Infrastructures for the use and reuse of research data

Seminario Consorcio Madroño 17th November 2008

Dr Celia Russell

Economic and Social Data Service and University of Manchester

The University of Manchester

Data - lifeblood of research

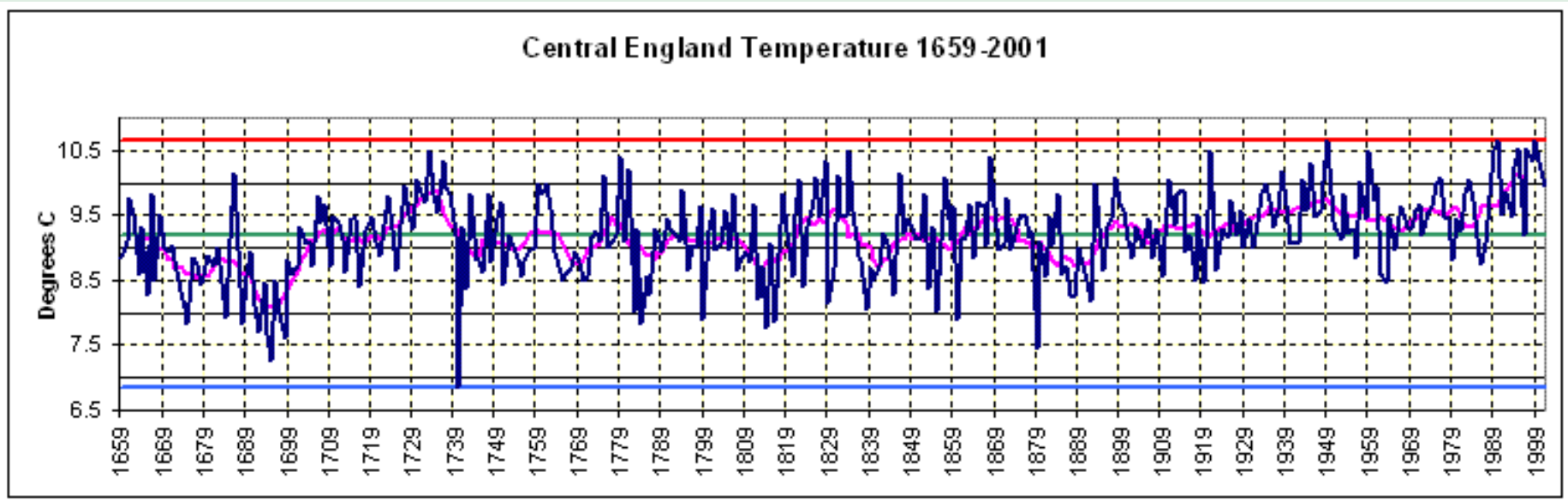
esds
Economic and Social Data Service

- Public bodies have invested considerable resources into producing high quality research data
- Value for secondary use is considerable
- Internet technologies have increased the research value of data and allow:
  - more equal access to data
  - improved quality and quantity of research
  - greater development of data handling capacities
- Signatories to the OECD's *Declaration on Access to Research Data from Public Funding*
- Store for future generations

The University of Manchester

Growing wealth through the generations

esds
Economic and Social Data Service



Levels of data infrastructures

- Faculty or individual level
- Institutional repositories e.g. Manchester University
- National, discipline specific e.g. Economic and Social Data Service
- National, big data e.g. National Grid Services
- International infrastructures e.g. CERN and EGEE

