Edinburgh e-Science TOE

Training Outreach and Education

http://www.nesc.ac.uk/training

NGS National Grid Service

http://www.ngs.ac.uk

Data and storage services on the NGS

egee

http://www.eu-egee.org/

Policy for re-use

- This presentation can be re-used for academic purposes.

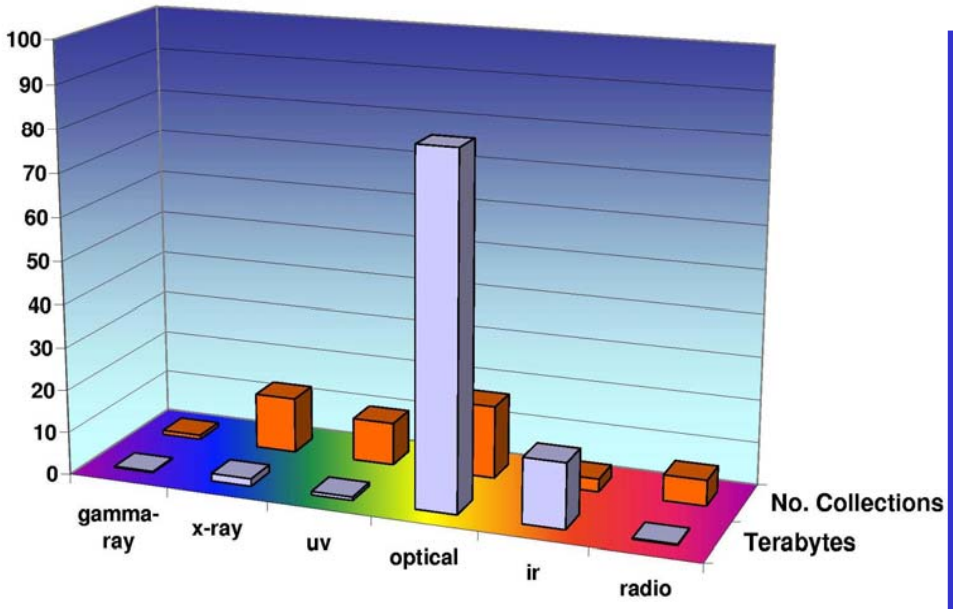
- However if you do so then please let training-support@nesc.ac.uk know. We need to gather statistics of re-use: no. of events, number of people trained. Thank you!!

