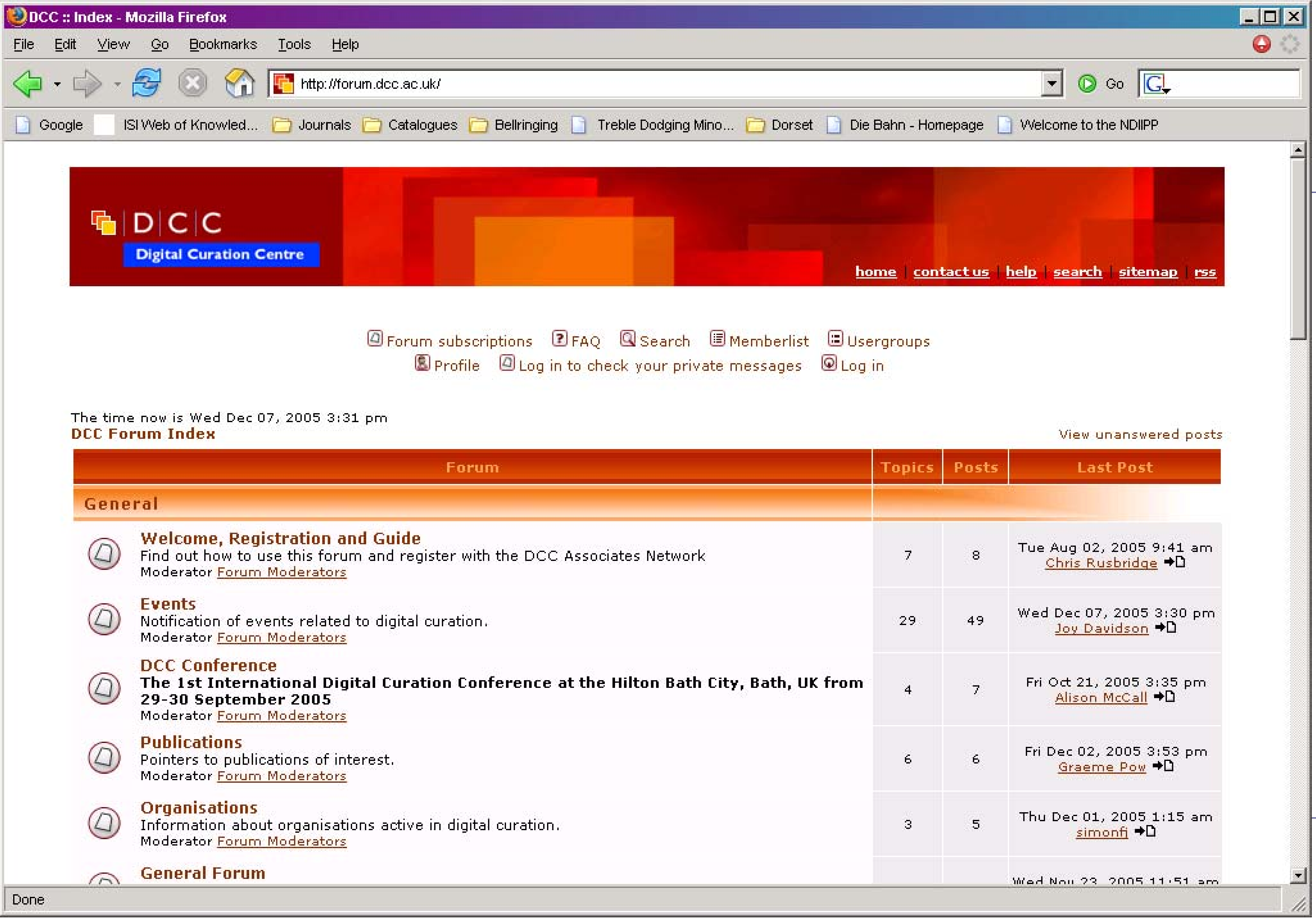
- *An audit checklist for the certification of trusted digital repositories* - RLG-NARA Task Force on Digital Repository Certification (draft for public comment, August 2005)
  - http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf
  - DCC collaborating with RLG in using the checklist to audit two UK scientific data repositories
  - RLG DigiNews article by Seamus Ross and Andrew McHugh: http://www.rlg.org/en/page.php?Page_ID=20793#article1
  - May eventually lead to DCC certification activity (?)

