

# The British Atmospheric Data Centre: Curation and Facilitation

Bryan Lawrence

NCAS/BADC

Rutherford Appleton Laboratory, CCLRC



- Introduction to the BADC & BADC data
- Issues for digital preservation:
  - Growth in users; growth in data volumes
  - Taxonomy of Metadata
    - Risk of WORN archives
  - Archive environment: producers and consumers
    - Important potential of e-science to break down barriers
- Summary

# The BADC (part of NCAS)!



The screenshot shows the BADC website with a blue header and a dark blue main content area. The header contains navigation tabs: Home, My BADC, Data, Search, Community, and Help. Below the tabs are links for 'More about the BADC', 'Login', and 'New User Registration'. The main content area features the BADC logo (a blue and white swirl) and the title 'BRITISH ATMOSPHERIC DATA CENTRE' in large white letters. To the right of the title is a colorful globe icon. Below the title, a paragraph describes the BADC's role as the NERC's Designated Data Centre for Atmospheric Sciences. A 'Latest News' section highlights 'BADC Launches great new look!'. The footer includes the NCAS logo, a description of the BADC's role within NCAS, and a 'Last Modified' timestamp.

Home My BADC Data Search Community Help

More about the BADC Login New User Registration

**BRITISH ATMOSPHERIC DATA CENTRE**

The British Atmospheric Data Centre (BADC) is the Natural Environment Research Council's (NERC) Designated Data Centre for the Atmospheric Sciences. The role of the BADC is to assist UK atmospheric researchers to locate, access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by NERC projects.

To find out more about the BADC, please follow the links on the sub-tab sections above.

**Latest News:** [BADC Launches great new look!](#)

**NCAS**  
NATURAL  
ENVIRONMENT  
RESEARCH COUNCIL

The BADC is one of the centres and facilities in the NERC Centres for Atmospheric Sciences, NCAS. NCAS carries out the core research programme in atmospheric science funded by NERC.

Home Contact Last Modified: 04/12/2002 17:16:56

**The Role: Key words: Curation and Facilitation!**



# Data Sets

**List of all datasets in BADC catalogue - Microsoft Internet Explorer provided by SSTO Office Systems**

Address: [http://www.badc.rl.ac.uk/cgi-bin/database/public/list\\_all\\_datasets.cgi?db\\_name=badcdb](http://www.badc.rl.ac.uk/cgi-bin/database/public/list_all_datasets.cgi?db_name=badcdb)

Start date: 02-apr-1950 End date: 30-dec-1950  
[Catalogue record](#) [Dataset web page](#)

**Chilbolton Radar data**  
Start date: 23-may-1998 End date: Not defined  
[Catalogue record](#) [Dataset web page](#)

**Climatology Interdisciplinary Data Collection**  
Start date: 01-jan-1854 End date: 31-dec-1996  
[Catalogue record](#) [Dataset web page](#)

**COAPEC HadCM3 model data**  
Start date: 01-jan-2079 End date: 31-dec-2178  
[Catalogue record](#) [Dataset web page](#)

**Cospar International Reference Atmosphere**  
Start date: 01-jan-1988 End date: 31-dec-1988  
[Catalogue record](#) [Dataset web page](#)

**Cryogenic Limb Array Etalon Spectrometer**  
Start date: 21-oct-1991 End date: 05-may-1993  
[Catalogue record](#) [Dataset web page](#)

**Earth Radiation Budget Experiment (ERBE)**  
Start date: 01-nov-1984 End date: 28-feb-1990  
[Catalogue record](#) [Dataset web page](#)

**ECMWF 15-year re-analysis data (ERA-15)**  
Start date: 01-jan-1979 End date: 28-feb-1994  
[Catalogue record](#) [Dataset web page](#)

**Met Office -Operational NWP Data Products (UM)**  
Start date: 26-oct-2000 End date: Not defined  
[Catalogue record](#) [Dataset web page](#) [Get data](#) [Register](#)

**Met Office Met. Research Flight C-130 data**  
Start date: 17-may-1993 End date: Not defined  
[Catalogue record](#) [Dataset web page](#) [Get data](#) [Register](#)

**Meteosat Images of Europe**  
Start date: 11-nov-1999 End date: Not defined  
[Catalogue record](#) [Dataset web page](#) [Get data](#)

**Microwave Limb Sounder (MLS L3)**  
Start date: 19-sep-1991 End date: 14-jun-1997  
[Catalogue record](#) [Dataset web page](#) [Get data](#)

**Microwave Limb Sounder (MLS) prototype H<sub>2</sub>O data**  
Start date: 18-sep-1991 End date: 13-apr-1993  
[Catalogue record](#) [Dataset web page](#) [Get data](#) [Register](#)

**Network for the Detection of Stratospheric Change (NDSC)**  
Start date: 01-jan-1991 End date: Not defined  
[Catalogue record](#) [Dataset web page](#) [Get data](#)

**Portable Unified Model (PUM) software from the Met Office**  
Start date: 21-nov-2001 End date: Not defined  
[Catalogue record](#) [Dataset web page](#) [Get data](#) [Register](#)



# Querying datasets

Query results - Microsoft Internet Explorer provided by SSTD Office Systems

File Edit View Favorites Tools Help Address <http://badc.nerc.ac.uk/data/atrs/>

Home My BADC Data Search Community Help

Get Data Access Rules Submit Data Dataset Index

## Along Track Scanning Radiometer (ATSR-1)

**i Introduction**

The data on this 2 CD set is derived from the the first Along-Track Scanning Radiometer (ATSR-1) which is a four-channel, dual-view, infra-red radiometer capable of measuring Sea Surface Temperature to very high accuracy (better the 0.3K). The instrument was launched on the ESA remote sensing satellite (ERS-1) in July 1991.

The dataset consists of two types of data product: (a) Spatially averaged sea surface temperatures (ASSTs) and (b) Time averaged global maps. The ASSTs are provided daily in half-degree cells together with with temporal and positional confidence information. The time-averaged global maps are provided at half degree resolution averaged over 5 day and 1 month periods. The data on the CDs cover the four year period from August 1991 to July 1995 inclusive.

The Principal Investigator for ATSR-1 is Chris Mutlow at the Rutherford Appleton Laboratory (RAL).

**Access to Data - File Format**

The data are held on CD-ROM at the BADC and they can be made available for browsing purposes. If you want to use the data in earnest you should order a copy of the CD from the contact address listed below.

The data are stored as 2-byte and 4-byte integers in fixed-length records. The choice of this simple format means that it should be a relatively simple matter to read the data on a variety of different computer set-ups and operating systems. Software is provided on the CDs to read and plot the data.

**Documentation**

Home

Done

Internet

Search for Met Office

## Query result

Searched for: 'sea surface' Matching datasets: 8

Online

**Along Track Scan**  
Start date: 01-aug-1991  
[Catalogue record](#)

**Climatology Inter**  
Start date: 01-jan-1854  
[Catalogue record](#)

**COAPEC HadCM3**  
Start date: 01-jan-2079  
[Catalogue record](#)

**ECMWF 40-year r**

**Catalogue**

This page gives detail contain additional info

**Short name:** ATSR

**Full name:** Along Tra

**Summary:** An infra-re consists c averaged positional

**Web page:** [/data/atrs/](#)

**Start date**

**End date**

**Bounding box**

Northerly latitude

Southerly latitude

Easterly longitude

Westerly longitude

**Min. altitude (Km)**

**Max. altitude (Km)**

**Altitude resolution**

**Latitude resolution**

**Longitude resolution**

**Time resolution**
























# Example: Met Office Products

21  
datasets

Query results - Microsoft Internet Explorer provided by SSTD Office Systems

Address: [http://badc.nerc.ac.uk/cgi-bin/search/select\\_search](http://badc.nerc.ac.uk/cgi-bin/search/select_search)