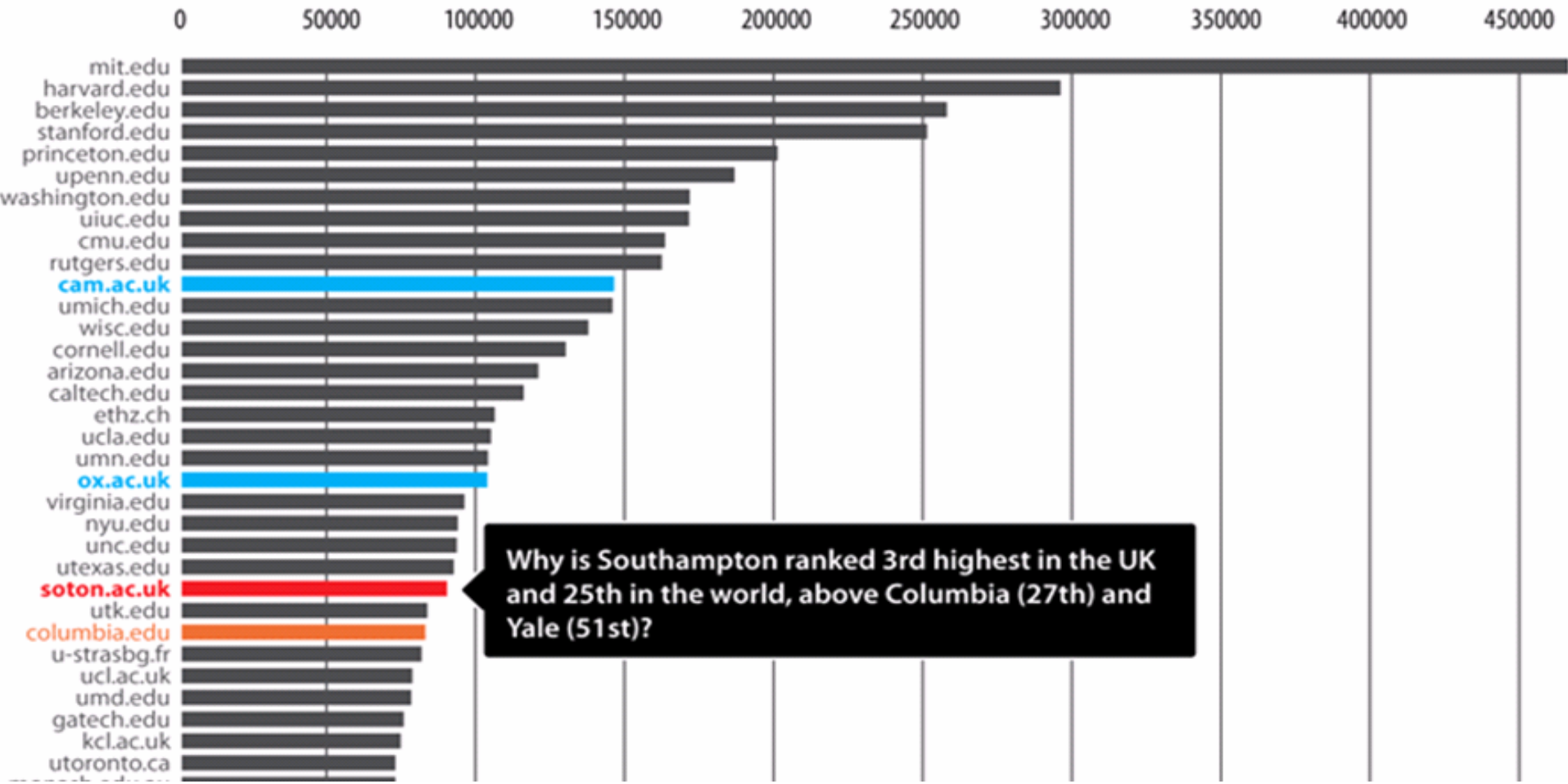
Institutional repositories

- Example Manchester University
- Goal of project:
  - To sustain and enhance the research representations of individual and organization affiliated with the University of Manchester
- For all research outputs including publications and data

The University of Manchester

g-factor rankings

esds
Economic and Social Data Service



Current situation

- 10-100 research outputs daily
- Citations per output below average
- 70,000 records currently held at faculty level
- Currently has two repositories, both with low uptake by researchers
  - Dspace repository mediated by the University library
  - Grid based repository developed by IT services

The University of Manchester

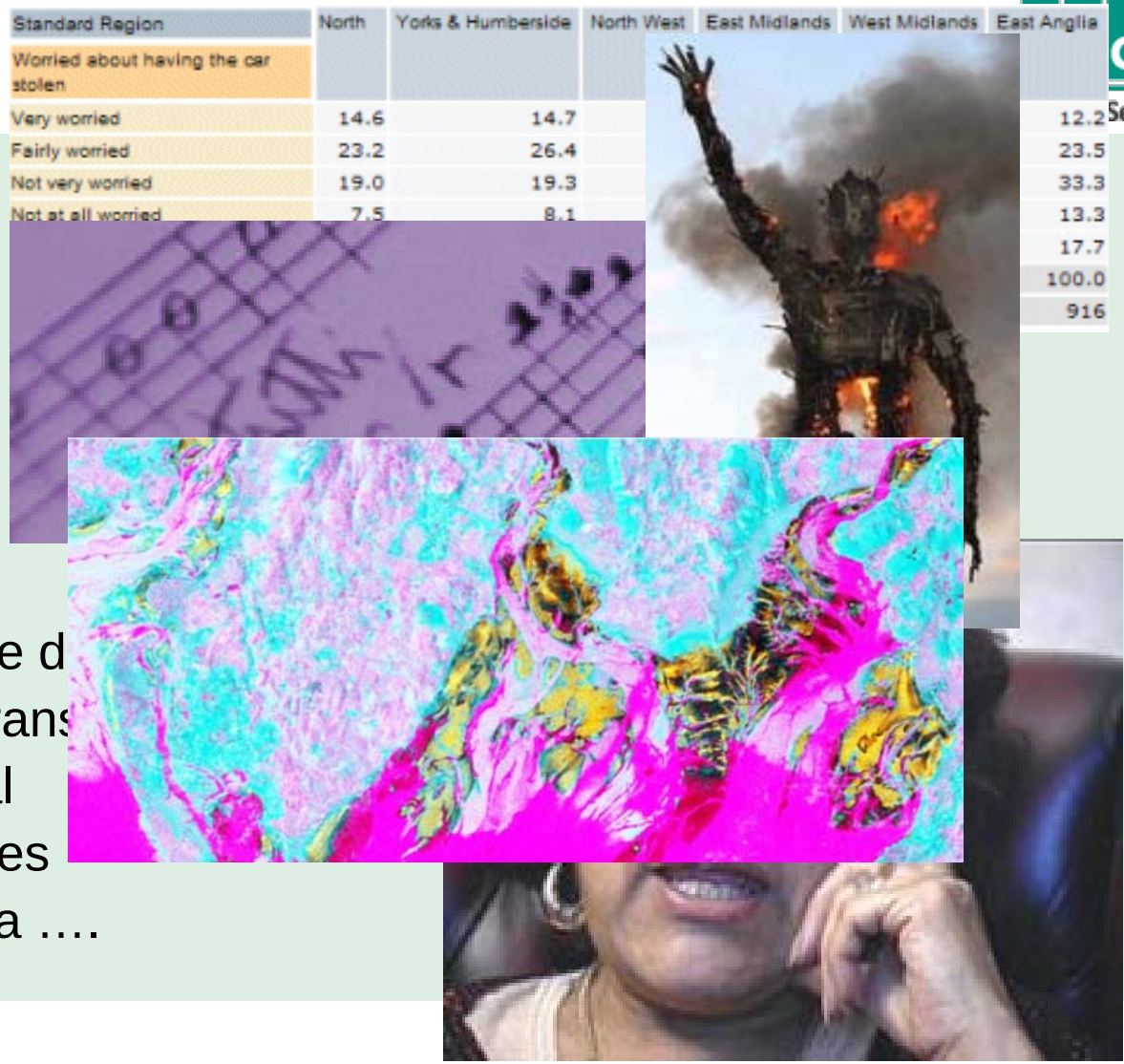
New repository

esds
Economic and Social Data Service

- Developed very closely with academic staff
- Faculty led at professorial level
- Depositing mandatory
- Will capture > 95% institutional research output
- Near zero barriers for content submission
- Supports **32 different content types**

