UK Digital Curation Centre

http://www.dcc.ac.uk

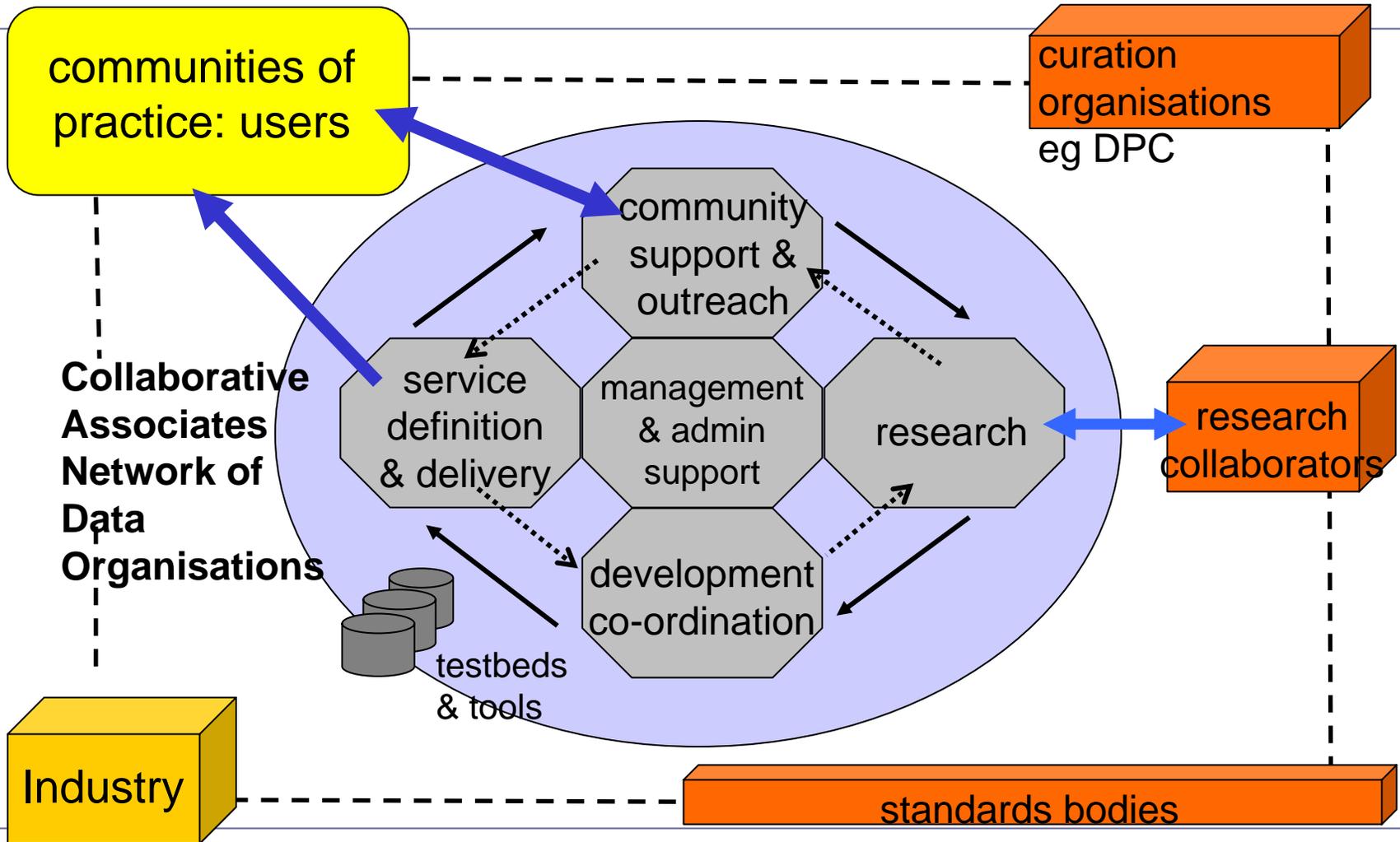## David Giaretta

Associate Director (Development)

Background to the DCC (1)

- The current generation of 'e-Science' experiments and computations will create more scientific data in the next five years than has been collected in the whole of human history.
- Exponential growth, combined with increasing diversity and complexity in all forms of electronic publication and resources used to support research and learning.
  - These electronic publications and resources are increasingly the primary corpus of record for research and education.
- The term 'digital curation' is increasingly being used for the actions needed to maintain and utilise digital data and research results over their entire life-cycle for current and future generations of users.
  - Data curation in science may be defined as the maintenance of a body of trusted data to represent the current state of knowledge in some area of research.
- **NB the Digital Curation Centre is <u>NOT</u> a repository of Primary Research data**