Back Forward Stop Home Search Favorites Media Print Mail

	Start date: 01-jan-1945 End date: Not defined		
	<a href="#">Catalogue record</a> <a href="#">Dataset web page</a> <a href="#">Get data</a> <a href="#">Apply for access</a>		
	<b>Met Office - Northern Hemisphere Mean Sea Level Pressure fields (1873-present)</b>		
	Start date: 01-jan-1873 End date: Not defined		
	<a href="#">Catalogue record</a> <a href="#">Dataset web page</a> <a href="#">Get data</a> <a href="#">Apply for access</a>		
	<b>Met Office - Stratospheric Assimilated Data</b>		
	Start date: 17-oct-1991 End date: Not defined		
	<a href="#">Catalogue record</a> <a href="#">Dataset web page</a> <a href="#">Get data</a> <a href="#">Apply for access</a>		
	<b>Met Office - TOVS Stratospheric Analyses</b>		
	Start date: 01-dec-1978 End date: 30-apr-1997		
	<a href="#">Catalogue record</a> <a href="#">Dataset web page</a> <a href="#">Get data</a> <a href="#">Apply for access</a>		
	<b>Met Office - UK High Resolution Radiosonde Data</b>		
	Start date: 01-apr-1990 End date: Not defined		
	<a href="#">Catalogue record</a> <a href="#">Dataset web page</a> <a href="#">Get data</a> <a href="#">Apply for access</a>		
	<b>Met Office - UK Land Surface Stations data (1900-present)</b>		
	Start date: 01-jan-1900 End date: Not defined		
	<a href="#">Catalogue record</a> <a href="#">Dataset web page</a> <a href="#">Get data</a> <a href="#">Apply for access</a>		
	<b>Met Office -Operational NWP Data Products (UM)</b>		
	Start date: 26-oct-2000 End date: Not defined		
	<a href="#">Catalogue record</a> <a href="#">Dataset web page</a> <a href="#">Get data</a> <a href="#">Apply for access</a>		

Internet



# BADC: Geographical Searching

The screenshot shows a Microsoft Internet Explorer browser window displaying the BADC website. The address bar shows the URL: [c.ac.uk/cgi-bin/ukmo/ukmo\\_list\\_station.cgi?src\\_id=5933](http://c.ac.uk/cgi-bin/ukmo/ukmo_list_station.cgi?src_id=5933). The page title is "Met Office station details - Microsoft Internet Explorer provided by SSTD Office Systems".

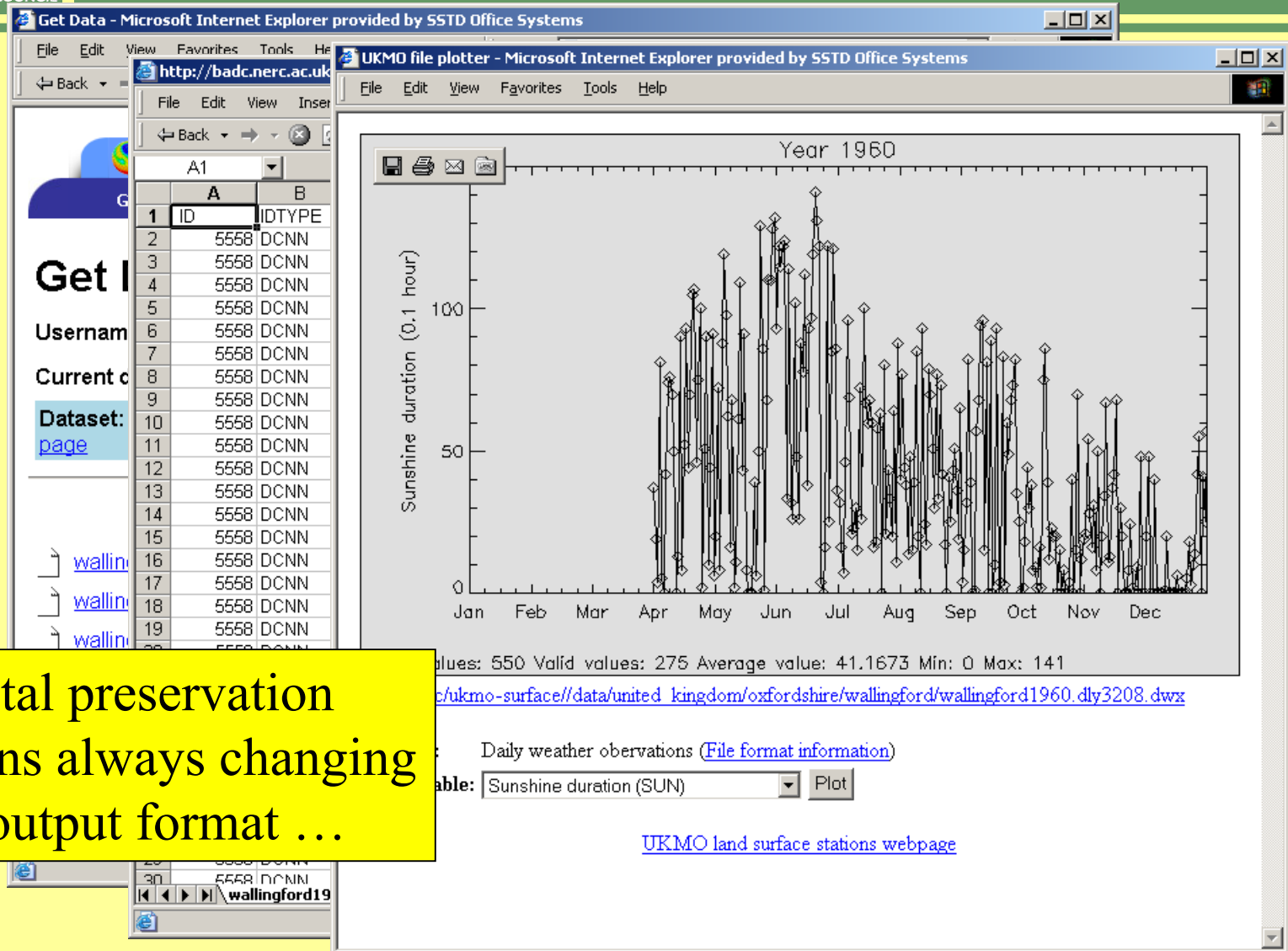
The website has a navigation bar with links: Home, My BADC, Data, Search, Community, Help. Below this is a search bar and links: Search for Met Office stations, NERC Metadata Gateway, A-Z page index.

The main heading is "Met Office station details". The details for station "BRITWELL SALOME" are as follows:

<b>Name:</b>	BRITWELL SALOME <a href="#">Locate on map</a>
<b>Geographic area:</b>	OXFORDSHIRE
<b>Latitude (decimal degrees):</b>	51.634 ( <a href="#">Original Met Office value</a> = 51.598)
<b>Longitude (decimal degrees):</b>	-1.027 ( <a href="#">Original Met Office value</a> = -1.048)
<b>Grid ref:</b>	SU 673934 (467.3km East, 193.4Km North of National grid origin)
<b>Grid ref type:</b>	Ordnance Survey
<b>Postcode:</b>	OX9 5
<b>Elevation:</b>	113 meters
<b>Time zone:</b>	0
<b>Drainage stream:</b>	THAME
<b>Hydrological area ID:</b>	392
<b>Station start date:</b>	01-01-1942
<b>Station end date:</b>	31-12-1975
<b>BADC data directory</b>	<a href="/badc/ukmo-surface/data/united_kingdom/oxfordshire/britwell_salome">/badc/ukmo-surface/data/united_kingdom/oxfordshire/britwell_salome</a>
<b>Measurements made:</b>	

On the left side of the browser window, there is a sidebar with a search bar and a list of search results. The results are sorted by "Number of stations" and "Results sorted by". The list includes: PYRTON MAN, BRITWELL SA, CUXHAM Loc, SHIRBURN MO, SWYNCOMBE, ASTON ROWA, and STONOR Loc.