Outreach objectives

- Raising awareness and dissemination
  - Web portal (http://www.dcc.ac.uk/)
  - International Journal of Digital Curation (http://www.ijdc.net/)
  - Annual International Conference
    - Bath, 29-30 September 2005
    - Glasgow, October 2006
- Understanding users and their needs
  - Gathering user requirements
- Associates Network

DCC Associates Network

- The DCC has identified the importance of engaging with
  - Institutions
  - Organisations
  - Individuals
- Across *all* disciplines and domains
- More information: http://www.dcc.ac.uk/associates/
- DCC Forum: http: http://forum.dcc.ac.uk/



Forum subscriptions FAQ Search Memberlist Usergroups
Profile Log in to check your private messages Log in

The time now is Wed Dec 07, 2005 3:31 pm

DCC Forum Index

View unanswered posts

Forum	Topics	Posts	Last Post
General			
Welcome, Registration and Guide Find out how to use this forum and register with the DCC Associates Network Moderator Forum Moderators	7	8	Tue Aug 02, 2005 9:41 am Chris Rusbridge
Events Notification of events related to digital curation. Moderator Forum Moderators	29	49	Wed Dec 07, 2005 3:30 pm Joy Davidson
DCC Conference **The 1st International Digital Curation Conference at the Hilton Bath City, Bath, UK from 29-30 September 2005** Moderator Forum Moderators	4	7	Fri Oct 21, 2005 3:35 pm Alison McCall
Publications Pointers to publications of interest. Moderator Forum Moderators	6	6	Fri Dec 02, 2005 3:53 pm Graeme Pow
Organisations Information about organisations active in digital curation. Moderator Forum Moderators	3	5	Thu Dec 01, 2005 1:15 am simonfj
General Forum			Wed Nov 23, 2005 11:51 am

Some issues

The future of DCC research

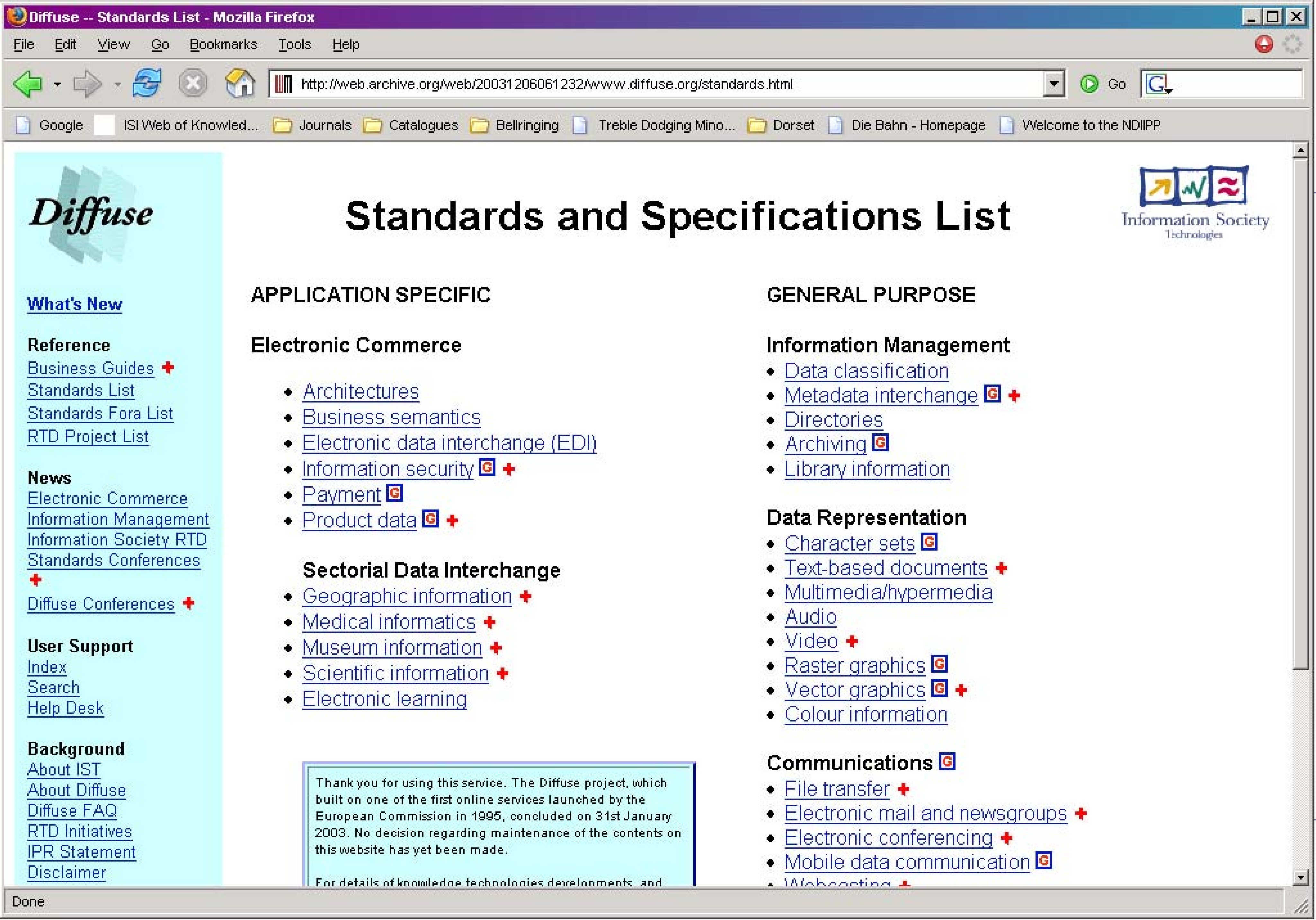
- DCC's EPSRC grant due to run until 2007
- Identifying a future research agenda:
  - US National Digital Infrastructure and Preservation Program, *It's about time* (August 2003) http://www.digitalpreservation.gov/
  - NSF-DELOS Working Group on Digital Archiving and Preservation, *Invest to save* (2003) http://eprints.erpanet.org/94/
  - Warwick Workshop on Digital Curation and Preservation (7-8 November 2005) - draft report available: http://www.dcc.ac.uk/training/warwick_2005/