2

D|C|C

Background to the DCC (2)

- **Funding: UK JISC** and **e-Science Core Programme**
  - £750K (+ £300K) /year (outreach, services & development)
  - £370K /year (research)
- Successful Consortium of four institutional partners
  - Universities of Edinburgh (lead), Glasgow & Bath together with CCLRC (Rutherford Appleton and Daresbury Laboratories)
- Initial 3-year awards made:
  - JISC funding started on 1st March 2004
    - Long term funding expected
  - Research grant-funded started on 1st September 2004
- Phase One set-up
  - some 'early deliverables' of website & helpdesk
  - preparation for full operation & launch of services
  - formal opening was on 5th November 2004

3

D|C|C

Developing the collaborative model

communities of practice: users

curation organisations
eg DPC

Collaborative Associates Network of Data Organisations

community support & outreach

service definition & delivery

management & admin support

research

development co-ordination

research collaborators

testbeds & tools

Industry

standards bodies

4

D|C|C

Aims & Objectives for the DCC

'quality improvement in data curation & digital preservation'

- – Initial focus: data as evidence for scholarly conclusions
- – Wider remit: worlds of scholarly communication & e-Learning
- – Even wider: advice to all

- twin aims:excellence in research & excellence in service

- need to bridge across communities:
  - – universities & research institutes
  - – scientific data tradition & document tradition
  - – multi-sectoral, international

digital curation = f(digital preservation, data curation)

- – actions needed to maintain and utilise digital information & research results over entire life-cycle, for current and future use.

5

D|C|C

Outreach Overview

- Raising awareness and dissemination

  - Web portal

  - Electronic Journal

  - Annual International Conference

- Understanding the users and their needs

  - Creating a taxonomy

  - Gathering requirements

- The Associates Network

Services Overview

- Service Planning and Monitoring
  - Plan and organise Services stream of the DCC
- Standards, tools, testbeds, and Repositories
  - Ensure access to standards, tools and repositories of Representation Information
- Advisory Services
  - Provide guidance on curation and preservation
    - Help Desk
    - Curation Manual (using expertise from within DCC and beyond)
- Audit and Certification
  - Deliver Audit and Certification Services to the UK HE, FE and research Community
    - and beyond…

Research Overview

- Data integration and publishing
  - Slowly coming to market. Publishing in community formats is a new twist
- Annotation
  - Everybody agrees this is important. No-one understands it.
- Archiving and Appraisal
  - What do we do about databases – they change!
- Legal issues
  - Can we at least help to clarify what's going on?
- Provenance and data quality
  - Again, we don't fully understand it.
- Ontologies, performance, registries, structure evolution…

8

D|C|C

Development Overview

- Not as "free" as Research but not as "restricted" as Services
- Will not duplicate existing useful work
- Need to identify niche for immediate services
- Need to identify collaborators worldwide

- Need to ask fundamental questions and develop new services
- Need Guidelines for prioritisation….

9

D|C|C

What can we rely on in the Long Term?

- The bits (original or migrated)
  - let us for the moment put to one side the issue of BIT PRESERVATION and authenticity etc (but it is an issue)
- Physical documents that people can read
  - e.g. ISO standards on paper
- Additional information we collect – either held by the DCC, its collaborators or successors

10

Preservation "vs" Current Use

- There are already very many architectures to support immediate use of information
  - Aim to support these
- Therefore chose to be guided by
  - long-term preservation aspects
    - try to ensure that components of the preservation architecture can supplement other "current use" architectures.
  - to promote this we should emphasise "interoperability" and "automated use" as far as possible.
  - based initially on OAIS Reference Model – but not limited to that
- maybe science data are easier to start with than documents…..