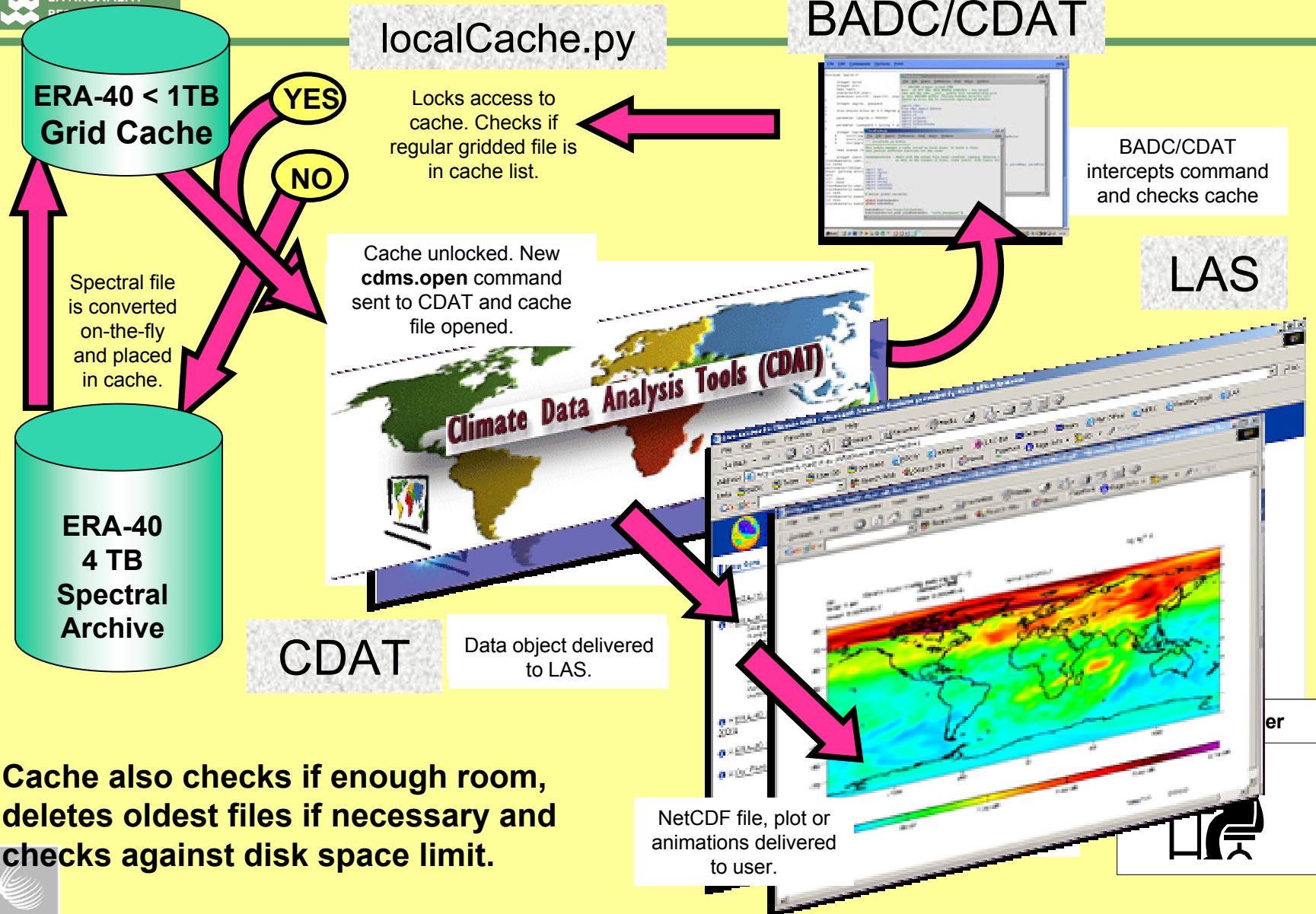
# Different types of data returned: Wallingford



Digital preservation  
means always changing  
the output format ...



## Cache management in CDAT



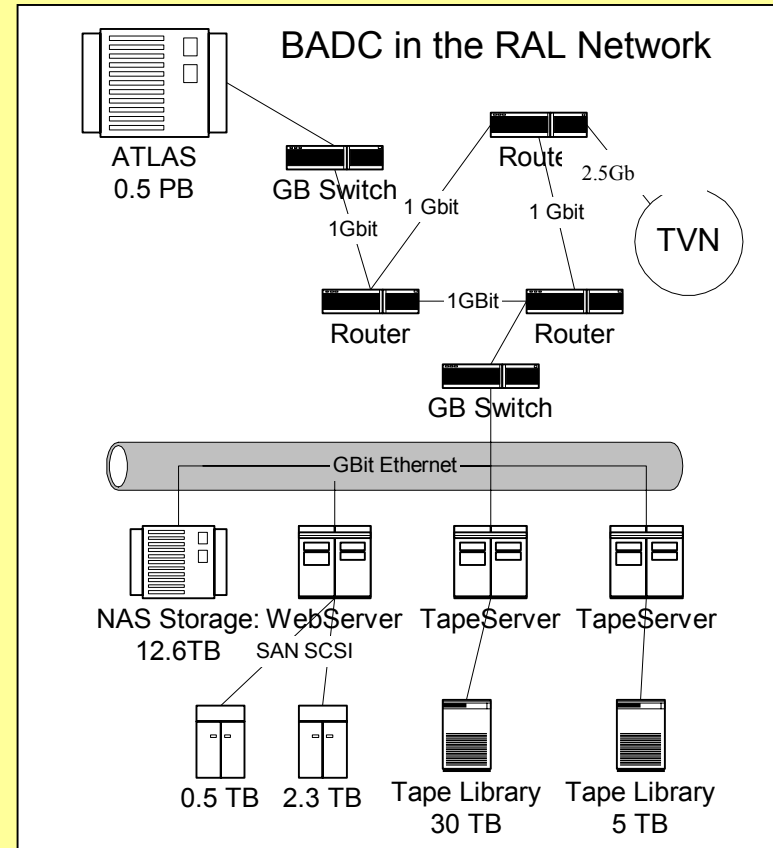
# Typical datasets and usage (2001)

Dataset	Number Files	Size (MB)	Active Users	Registered Users
<b>ACSOE</b>	2558	318	26	Public
<b>MO assimilated data</b>	62174	172692	109	221
<b>ECMWF ERA15</b>	109445	165436	80	327
<b>ECMWF operational data</b>	143437	240095	97	376
<b>ECMWF trajectories</b>	2000	772	51	284
<b>Meteosat</b>	71851	7963	42	Public
<b>MST Radar</b>	61460	33498	39	29
<b>MO Radiosondes</b>	1457695	44750	160	433
<b>MO Surface data (MIDAS)</b>	307738	41974	479	936
<b>Ozone Climatology</b>	723	335	7	Public
<b>MO Unified Model</b>	635	292	23	26
<b>URGENT</b>	2153	340	30	34

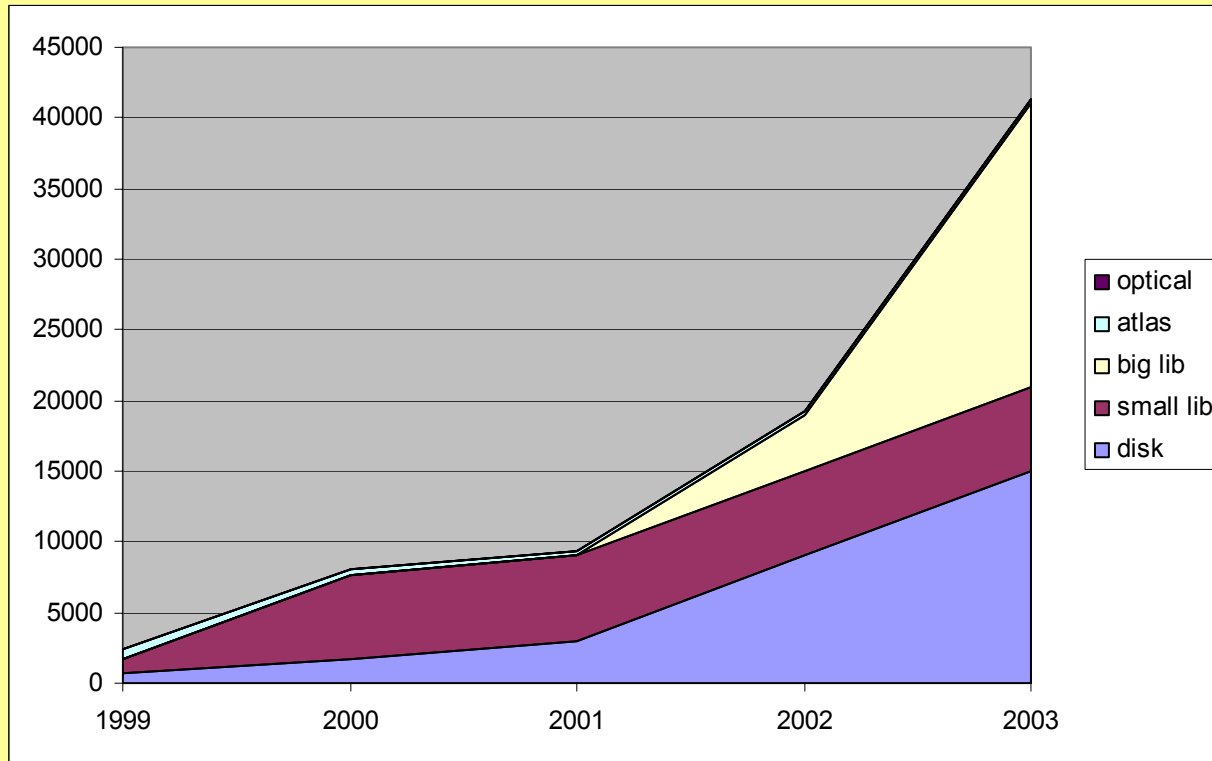
- over 1 TB of data delivered to users in the year!

