CARMEN and the human infrastructure of research data sharing

Graham Pryor
eScience Liaison
Digital Curation Centre, Edinburgh

Repository Curation Service Environments Workshop, 1st December 2008

Agenda

- Introduction to CARMEN and the DCC study
- Key elements of the CARMEN solution
- Principal observations and conclusions
  - Metadata
  - Data sharing
  - General conclusions of study

DCC eScience Liaison

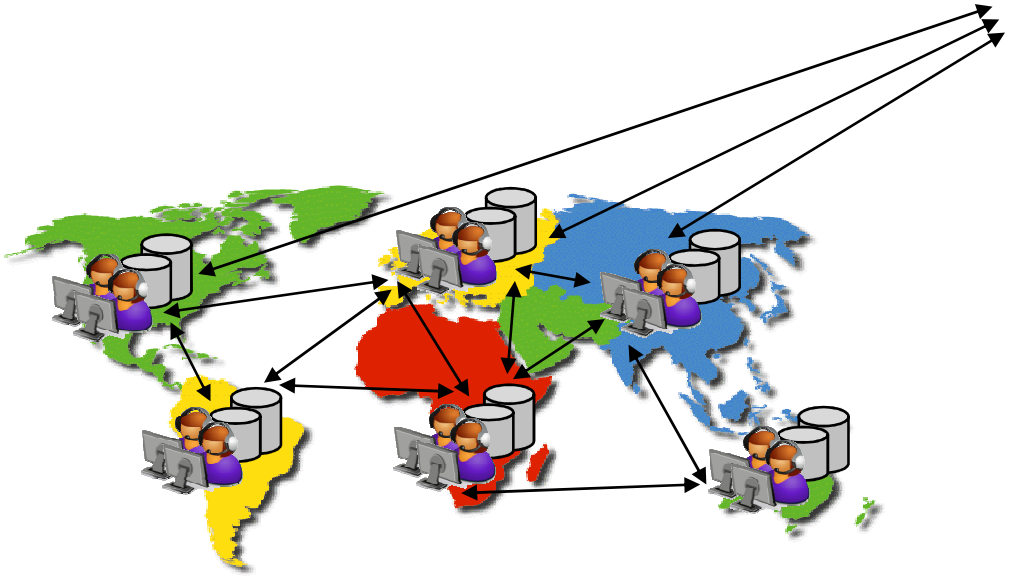
- Understand, promote and support the data needs of researchers
- Build the curation/eScience /research community
- Studies of eScience projects to identify innovative solutions/ good practice
- Collaborative workshops
- Research Data Management Forum
- Investigative projects – e.g.
  - OECD data sharing infrastructure
  - Case studies in the life sciences

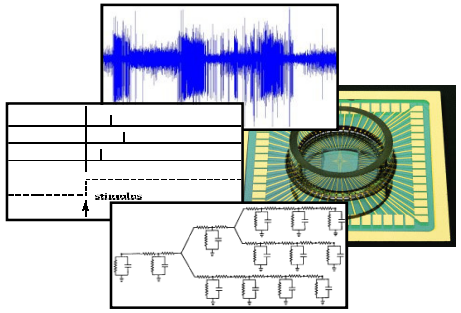
CARMEN

- £4.5M, 4 year eScience pilot project, commenced 1st October 2006
- 20 academic investigators, 11 UK institutions, plus commercial associates and international observers
- Aims to deliver new and more effective practices in the conduct of neurophysiological research
- Primary objective: the introduction of technology-enabled methods for sharing experimental data

CARMEN – http://www.carmen.org.uk/

Enabling sharing and collaborative exploitation of data, analysis code and expertise that are not physically collocated





Source: CARMEN SFN

The DCC longitudinal study

- DCC objectives
  - To understand the data curation requirements of the eScience community
  - To promulgate good curation practice and proven solutions, and to orientate DCC tools around real user requirements
- DCC study objectives
  - How effectively CARMEN integrates heterogeneous data
  - The nature of protocols and services for managing access
  - Mechanisms for the assignment of appropriate metadata
  - Legal compliance across a dispersed and diverse community
  - Analysis of the informatician / researcher partnership

The DCC longitudinal study

- Method
  - a series of observations over twelve months, based upon
  - participation in consortium meetings;
  - meetings with key project team members;
  - interviews with experimental neuroscientists;
  - the CARMEN Web, planning documents, email and blog.
- Leading to
  - an organic representation of project evolution and solutions;
  - a record of attitudes, needs, processes and relationships;
  - an analysis of the impact these human factors may have on data curation.