D | C | C

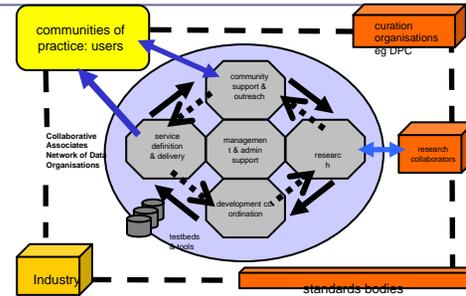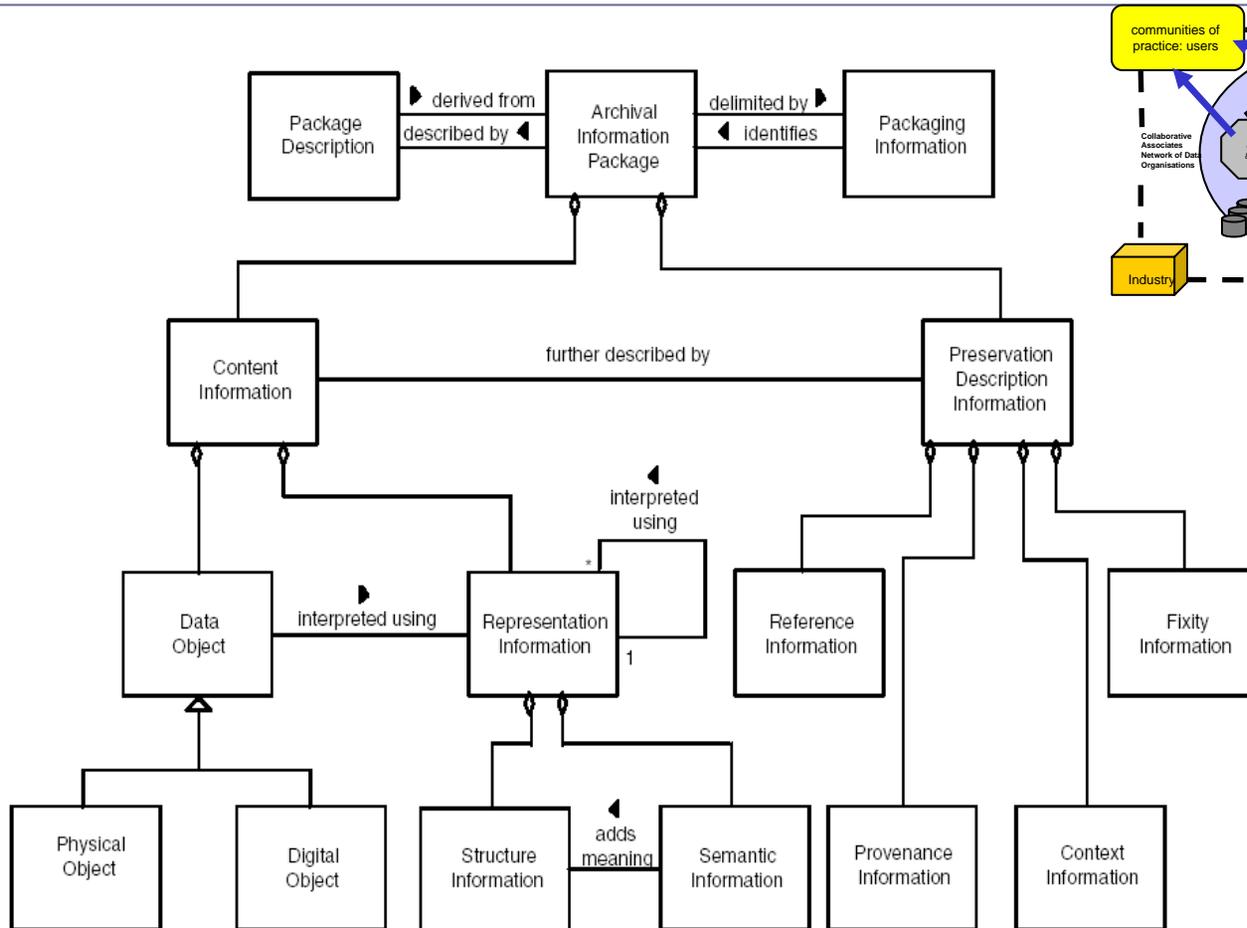# OAIS Reference