- Primarily digital data
  - All data delivered via the Internet
- Security of data
  - Maintained using login ids where necessary
- >20 TB of data online, more near-line, growing incredibly rapidly.
  - Comprehensive archive for all NERC atmospheric data.
  - Accessible archive of Met Office and ECMWF data
  - UK Accessible archive of appropriate U.S. datasets
- Embedded in scientific community
  - Head and research group actively involved
  - Part of NCAS, part of NERC, part of CLRC
  - Intimately involved in e-science

- Approx 50 TB (Nov02)
- Projected to double well within next couple of years given existing commitments
- Committed to keeping as much as possible on spinning disk
- Further backup and extra storage at national archival centre (ATLAS, PB soon)



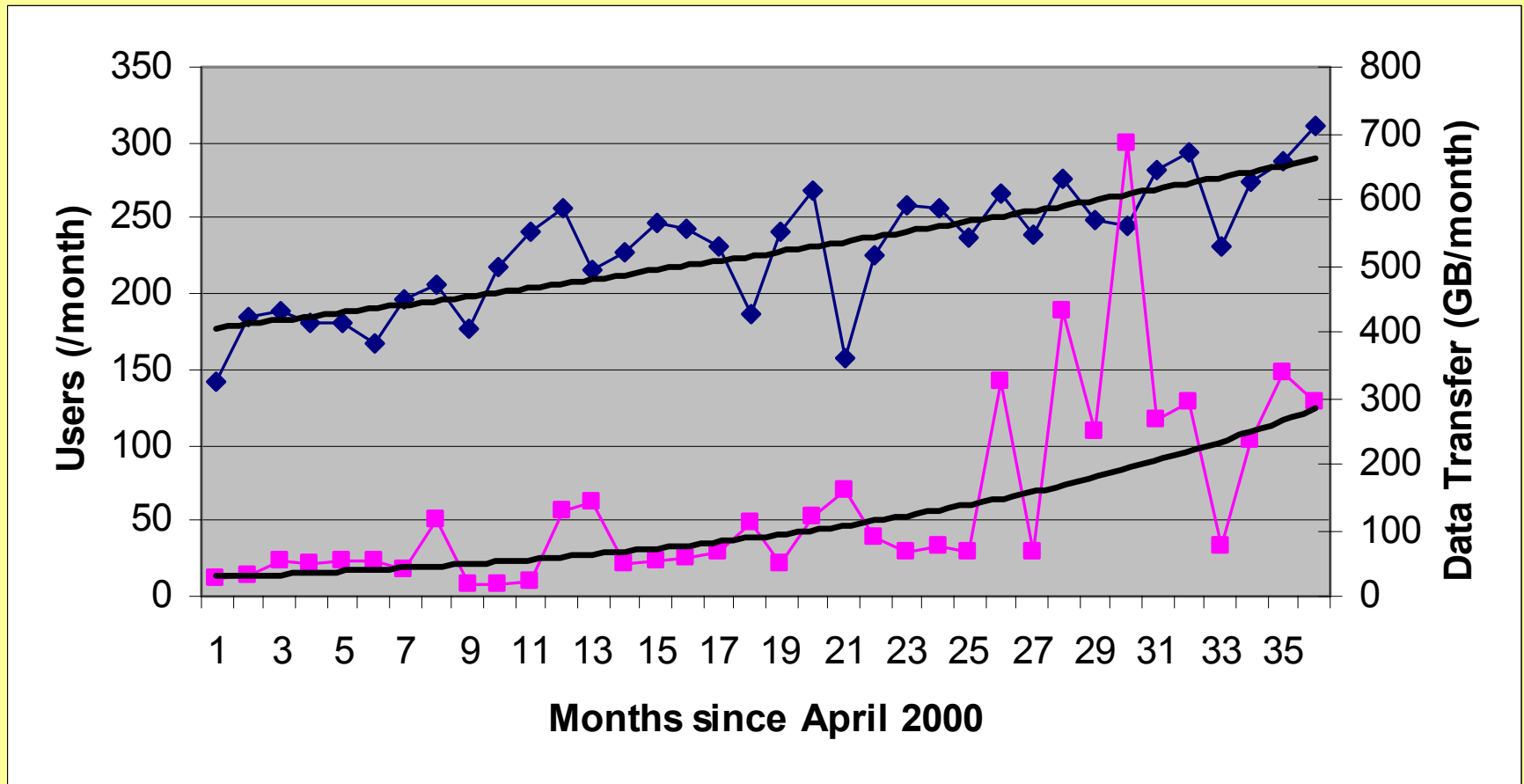
# BADC Data Storage



Estimated data volumes (including backup). The key issue is growth. BADC not always “official archive of last resort”, but advent of “data publication” has and will change that.