Standards (1)

- The DCC is interested in a range of different standards:
  - The OAIS model (ISO 14721:2003)
  - Metadata (e.g., PREMIS Data Dictionary, METS, ISO 23081 Metadata for records, etc.), data description standards (e.g., EAST)
  - Content packaging standards (e.g., METS, MPEG-21 DIDL, XFDU (CCSDS), IMS Content Packaging)
  - Identifiers
  - …

Standards (2)

- DCC Standards Watch activity:
  - Still under development
  - Possibly to be based on an updated version of the Diffuse standards and specifications list (an EU funded project), currently only available via the Internet Archive's Wayback Machine
  - There are overviews of content packaging standards on DCC development wiki: http://dev.dcc.ac.uk/twiki/bin/view/Main/ContentPackaging



Diffuse -- Standards List - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://web.archive.org/web/20031206061232/www.diffuse.org/standards.html
Go
Google
ISI Web of Knowled...
Journals
Catalogues
Bellringing
Treble Dodging Mino...
Dorset
Die Bahn - Homepage
Welcome to the NDIIPP
Diffuse
Information Society Technologies
Done

What's New

Reference
Business Guides +
Standards List
Standards Fora List
RTD Project List

News
Electronic Commerce
Information Management
Information Society RTD
Standards Conferences +
Diffuse Conferences +

User Support
Index
Search
Help Desk

Background
About IST
About Diffuse
Diffuse FAQ
RTD Initiatives
IPR Statement
Disclaimer

Standards and Specifications List

APPLICATION SPECIFIC

Electronic Commerce

- Architectures
- Business semantics
- Electronic data interchange (EDI)
- Information security G +
- Payment G
- Product data G +

Sectorial Data Interchange

- Geographic information +
- Medical informatics +
- Museum information +
- Scientific information +
- Electronic learning

Thank you for using this service. The Diffuse project, which built on one of the first online services launched by the European Commission in 1995, concluded on 31st January 2003. No decision regarding maintenance of the contents on this website has yet been made.

For details of knowledge technologies developments, and

GENERAL PURPOSE

Information Management

- Data classification
- Metadata interchange G +
- Directories
- Archiving G
- Library information

Data Representation

- Character sets G
- Text-based documents +
- Multimedia/hypermedia
- Audio
- Video +
- Raster graphics G
- Vector graphics G +
- Colour information