Supported content

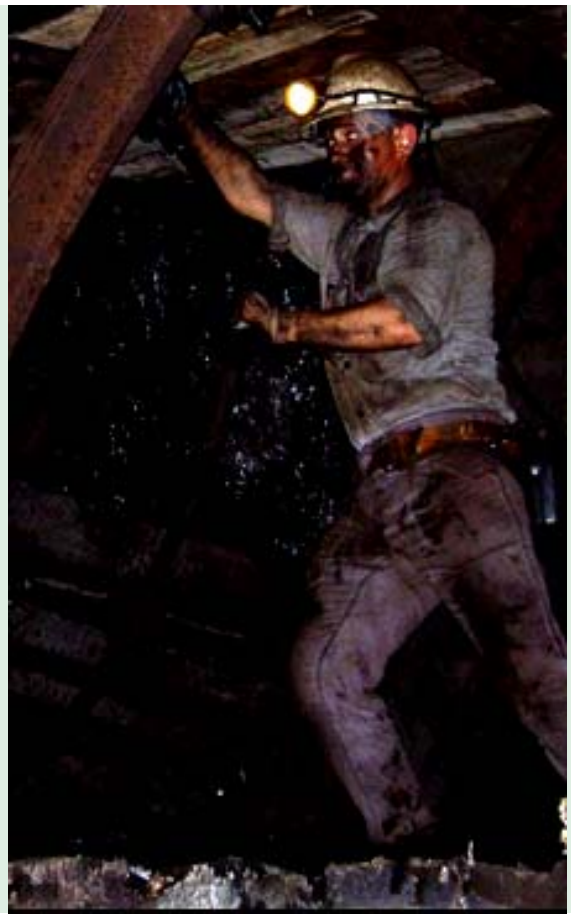
- Publications
- Software
- Grey literature
- Presentations
- Dissertations
- Data including:
  - Quantitative d
  - Interview trans
  - Audiovisual
  - Music scores
  - Spatial data ….



Standard Region	North	Yorks & Humberside	North West	East Midlands	West Midlands	East Anglia
Worried about having the car stolen						
Very worried	14.6	14.7				12.2
Fairly worried	23.2	26.4				23.5
Not very worried	19.0	19.3				33.3
Not at all worried	7.5	8.1				13.3
						17.7
						100.0
						916

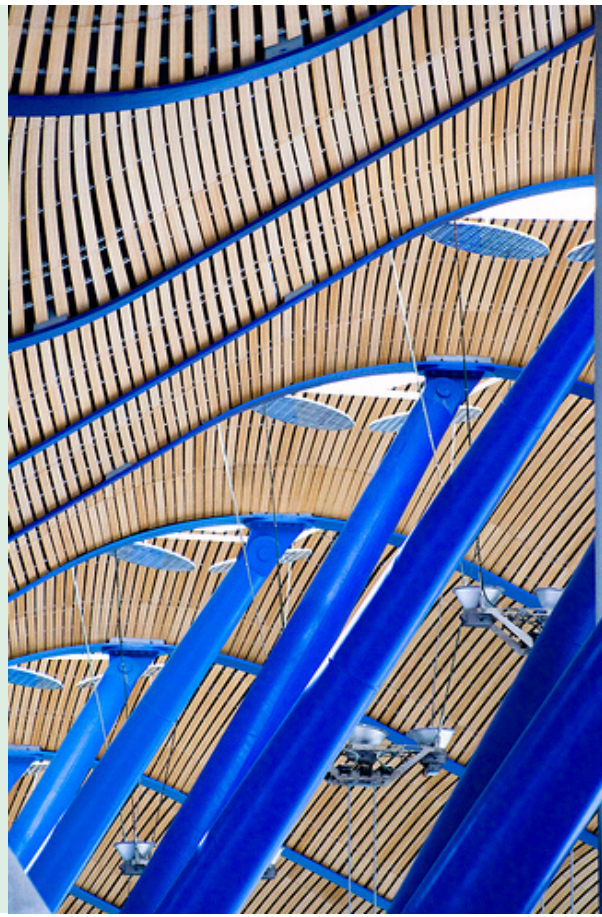
Building the repository

- Set up over 2 years
- Initial £500,000 budget
- Includes set-up costs, hardware, 7.5 people - years
- Business case includes estimates of recurrent costs
- Funded by University rather than funding council



Architectural requirements

- Scalability - to petabyte levels
- Sustainability
- Sensitivity to future
- Flexibility
- Granularity
- Secure
- Robust



Software platforms

Open source:

- Eprints **Runs on perl**
- Dspace **Runs on java**
- Fedora **Runs on java**

University as a lot of local expertise with Java

Large enough to do own technical development

Fedora had the granularity and scalability required

Access and copyright

- Outputs are harnessed irrespective of copyright
- If subject to a restrictive copyright policy, output still deposited but embargoed
- Checking copyright is researcher's responsibility
- Researchers are taken step by step through process as part of depositing workflow
- Dark archive kept within living environment and inherits same preservation model as living content

The University of Manchester

Keeping it safe

esds
Economic and Social Data Service

- Dual sites in two buildings on campus
- Load balanced network
- Terabyte backup
- Mutual mirroring service with another university

National Data Infrastructures

- Tend to be discipline based
- Research council funded
- Most based on grant horizons of 2-5 years
- But some have been in place for decades