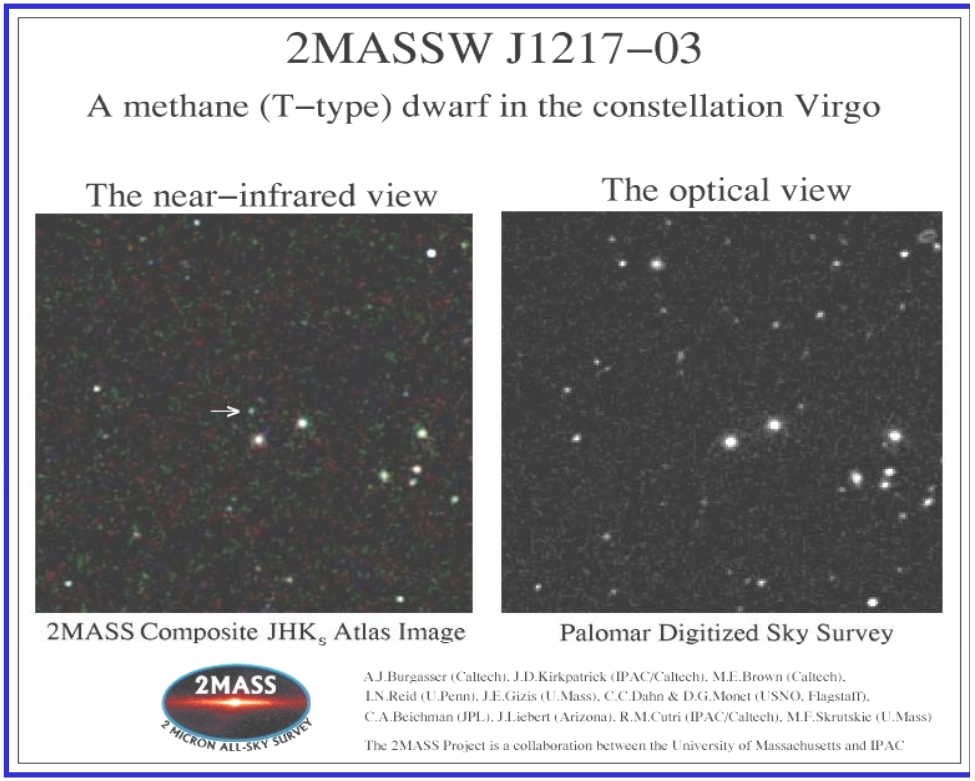
NGS
National Grid Service

Data and storage

- This talk: Background and NGS services
- Yesterday: how to run jobs
- **BUT**
  - Can't have computation without data!
- AND **BUT**
  - It's the data that drives research
  - Data → Information → Knowledge
- AND **YEAH BUT NO BUT**
  - Can't have digital data without computation!

Astronomy





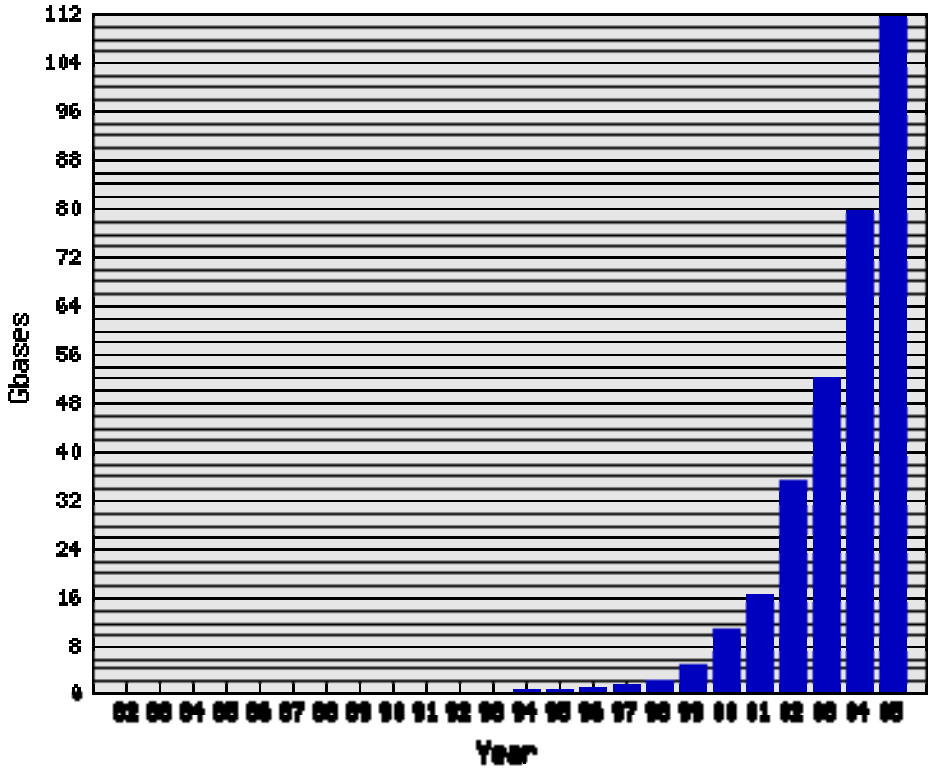
No. & sizes of data sets as of mid-2002, grouped by wavelength

- 12 waveband coverage of large areas of the sky
- Total about 200 TB data
- Doubling every 12 months
- Largest catalogues near 1Billion objects