Rationale and reality

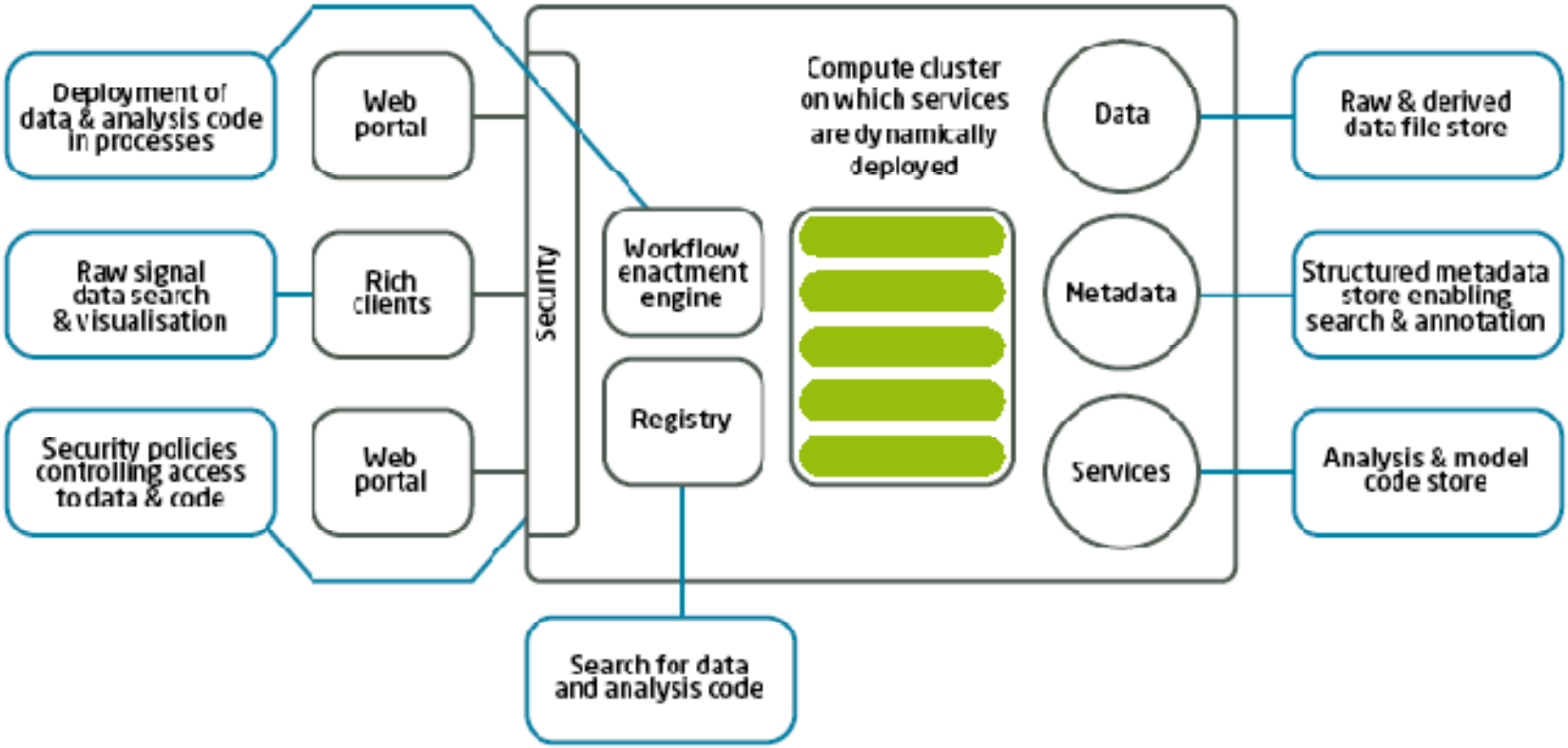
- CARMEN an eScience not a digital curation project
- Driven by the purpose to make experimental data available for further modelling and exploration
- Yet it is dependent upon
  - appropriate metadata to enable meaningful access;
  - a common structure for archiving datasets;
  - optimal integration and re-use of data;
  - diligent and long term curation of data and tools.