Example: ESDS

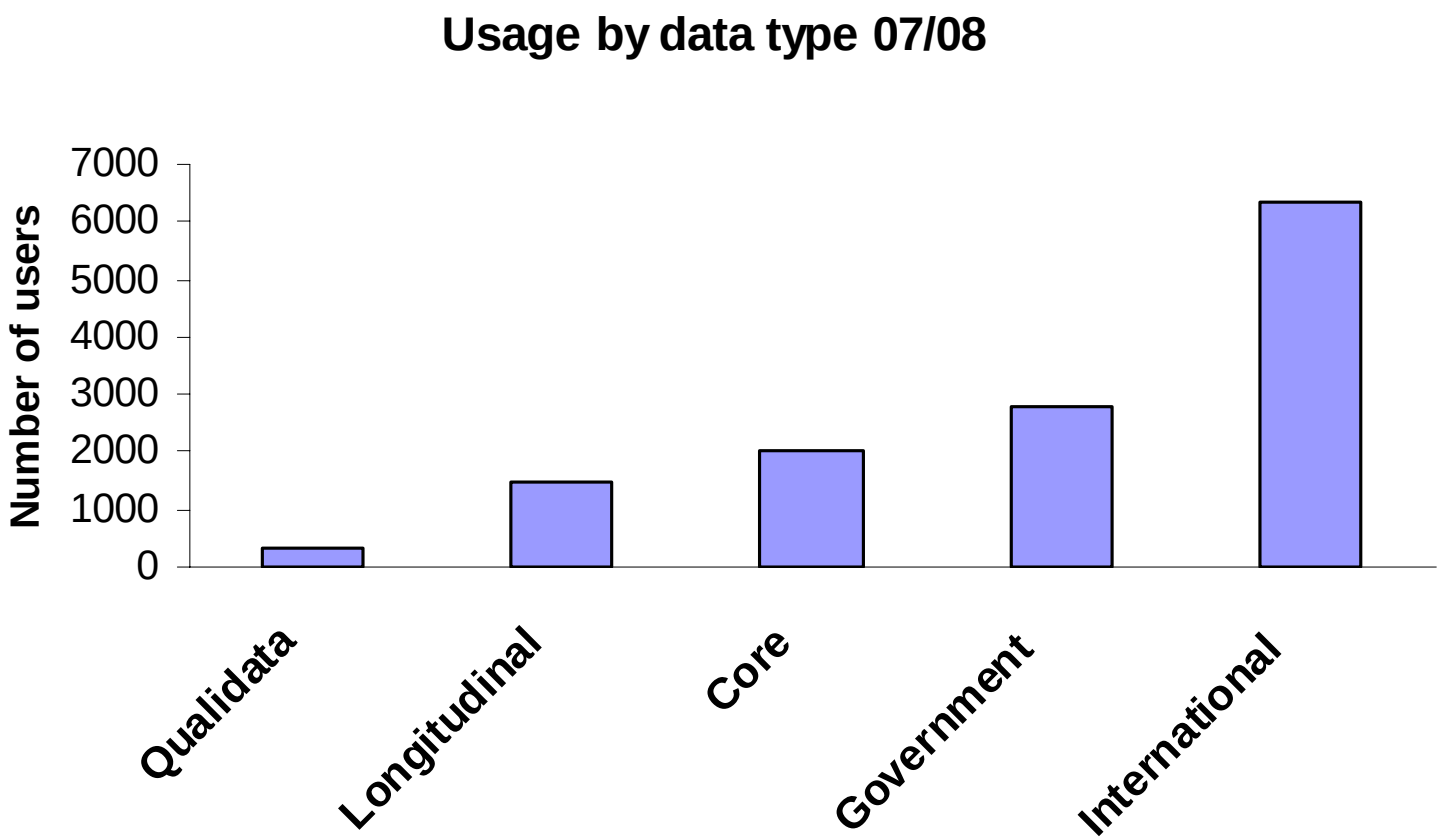
Economic and Social Data Service

- Economic and Social Data Service
- Principal national data service for UK social science data
- Distributed service
- All academic institutions have equal access
- Funded until 2013
- In 2009, will write funding requirements for 2013-2018

ESDS – specialist data services

- Split into 5 specialist services by data type
  - Core (researcher deposited data)
  - Qualitative data
  - Longitudinal data
  - Government data
  - International data

Usage by data type



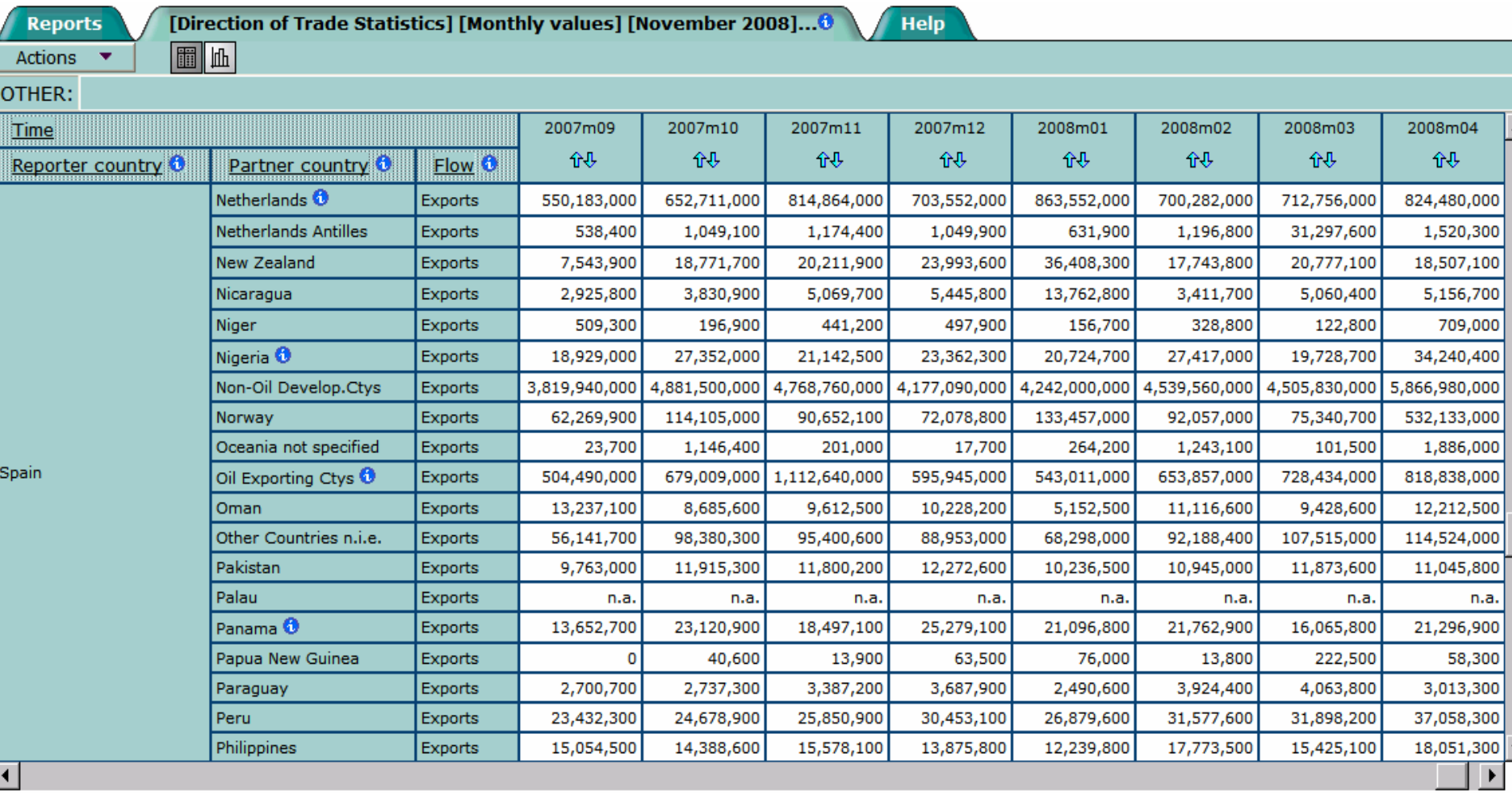
Data for research

- Research data comes from many sources
- Data produced by non-academic organisations
  - Cohort studies
  - Government survey and national laboratory data
  - global databanks produced by intergovernmental organisations
- These kinds of data extremely widely used in research