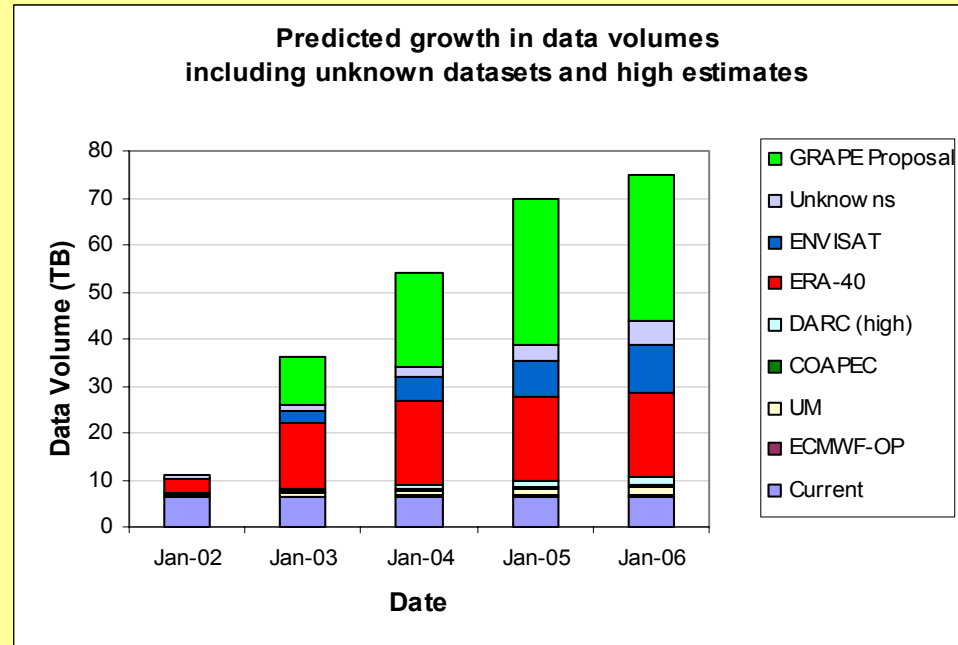
# Usage Growth



1151 distinct users from 12/01 until 11/02 downloaded 2.5 TB



# BADC Storage Planning

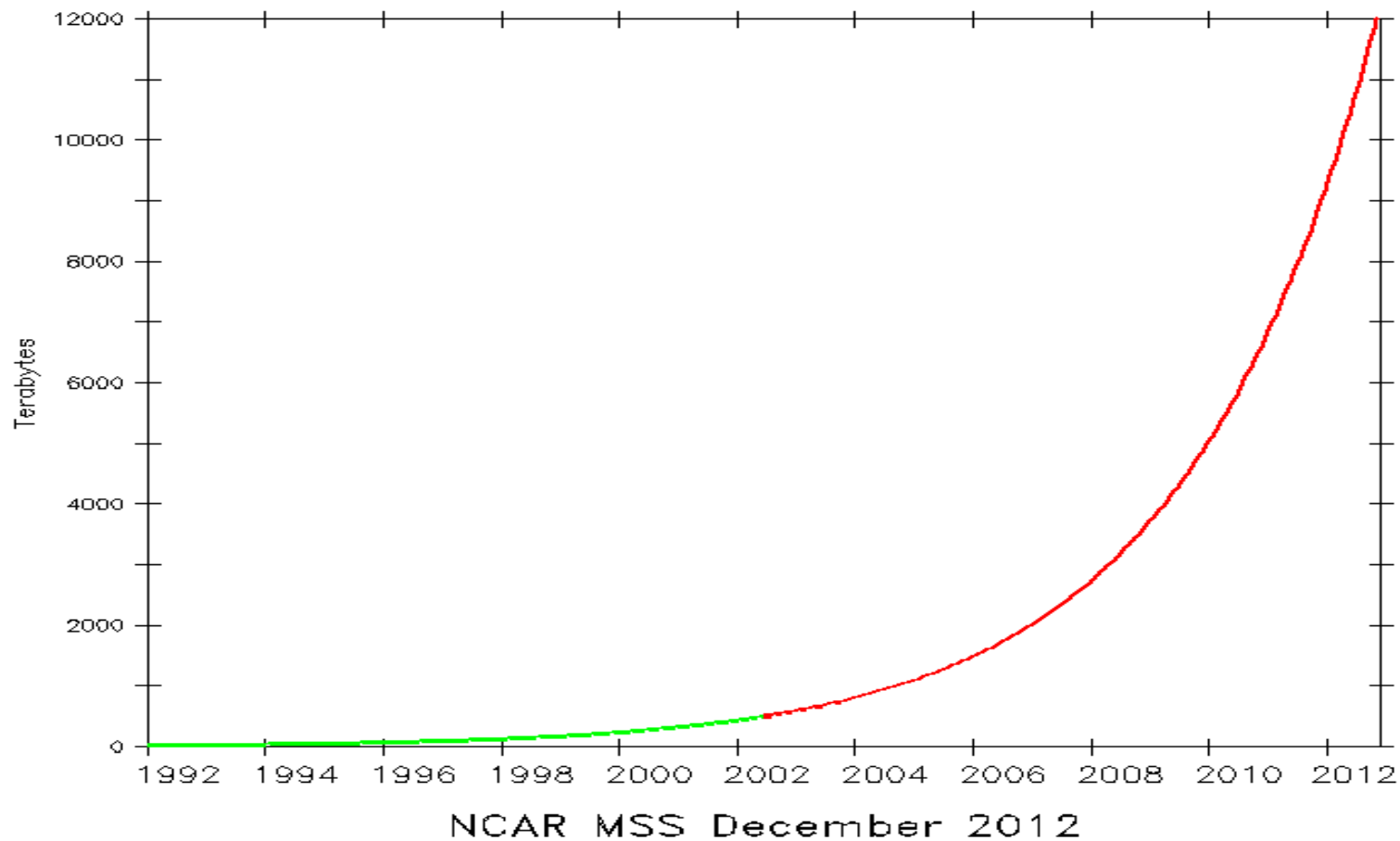


Data Growth estimates (August 01) - slightly above where we are now, but GRAPE hasn't started. No MSG included. No new Hadley Centre data. About to go horribly wrong.

- can predict about 1.5-2 years, no more!



Figures from Gary Strand, NCAR, ESG website





## Data

- Nasa Ames
- GRIB
- BUFR
- NetCDF
- HDF

Relatively little “picture image” data:  
(.gif, .jpg ...)

Database formats:

- Ingres, hopefully migrating to MySQL (licensing issue)

## Documents

- HTML
- ASCII
- PDF

And then the problems:

- .doc, .ppt, and many others

Desirable aim: all data and all documentation in open source formats ... we can't afford the risk associated with predatory licensing policies ...

# Easily catalogued, but successful preservation?

	a	e	i	o/u
	†; (	*A	†	†; F
y	目	X	*田	*王; *M
w	田	S	*田	*A, R
r	h; &	ψ	*田	+; φ
m	田	q; 7	√	*▽; *A
n	▽, T, T	ψ	Y	ψ; H
p	†; H	*Σ (1)	△, △, △	†; A; ψ
t	□; △△	ψ	△, ^	T; △; *A
d	†	△	†	*; *A
k	⊕, ⊖	2, 2, 2, 2	▽	†; †
q	▽	⊙	q	†(1)
s	Y	ψ, ψ, ψ	*G	*A, I, I, E
z	†	A		†

non-placés: L8 田 (yat?); e1 7 (q?); 35 7 (naa?); 36 4 (ko?)  
L37 7 (qa?); 43 田, 田 (wa?); 65 7 (ki?); 90 7 (ka?).

**filum of Linear A'**



Phaistos Disk, 1700BC

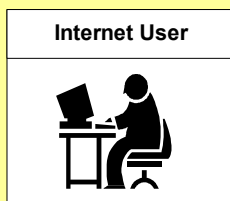
One could argue that the writers of these documents did a brilliant job of preserving the bits-and-bytes of their time ...

And yes they've both been translated ... many times, it's a shame the meanings are different ...

# What is metadata for?

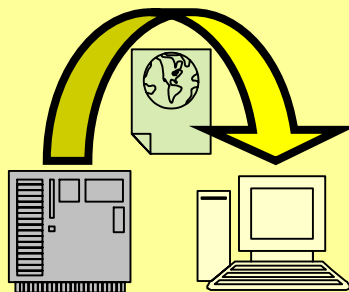


Firstly: information to *help one use* one's own data: e.g. calibration data (A)

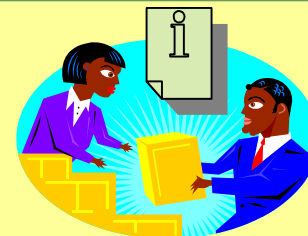


Metadata can help one find other people's data  
... and then help one *obtain and use* it. (C)

The answer depends on who you are!



Metadata can be used to *enable automatic software* to manipulate data. (D)



It is information passed with the data to *enable* someone *else* to use it. It *describes* the data. (B)



Metadata can be used to enable the *preservation* of data for posterity (all of ABCD)

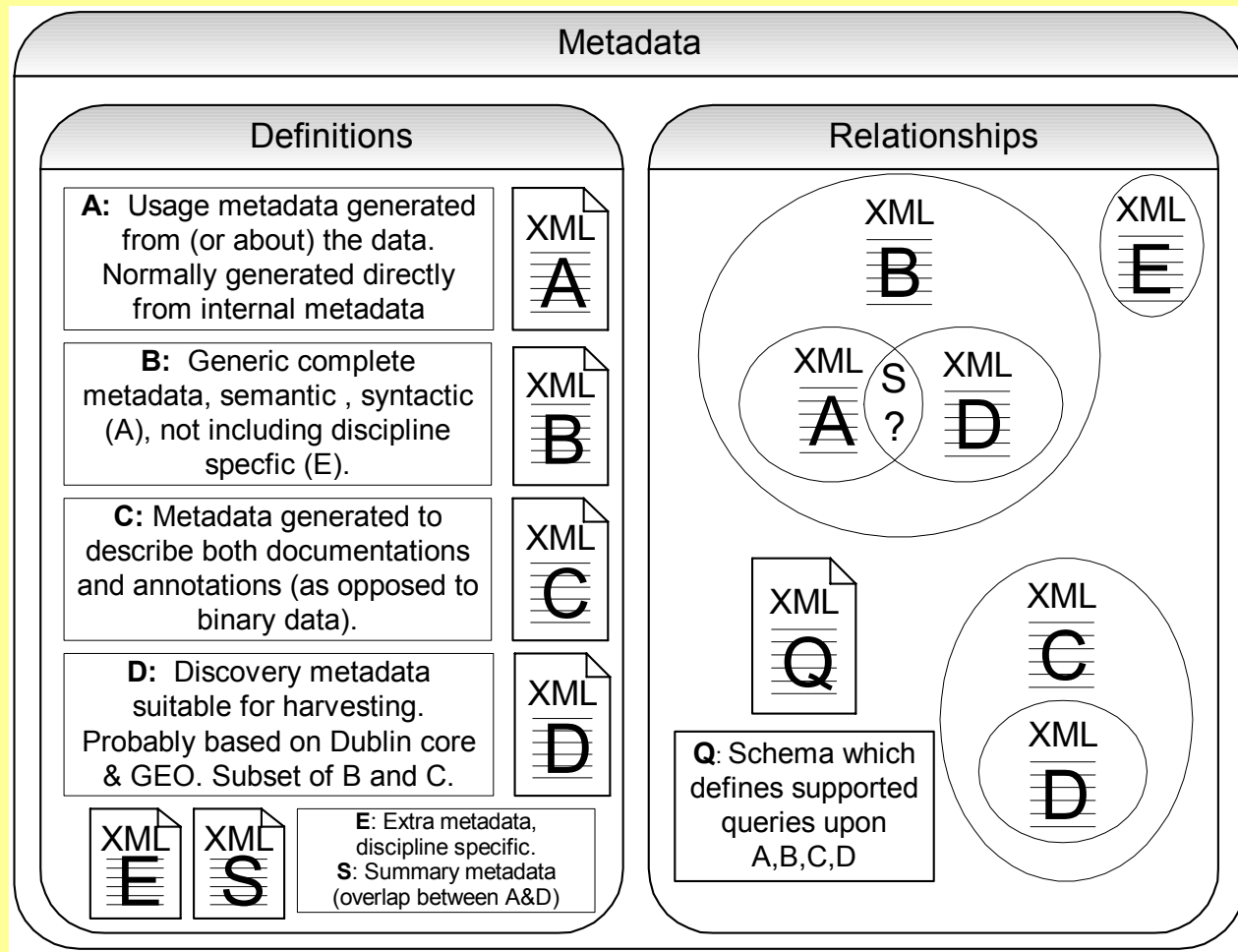
1. Discovery Metadata
  - “dataset” descriptions ... “high-level”.
2. Syntactic Metadata
  - the “format” of the data.
3. Semantic Metadata
  - what the data means.
4. Referential Metadata
  - metadata which refers to ancillary metadata.
5. Ancillary Metadata
  - Tables of “codes”, thesauri, ... ontologies.
  - Publications, “Grey Matter”

No one metadata “standard” applies to all of these! The lower down the hierarchy, the less a standard will be widely applicable.