Model – Functional Model
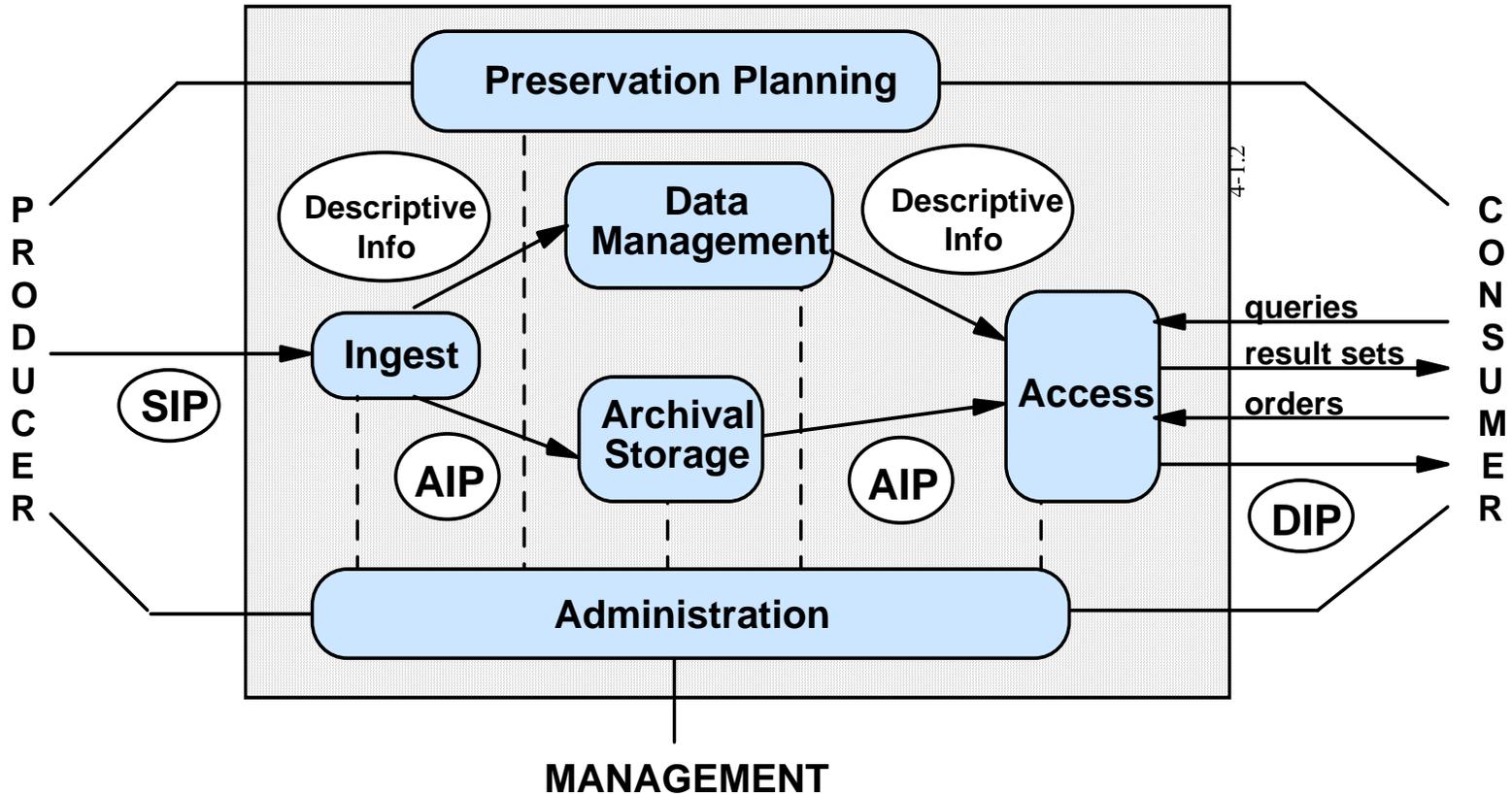
OAIS – **Preservation Planning** - key aspects