ESDS International

- national licensing agreements
- data free at the point of use
- single web interface
- effective dissemination and outreach
- updated latest releases
- access by federated access
- common user interface
- created a new community of users

Common user interface



Reports | [Direction of Trade Statistics] [Monthly values] [November 2008]... | Help

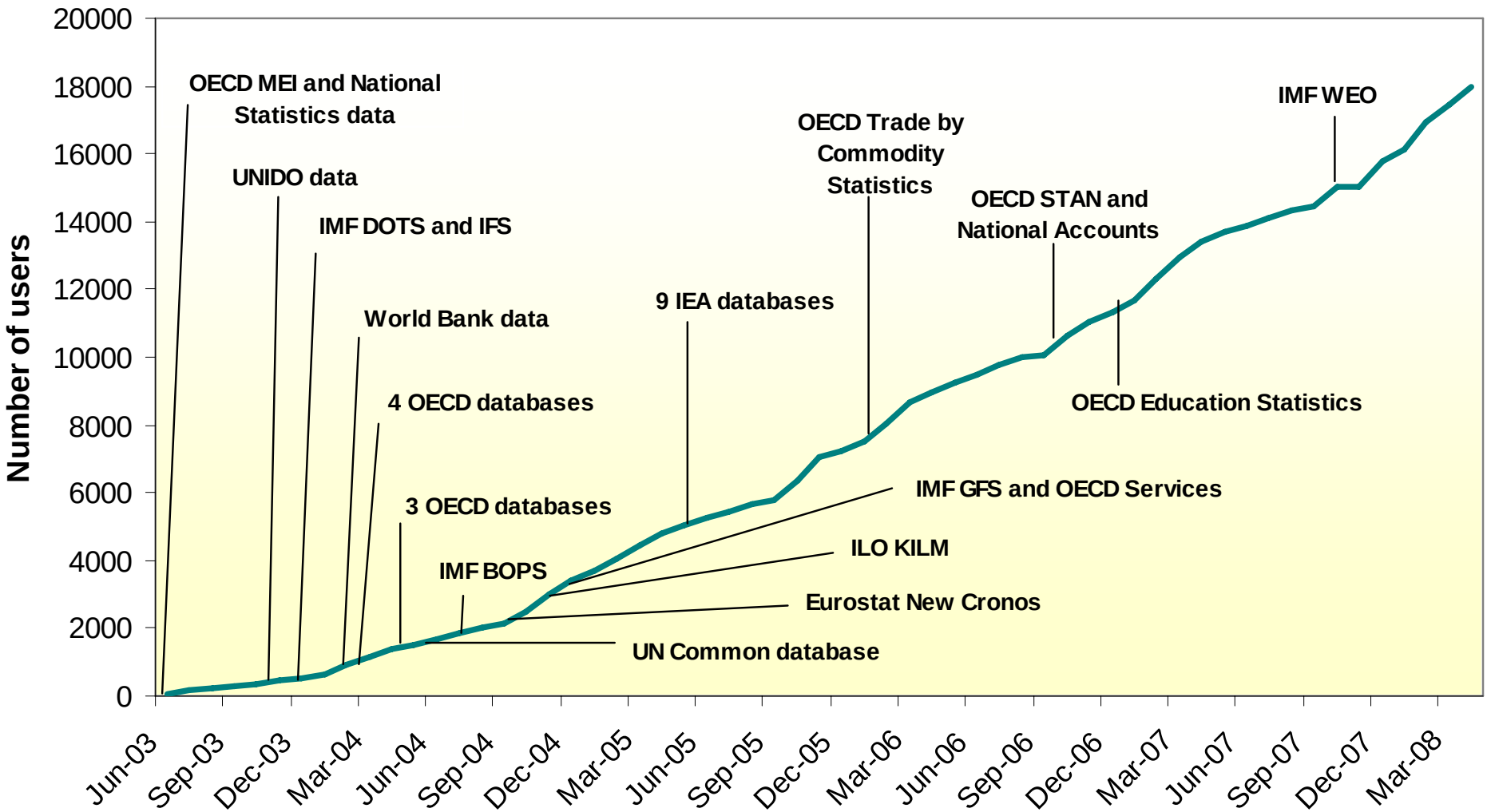
Actions

OTHER:

Time / Reporter country	Partner country	Flow	2007m09	2007m10	2007m11	2007m12	2008m01	2008m02	2008m03	2008m04
Spain	Netherlands	Exports	550,183,000	652,711,000	814,864,000	703,552,000	863,552,000	700,282,000	712,756,000	824,480,000
	Netherlands Antilles	Exports	538,400	1,049,100	1,174,400	1,049,900	631,900	1,196,800	31,297,600	1,520,300
	New Zealand	Exports	7,543,900	18,771,700	20,211,900	23,993,600	36,408,300	17,743,800	20,777,100	18,507,100
	Nicaragua	Exports	2,925,800	3,830,900	5,069,700	5,445,800	13,762,800	3,411,700	5,060,400	5,156,700
	Niger	Exports	509,300	196,900	441,200	497,900	156,700	328,800	122,800	709,000
	Nigeria	Exports	18,929,000	27,352,000	21,142,500	23,362,300	20,724,700	27,417,000	19,728,700	34,240,400
	Non-Oil Develop.Ctys	Exports	3,819,940,000	4,881,500,000	4,768,760,000	4,177,090,000	4,242,000,000	4,539,560,000	4,505,830,000	5,866,980,000
	Norway	Exports	62,269,900	114,105,000	90,652,100	72,078,800	133,457,000	92,057,000	75,340,700	532,133,000
	Oceania not specified	Exports	23,700	1,146,400	201,000	17,700	264,200	1,243,100	101,500	1,886,000
	Oil Exporting Ctys	Exports	504,490,000	679,009,000	1,112,640,000	595,945,000	543,011,000	653,857,000	728,434,000	818,838,000
	Oman	Exports	13,237,100	8,685,600	9,612,500	10,228,200	5,152,500	11,116,600	9,428,600	12,212,500
	Other Countries n.i.e.	Exports	56,141,700	98,380,300	95,400,600	88,953,000	68,298,000	92,188,400	107,515,000	114,524,000
	Pakistan	Exports	9,763,000	11,915,300	11,800,200	12,272,600	10,236,500	10,945,000	11,873,600	11,045,800
	Palau	Exports	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Panama	Exports	13,652,700	23,120,900	18,497,100	25,279,100	21,096,800	21,762,900	16,065,800	21,296,900
	Papua New Guinea	Exports	0	40,600	13,900	63,500	76,000	13,800	222,500	58,300
	Paraguay	Exports	2,700,700	2,737,300	3,387,200	3,687,900	2,490,600	3,924,400	4,063,800	3,013,300
	Peru	Exports	23,432,300	24,678,900	25,850,900	30,453,100	26,879,600	31,577,600	31,898,200	37,058,300
	Philippines	Exports	15,054,500	14,388,600	15,578,100	13,875,800	12,239,800	17,773,500	15,425,100	18,051,300