- ALL data activities need to consider all of these from the beginning!

# NDG: Required "Data" Metadata

Need a tool  
to generate  
B!



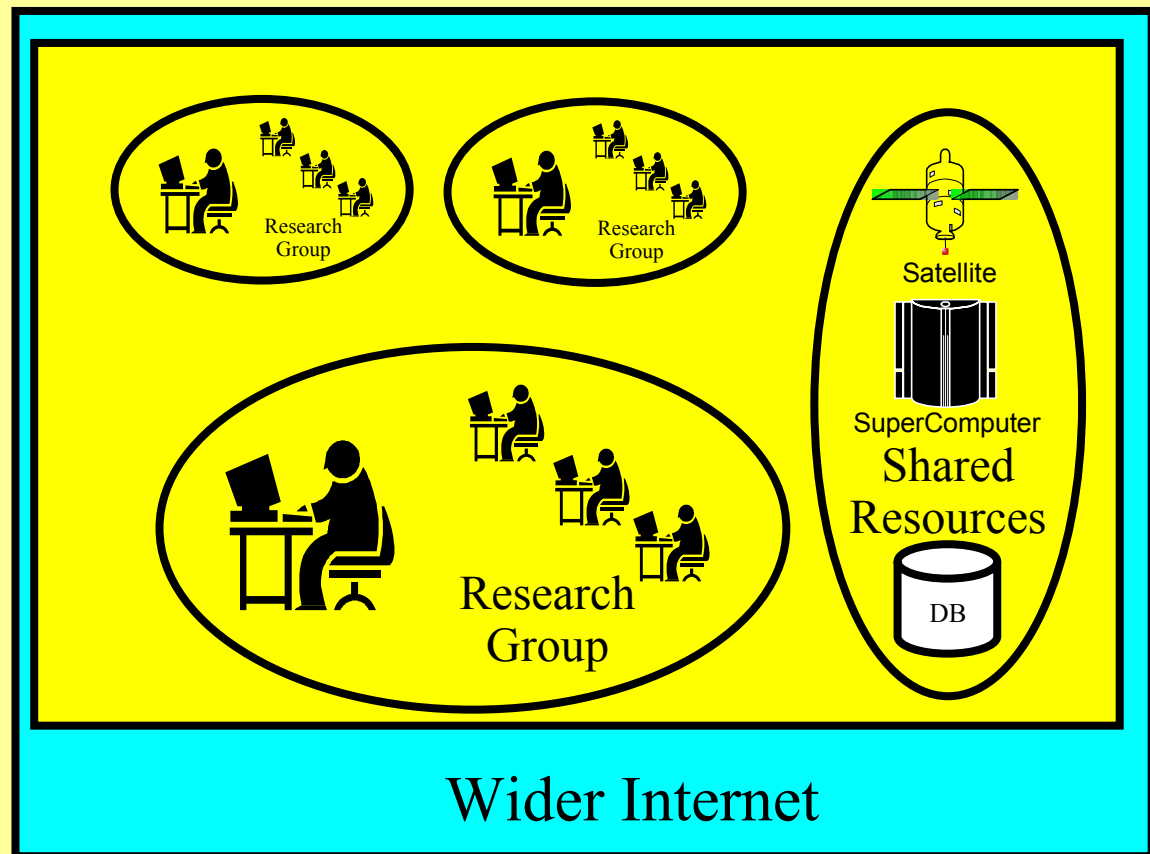
# Where does the data come from?

Consider a hierarchy of data users beginning with an individual scientist, who may herself be part of a research group, itself part of a community sharing resources, lying in the wider internet ...

To be well integrated the metadata should have a role at each level!

(The data portal client and server interface may be different at each level).

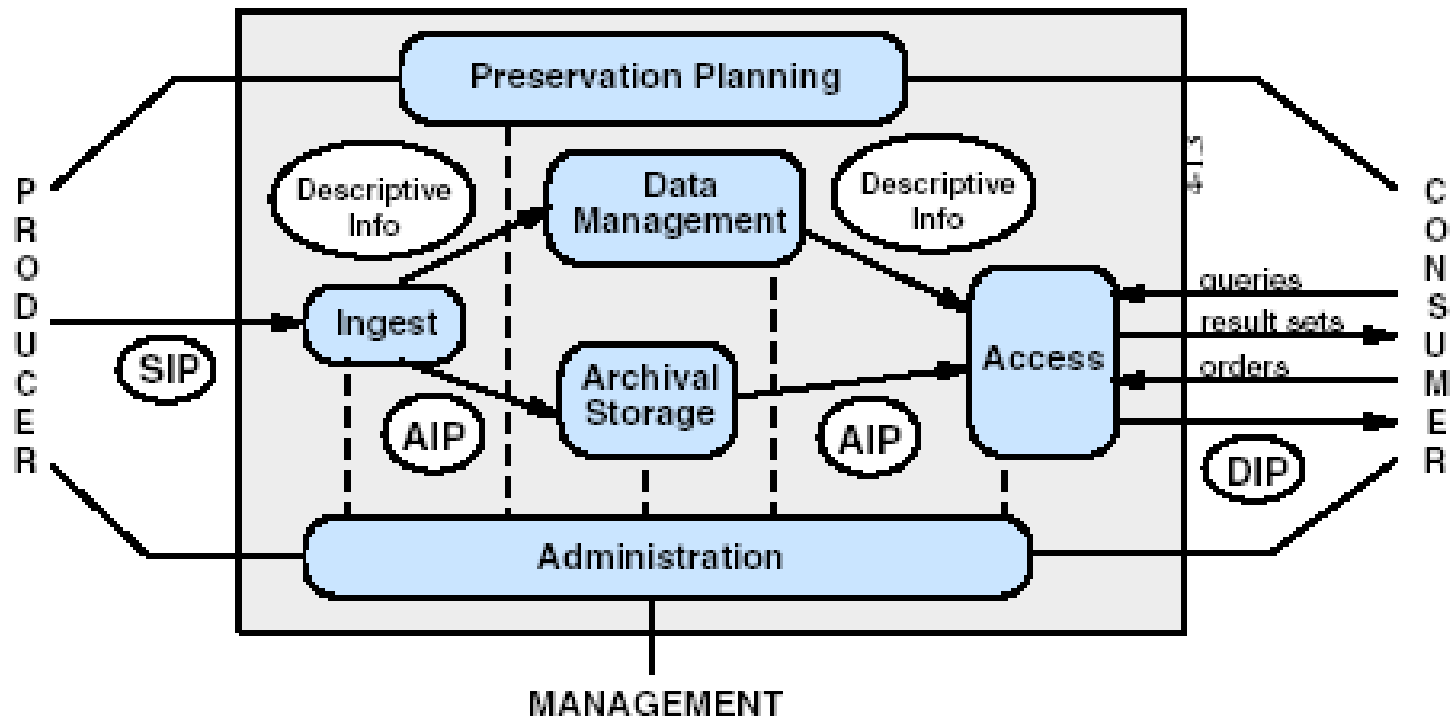
At each level “extra” metadata will be required, probably produced by dedicated staff at the research group, or data centre.



The world is full of WORM drives  
(Write Once Read Many).

- It is also full of WORN archives (and databases)
  - (Write Once Read Never)
- Little or No metadata leads to WORN archives!
  - Metadata is key to preserving data for posterity!
  - It's also a key to making use of the data now!
- It is possible to spend an infinite sum of money on metadata without doing the job right!
- How do you know what to do?
  - Start with a sensible expandable database schema!
  - Be involved in the use of the data and try and identify and preserve institutional wisdom.
  - Accept the extra labour overheads will exist for the data **producer** as well as **archiver**!