Tangible deliverables

- Federated data environment accessed via single interface (Web portal)
- Repositories (CAIRN[1]) containing data and metadata
- Identical infrastructure but data not mirrored
- A suite of neuroinformatics services that can be executed on data within the federation
- A managed environment for the long term storage of data and services
- Data preservation; metadata evolution (reflecting the science)

[1] CARMEN Active Information Repository Node

CARMEN system schematic



Metadata – a common cause

Urgent and universal drive to

- Bridge the disparate 'dialects' in neurophysiology
- Create or adopt an ontology to sustain the shared and consistent understanding of terms used
- Achieve an optimal fit with working practices in experimental neuroscience (community involvement)
- Use a metadata architecture that supports
  - the combination of different datasets
  - discovery, interrogation and access via any legitimate route
  - uninhibited data submission or use
- Employ metadata to assist data interpretation and select appropriate analysis services
- Preserve data whilst allowing metadata to evolve (accommodating new scientific method)

Framework for a common language

- Compliance with criteria for minimum metadata set as fundamental rule (otherwise no data upload)
- Build on existing technologies and standards
- Represent critical elements of the experimental process
  - Familiar Web services interface
  - Employ design principles of MIAPE guidelines[1]
  - Use of MINI[2] as technology independent checklist of essential information (no data format or repository structure identified)
  - Modules structured using the FuGE[3] data model

[1] Minimum Information About a Proteomics Experiment – Taylor et al
[2] Minimum Information about a Neuroscience Investigation
[3] Functional Genomics model – http://fuge.sourceforge.net

Metadata - dissonance

- MINI-inspired metadata proforma '*excruciatingly tedious*' to complete; too many fields not relevant '*to me*'
- Neurophysiological data infamously heterogeneous
- A truly generic minimum metadata architecture satisfies no-one
- Where experimental parameters frequently change there is resistance to a consistent set of metadata
- Need for much greater flexibility – but to enable the *addition* of terms!
- Threat to data ownership and research integrity

Data security *and* data sharing

- Social networking concept of groups an attempt to build on recognised layers of trust

Researcher/Supervisor (trust)

Collocated Researchers (may be trust)

Networking (new trust)

Data security *versus* data sharing

- Domain norms underpinning sharing are directly related to the cost of data acquisition (financial and intellectual)
- Highly individualised research
- High perceived value (including commercial)
- Culture of trust predicated by long 'courtships' before resources (data and code) are shared
- Reluctance even to share metadata where this provides clues to new research or experimental methods
- View of RC data sharing edicts as naïve and remote from reality of competitive research environment

Data security *versus* data sharing

- Data security regarded as pre-eminent
- Data 'owners' to exercise full control over data sharing
- Data released to CARMEN members contingent upon publication (potentially long delays)
- Expectation of being directly involved in further analysis of data sets (managed collaborations)
- Tradition of closed networks will not accept imposition of cultural change
- CARMEN data sharing strategy seeking incentives for early release

Six selected conclusions from study

- Familiar practices/tools will alleviate distrust when IT literacy is inconsistent
- National data management strategies are unlikely to succeed without restructuring in funding/support to informatics-dependent initiatives
- As science is provisional it is crucial to adopt routines that preserve the integrity of the data whilst facilitating the evolution of the metadata
- Be candid: the assignment of metadata requires discipline knowledge and is likely to prove burdensome to researchers
- The imposition of a rate of change designed only to effect new political, economic or technological strategies could prove intolerable
- Researchers focus upon data to enable scientific endeavour, not to engage with curation issues (where the benefits have still to be demonstrated)