Building an International Data Community
Unique users and dataset release dates

Source: Athens usage statistics April 2008



Running the service

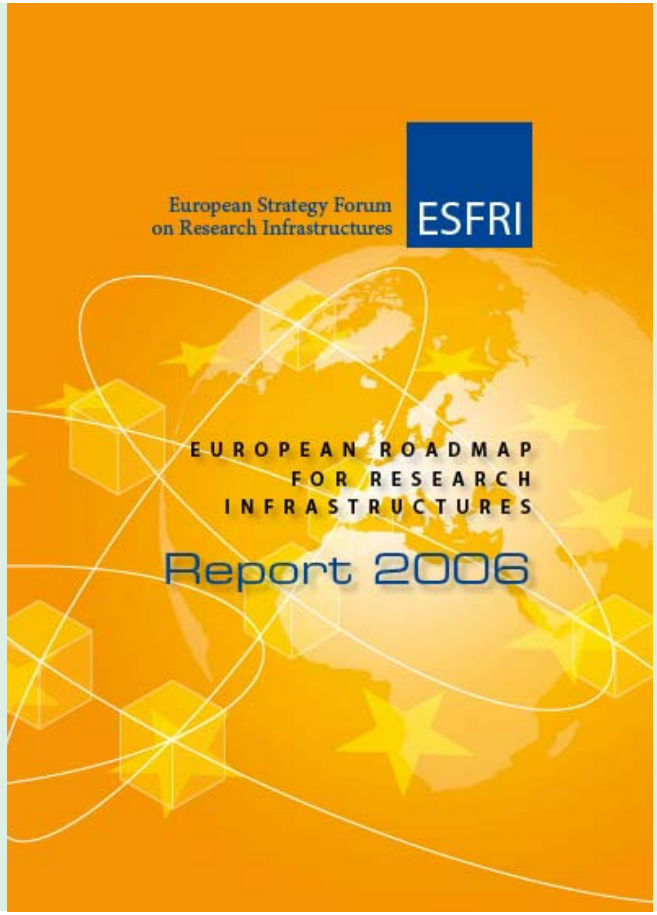
- Service used by 18,000 researchers
- Run about 100,000 data analysis sessions a year
- Runs on 2 servers – networked balanced load
- Main cost national licensing agreements
- From 2007 onwards, 4.2 staff
- Large economies of scale
- Don’t need many people to run a national service!

Other national data initiative examples

- US Datanet
  - Funded by NSF
  - Will create 5 research data networks
  - $100 million over next 5 years
  - University consortia
  - Self financing in the long term
- Australian ANDS
  - Developing frameworks
  - Providing utilities e.g. umbrella services
  - Seeding the data commons
  - Building capabilities
  - Looking to collaborate internationally

European infrastructures

- Significant work at EU level in supporting research data infrastructures
- Discipline based
- ESFRI roadmap identifies 35 key Research Infrastructure projects



Example ESFRI project: Cessda

Big data

- Many projects worldwide now generate terabytes of data per year
- Leading edge data storage and handling
- Big data has special requirements
  - Unmanageable in web environment
  - Collaborative working
  - Complex analyses
  - Visualisation
  - Usually fits within well defined schemas

Looking after big data

Grid infrastructures:

- host large and complex data resources
- include the computing resources to handle and analyse data
- distributed federated systems
- heterogeneous environments
- appear as single system to end user
- single sign on – grid certificate
- shaped by user community
- large public investment