# The OAIS Archive Model

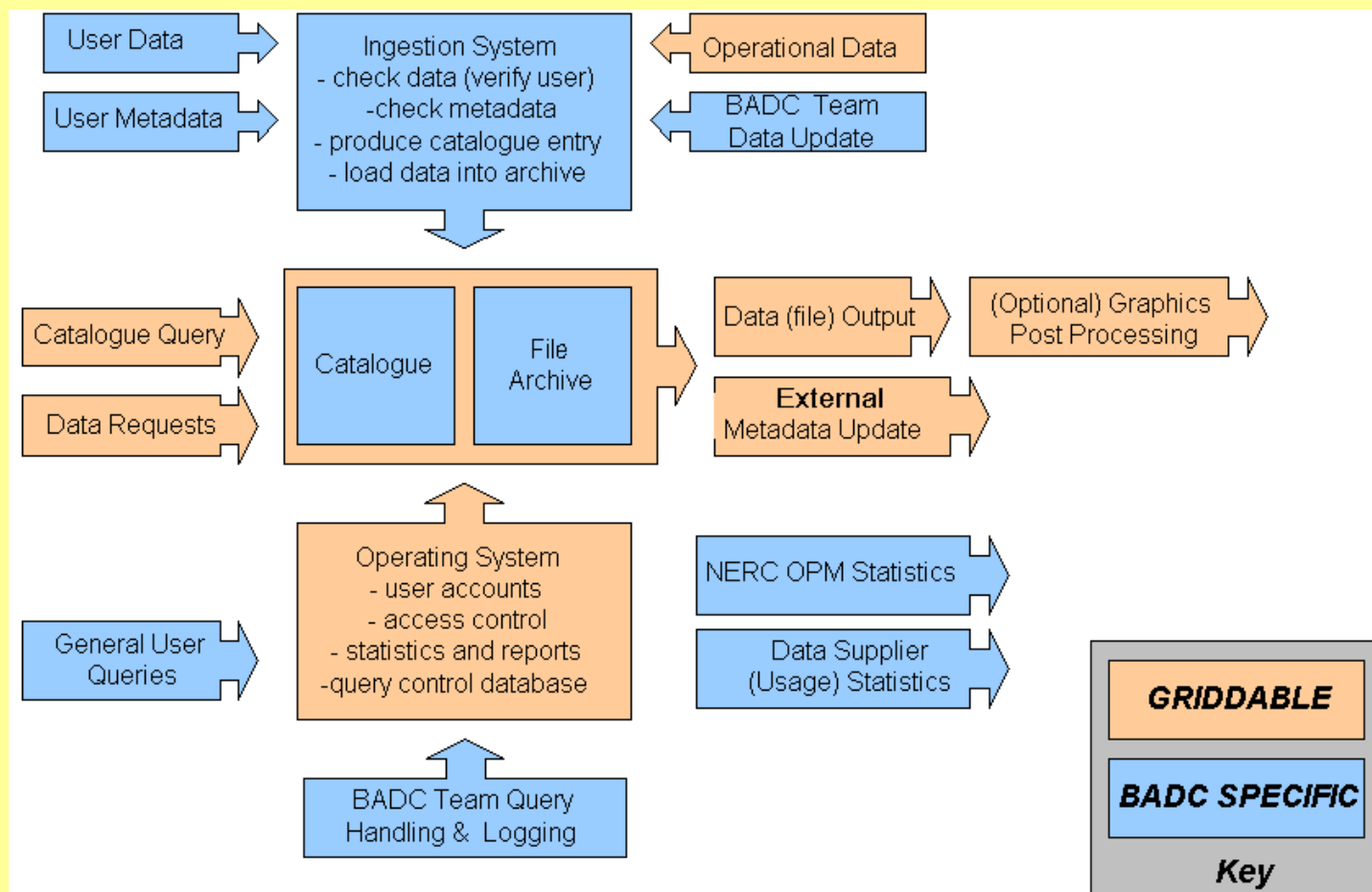


Open Archival Information System:

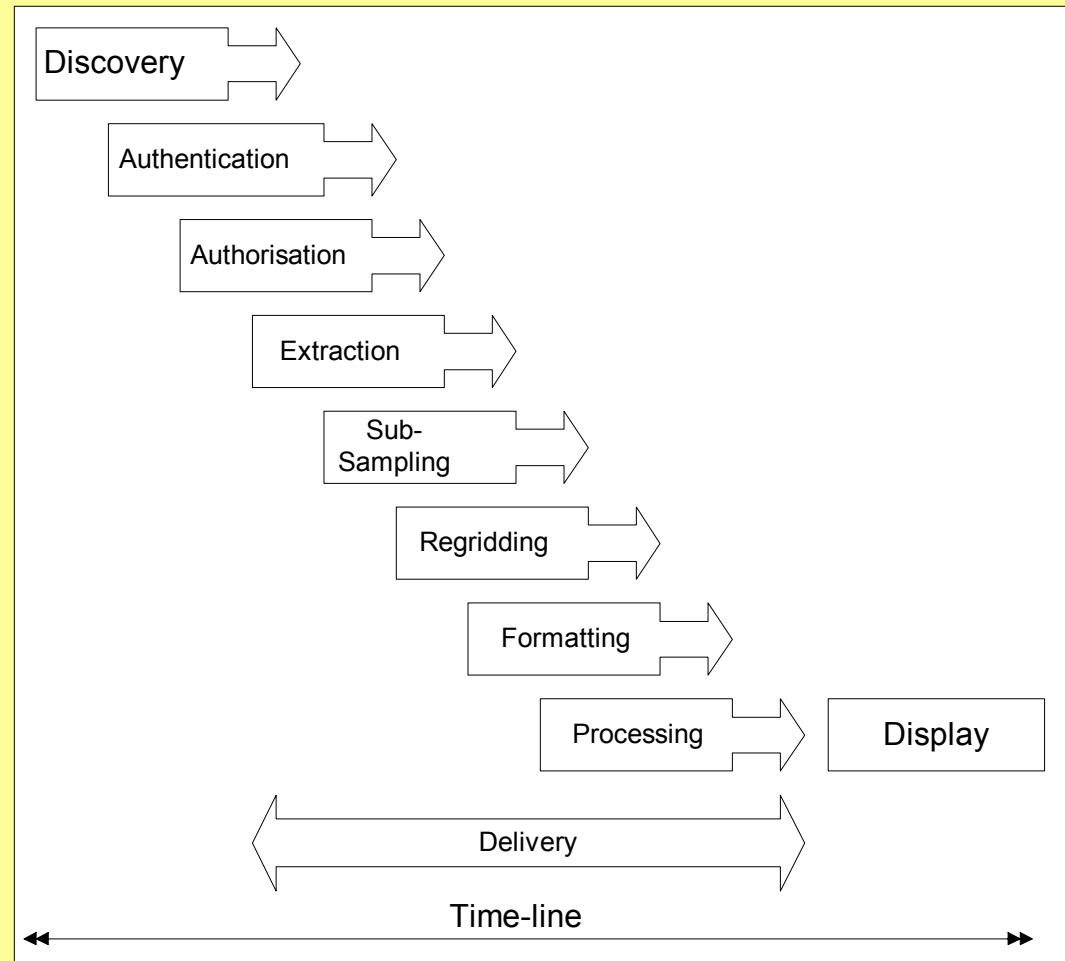
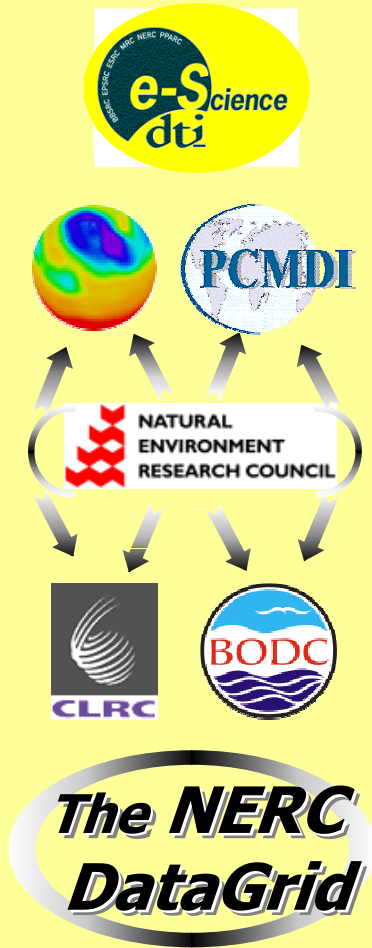
- the producers and the consumers are a key part of the structure!!