Data and images courtesy Alex Szalay, John Hopkins University

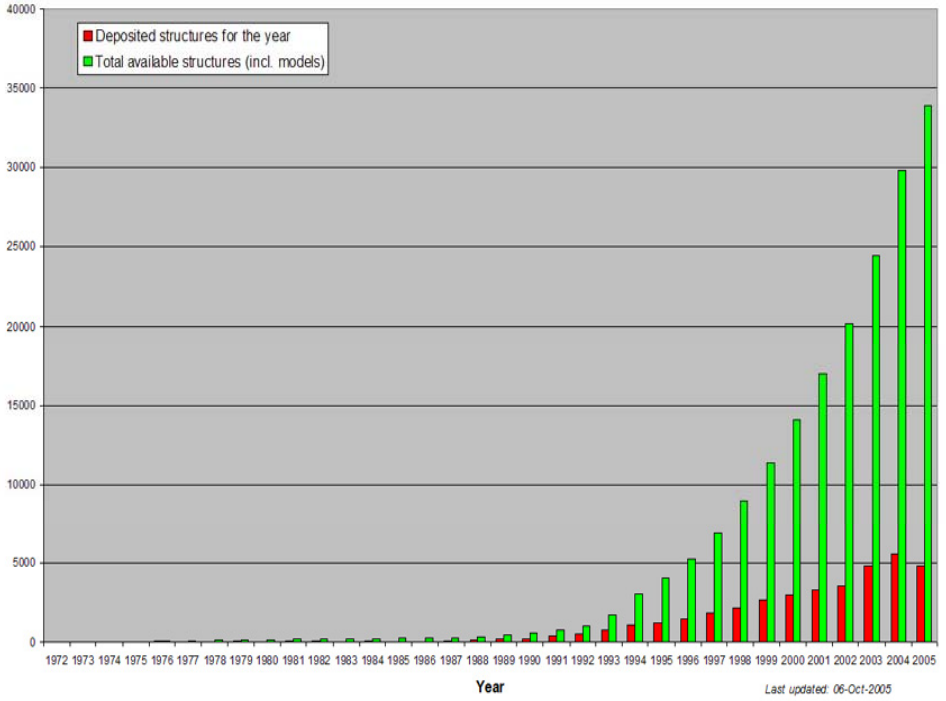
Biomedical

EMBL DB 111,416,302,701 nucleotides



EMBL Nucleotide Sequence Database

PDB 39,853 Protein structures



PDB - Protein Data Bank

NGS
National Grid Service

Data, Data Everywhere

- Entering an age of data: data explosion – growing volumes
  - CERN: LHC will generate 1GB/s = 10PB/y
- Data stored in many different ways – growing diversity
  - Relational databases
  - XML databases
  - Flat files
- From
  - Legacy datasets
  - Simulations
  - New measurements, sources, analyses….

NGS
National Grid Service

So how do we....

- Mine data riches for nuggets of information?
  - Discovery depends on insights
  - Unpredictable or unexpected use of data
- **Facilitate**
  - Data discovery
  - Data understanding
  - Data access
  - Data integration
- **Empower e-Business and e-Research**
- **A Grid is a vehicle for achieving this**

NGS
National Grid Service

A Grid and data

- Grid: diverse services sharing AuthN and AuthZ across admin domains, including:
  - Data storage
    - Controlled AA
  - Data transfer
    - Between stores, stores and compute nodes
  - Data catalogues, registries,…
    - How can I find the data in the resources that I can access?
- Key issue:
  - Move data to where computation will happen?
  - Move computation to be close to data?

Move computation to the data

- Assumption: code size << data size
  - Minimise data transport
- Provision combined storage & compute resources
- Develop the database philosophy for this?
  - Queries are programs safe to run near data
- Develop the storage architecture for this?
  - Computation hosted close to storage
- Develop experiment, sensor & simulation architectures
  - That take code to select and digest data as an output control
  - That attach the provenance & metadata

NGS
National Grid Service

Meta-data: describing data

- Choosing data sources
  - How do you find them?
  - How are they described and advertised?

- Meta-data is required describing
  - Content
  - Provenance
  - Structure
  - Types, Formats & Ontologies
  - Operations available
  - Access requirements
  - Quality of service

Necessary to find, understand, access, automate, assess quality and manage
But very hard to create & maintain: Incentives & tools

2 main types of data services on Grids

- **Simple data files on grid-specific storage**
- **Middleware supporting**
  - **Replica files**
    - to be close to where you want computation
    - For resilience
  - **Logical filenames**
  - **Catalogue**: maps logical name to physical storage device/file
  - **Virtual filesystems**, POSIX-like I/O
  - Services provided: storage, transfer, catalogue that maps logical filenames to replicas.
- **Solutions include**
  - **gLite data service**
  - **Globus: Data Replication Service**
  - **Storage Resource Broker**

- **Other data! e.g. ….**
  - Structured data: RDBMS, XML databases,…
  - Files on project's filesystems
  - Data that may already have other user communities not using a Grid
- Require extendable middleware tools to support
  - Computation near to data
  - Controlled exposure of data *without replication*
- Basis for integration and federation
- **OGSA – DAI**
  - In Globus 4
  - Not (yet...) in gLite

National Grid Service

Oracle and the NGS

- The requirement for data hosting will grow
- Oracle database: for both users and services offered by the NGS.

The NGS Oracle service is ready to host data for your project!

- Additional application needed after joining the NGS

Data & storage services on the NGS

- The Storage Resource Broker
- GridFTP: a protocol for large file transfer

- OGSA-DAI: data access and integration

- And also: ORACLE – available for data storage