The University of Manchester

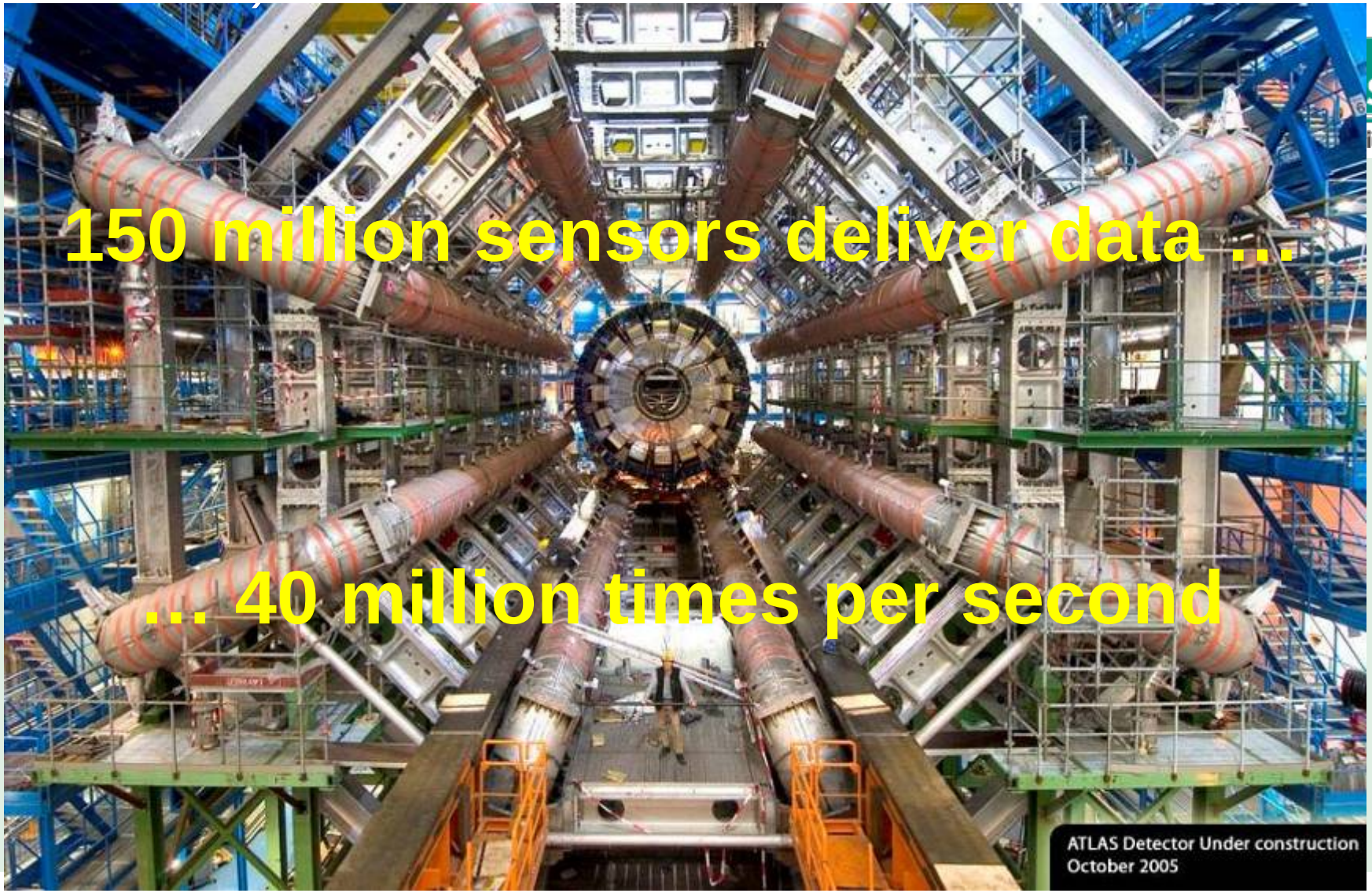
National Grid Services

esds
Economic and Social Data Service

- National grid services host big databases too large to host at local institution (around 50MB up to 5TB)
- Usually free to use, light peer review
- Multi-disciplinary

Why are we talking about CERN?

- CERN has a long history in driving research data infrastructures
- The Large Hadron Collider Computing Grid formed the basis of a wider global grid infrastructure
- This now attracts users from many disciplines


150 million sensors deliver data …
… 40 million times per second
ATLAS Detector Under construction
October 2005

LHC Data Grid Hierarchy:

Experiment

~PByte/sec

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1

Online System

~200-1500 MBytes/sec

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

Tier 1

10 - 40 Gbps

IN2P3 Center

RAL Center

INFN Center

FNAL Center

~10 G

Tier 2

Tier2 Center

1-10 Gbps

Tier 3

Physics data cache

Institute

1 to 10 Gbps

Workstations

Tier 4

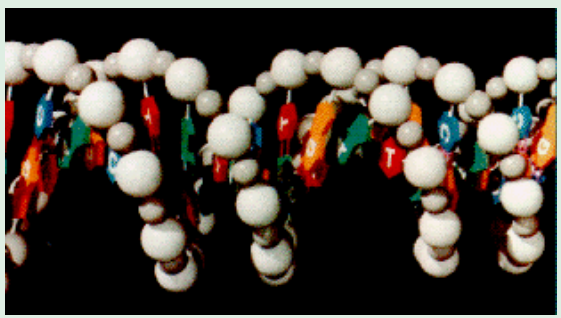
11 tier 1 sites
PIC at Barcelona

~ 140 tier 2 sites
e.g. Manchester

Tens of Petabytes by
An Exabyte ~5-7 Years later.

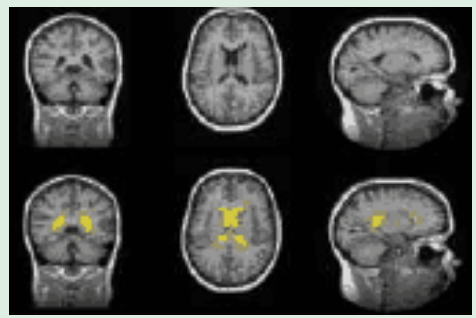
Impact of the LHC Computing Grid in Europe

- LCG has been the driving force for the European multi-science Grid EGEE (Enabling Grids for E-sciencE)
- EGEE is now a global effort, and the largest Grid infrastructure worldwide
- Co-funded by the European Commission (Cost: ~130 M€ over 4 years, funded by EU ~70M€)
- EGEE already used for >20 applications, including…