- Designated Communities & Knowledge Base
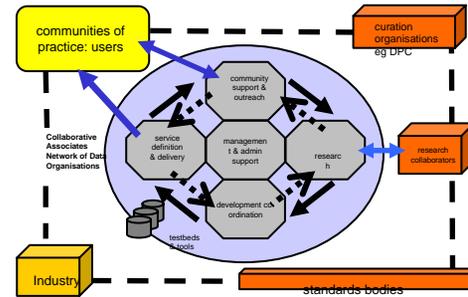- Representation Net

D|C|C

Representation Net
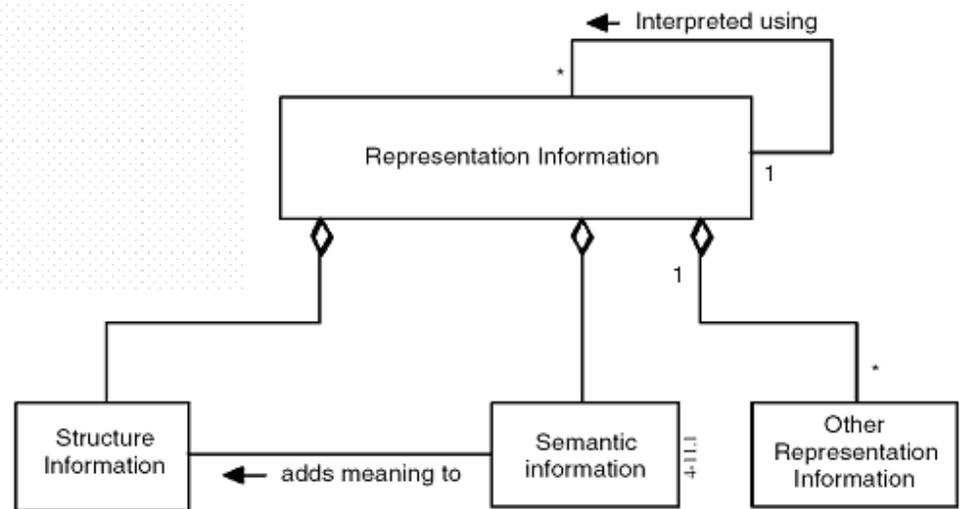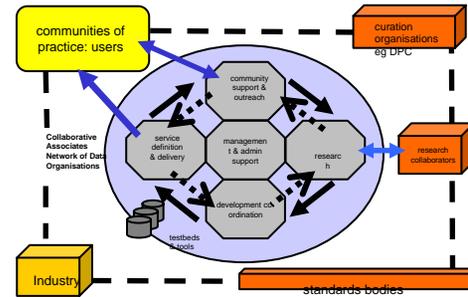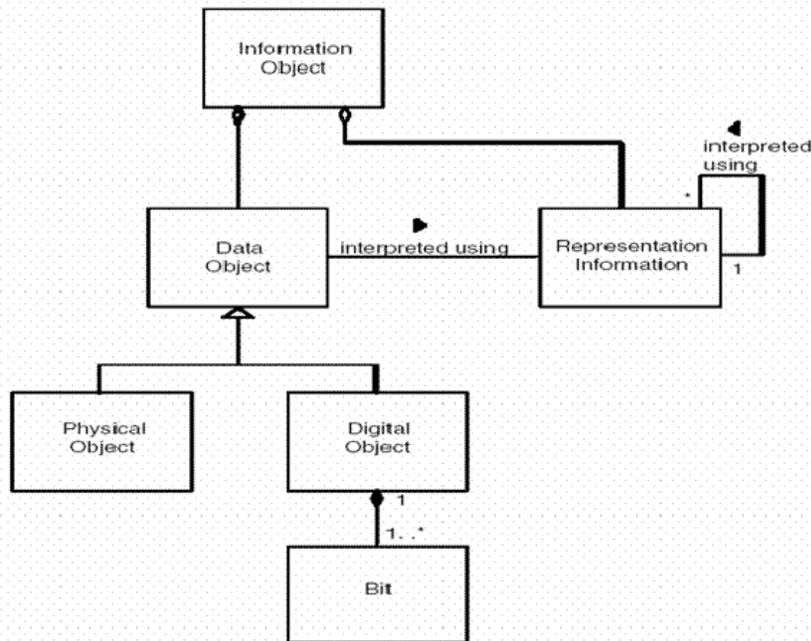
Archival Information Package

Testbeds

- Hardware used by "curators" in the wild
  - Details from projects
- Hardware suppliers
- Software suppliers
 - Commercial
  - Non-commercial

Standards and Audit & Certification

- How can people know to whom their information can be entrusted?
- OAIS follow-on standard(s) underway
  - on which a certification program can be based
- From the standards:
  - need to establish accreditation and certification bodies in preparation for offering audit and certification services
  - audit, certification and accreditation are potential sources of long term funding for the DCC
  - Testbeds and testing procedures
    - for software certification
    - hardware and software systems will need to be purchased, hired or borrowed.
      - we expect to work with hardware and software manufacturers to certify hardware and software components

D|C|C
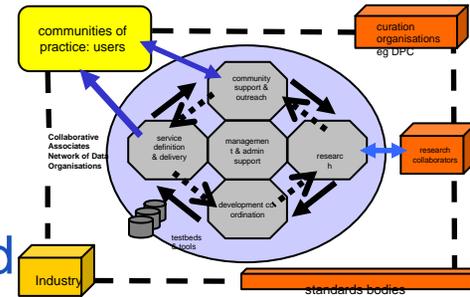
Working with Others

- Digital Library Federation
- The National Archives
- Global Grid Forum
- NARA
- Library of Congress
- Research Library Group
- Digital Preservation Coalition
- JISC community
- E-Science Community
- Associates Network
- …and many more
- InterPARES(?)

Development info – see

http://dev.dcc.rl.ac.uk/twiki/bin/view
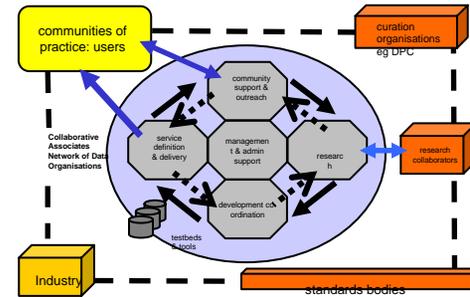
for details of Wiki and email list open to all