Communications G

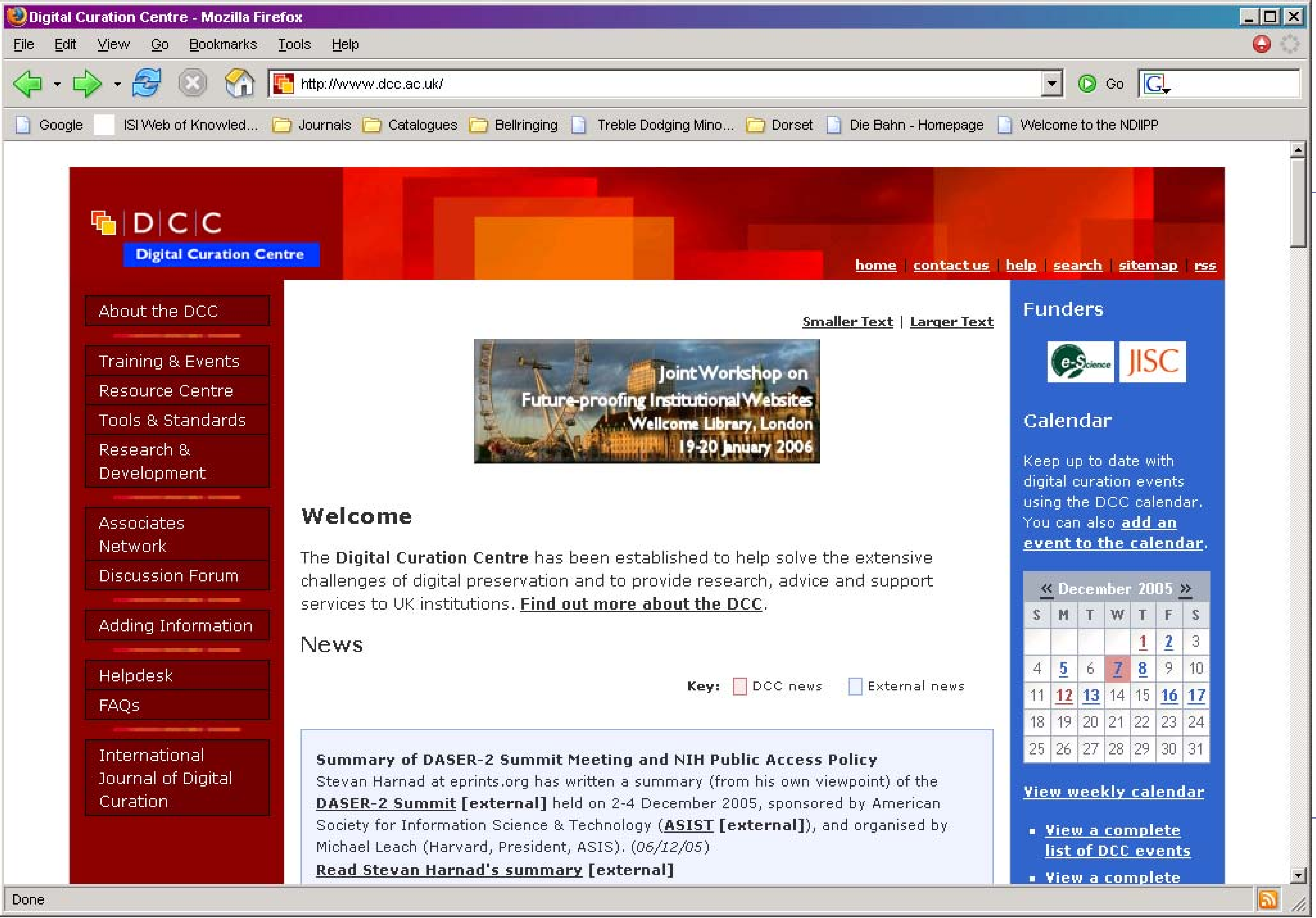
- File transfer +
- Electronic mail and newsgroups +
- Electronic conferencing +
- Mobile data communication G
- Webcasting +

Collaboration with others

- Collaboration is very important for the DCC
  - We are aware that the DCC is not the only source of expertise in the digital curation domain
- DCC responses:
  - Membership of (and participation in) Digital Preservation Coalition
  - Associates Network
  - Joint organisation of events
  - Information Days and other outreach activity

Further information

– Digital Curation Centre (DCC) Web portal: http://www.dcc.ac.uk/



DCC
Digital Curation Centre

home | contact us | help | search | sitemap | rss

About the DCC

Training & Events
Resource Centre
Tools & Standards
Research & Development

Associates Network
Discussion Forum

Adding Information

Helpdesk
FAQs

International Journal of Digital Curation

Smaller Text | Larger Text

Joint Workshop on Future-proofing Institutional Websites
Wellcome Library, London
19-20 January 2006

Welcome

The **Digital Curation Centre** has been established to help solve the extensive challenges of digital preservation and to provide research, advice and support services to UK institutions. **Find out more about the DCC**.

News

Key: □ DCC news □ External news

Summary of DASER-2 Summit Meeting and NIH Public Access Policy
Stevan Harnad at eprints.org has written a summary (from his own viewpoint) of the **DASER-2 Summit** [external] held on 2-4 December 2005, sponsored by American Society for Information Science & Technology (**ASIST** [external]), and organised by Michael Leach (Harvard, President, ASIS). *(06/12/05)*
Read Stevan Harnad's summary [external]

Funders

e-Science JISC

Calendar

Keep up to date with digital curation events using the DCC calendar. You can also **add an event to the calendar**.

« December 2005 »

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

View weekly calendar

- **View a complete list of DCC events**
- **View a complete**

Questions?

Acknowledgements

- The Digital Curation Centre is an initiative of the the Joint Information Systems Committee (JISC) and the e-Science Core Programme of the UK research councils. The consortium is led by the University of Edinburgh and includes the University of Glasgow (HATII), the Council for the Central Laboratory of the Research Councils, and UKOLN, University of Bath: http://www.dcc.ac.uk/
- UKOLN is funded by the Museums, Libraries and Archives Council (MLA) and the JISC, as well as by project funding from the JISC, the European Union and other sources. UKOLN also receives support from the University of Bath, where it is based (http://www.ukoln.ac.uk/)

UKOLN