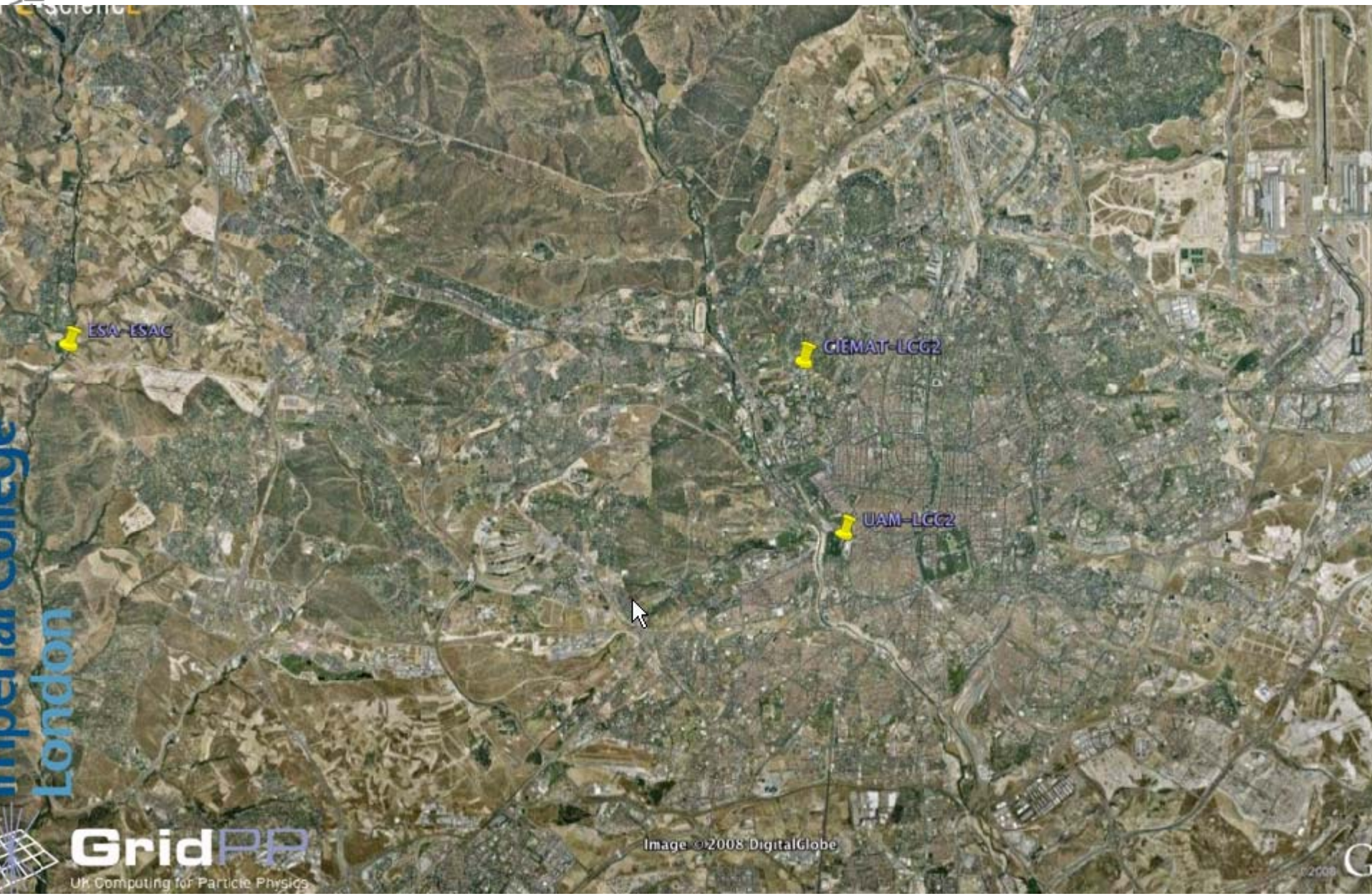
Bio-informatics

Earth Sciences



Medical Imaging

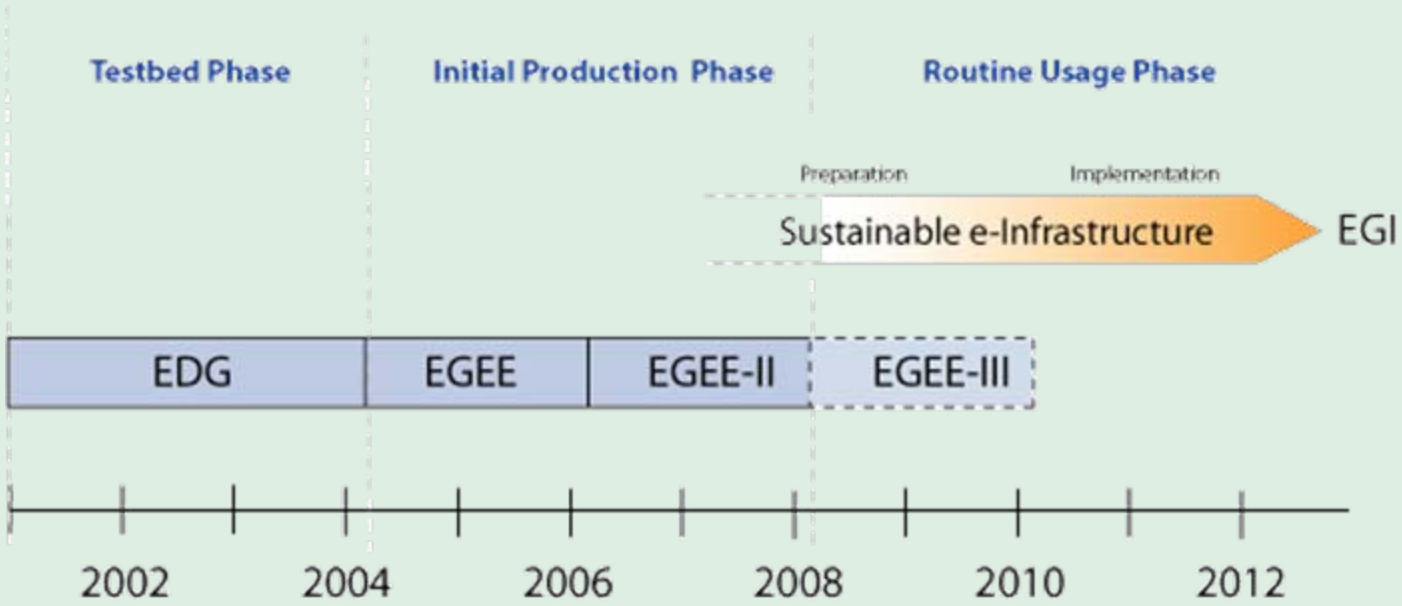

nabling Grids
r E-sciencE
London
GridPP
UK Computing for Particle Physics
Image © 2008 TerraMetrics
Image NASA
© 2008 Europa Technologies
© 2008 Tele Atlas
Google


ESA-ESAC
CIEMAT-LCG2
UAM-LCG2
Image ©2008 DigitalGlobe
GridPP
UK Computing for Particle Physics
Imperial College London

The University of Manchester

Time frames

esds
Economic and Social Data Service



Future Sustainability

- Need to prepare for permanent Grid infrastructure
- Ensure a high quality of service for all user communities
- Independent of short project funding cycles
- Infrastructure managed in collaboration with National Grid Initiatives (NGIs)
- European Grid Initiative (EGI)
- Is there an cloud alternative?

The University of Manchester

esds
Economic and Social Data Service

Gracias a:

- Dave Bailey, University of Manchester
- Roger Barlow, University of Manchester, CERN and Stanford
- Phil Butler, University of Manchester
- Paul Murphy, University of Manchester
- Kier Hawker, Rutherford Appleton Laboratory