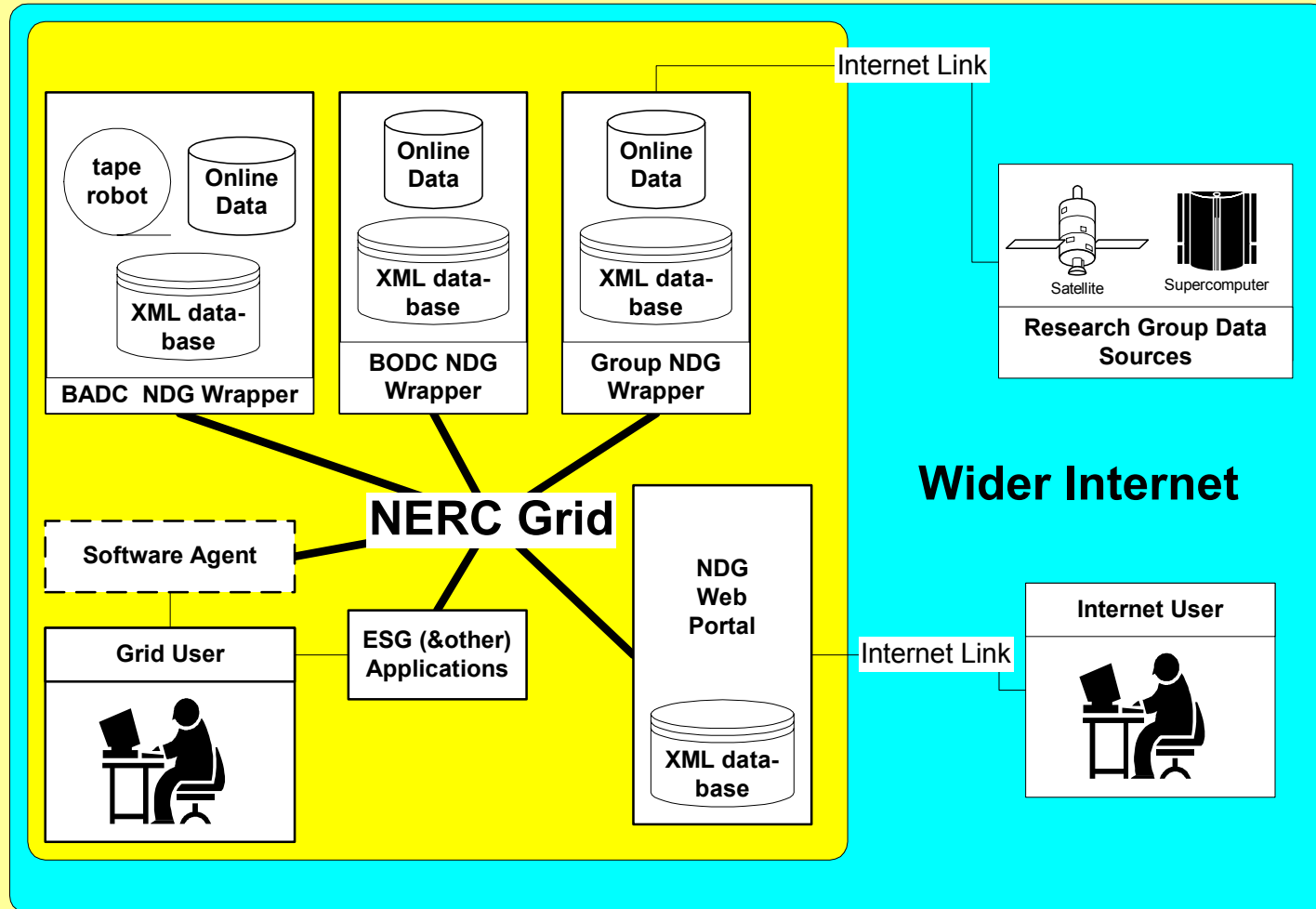
# BADC and e-science



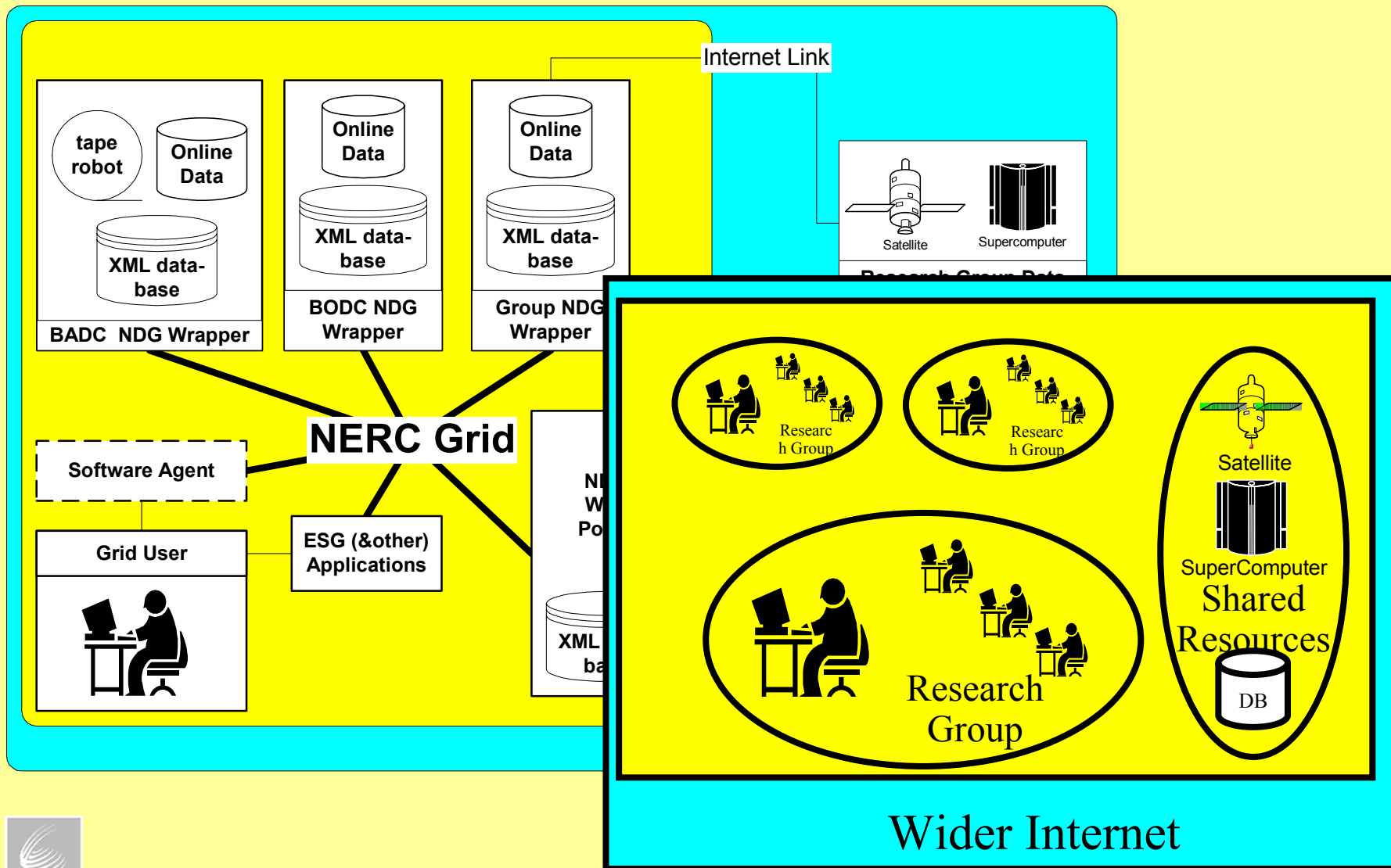
# The Data Use Chain



# The NERC Data Grid & the NERC Metadata Gateway



# A google for data; the metadata carrot!



- Exponential growth in input data
  - Not just volume, variety ...
- Linear growth in users
- Exponential growth in data transfers
- Moore's Law is both our sword and our saviour – storage cost of old data is essentially free.
  - Must leverage off community tool development
  - Can't rely on increasing software license budget, must innovate and rely on free solutions ...
  - Must stay with the community: key issues are interdisciplinarity and annotation.

- BADC: ~ 4000 users, data volumes growing rapidly.
- Embedded in the research community.
- Depending on open source software and formats.
- Relying on rapidly evolving metadata standards to both obtain data and deliver data.
- Looking to the e-science future to minimise barriers between data producers and consumers
  - Long time preservation of archived data will depend on this; we can't rely on the wider community still having the tools and knowledge to “read” scientific data.
  - Hoping that “ontology” will become more than a buzzword ...