# Preserving the Living – experiences and reflections on semi-current ‘records’



With thanks to Sarah Jones & other DCC colleagues

**Kevin Ashley**
**Director, Digital Curation Centre**
**www.dcc.ac.uk**
**@kevingashley**
**Kevin.ashley@ed.ac.uk**

# My topics for today

- What are we talking about, and in what language?
- Some personal experiences
- Some reflections on more general issues

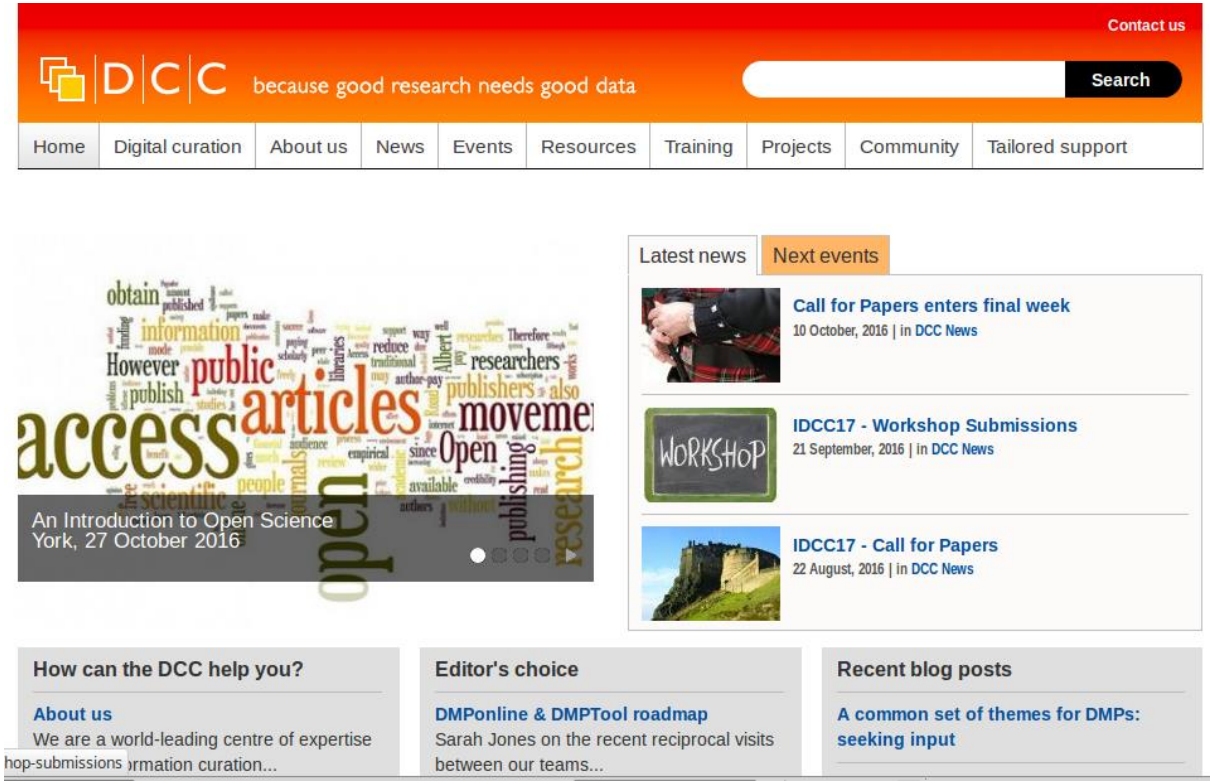
# ‘Semi-current records’

- The language of archives & records management
- Material that (to some extent) still serves the purpose for which it is created
- Implication of less use – therefore perhaps greater neglect
- Bound up with world of records as physical artefacts
  - Cannot transfer custody
  - Cannot transfer governance
- But if ‘records’ – still immutable?
- I think in many cases they aren’t…

# My home – the DCC

Mission – to help organisations worldwide make the best use of data in and for research

Training, shared services, consultancy, research, publications



# What is data curation ?

"M
res

More than

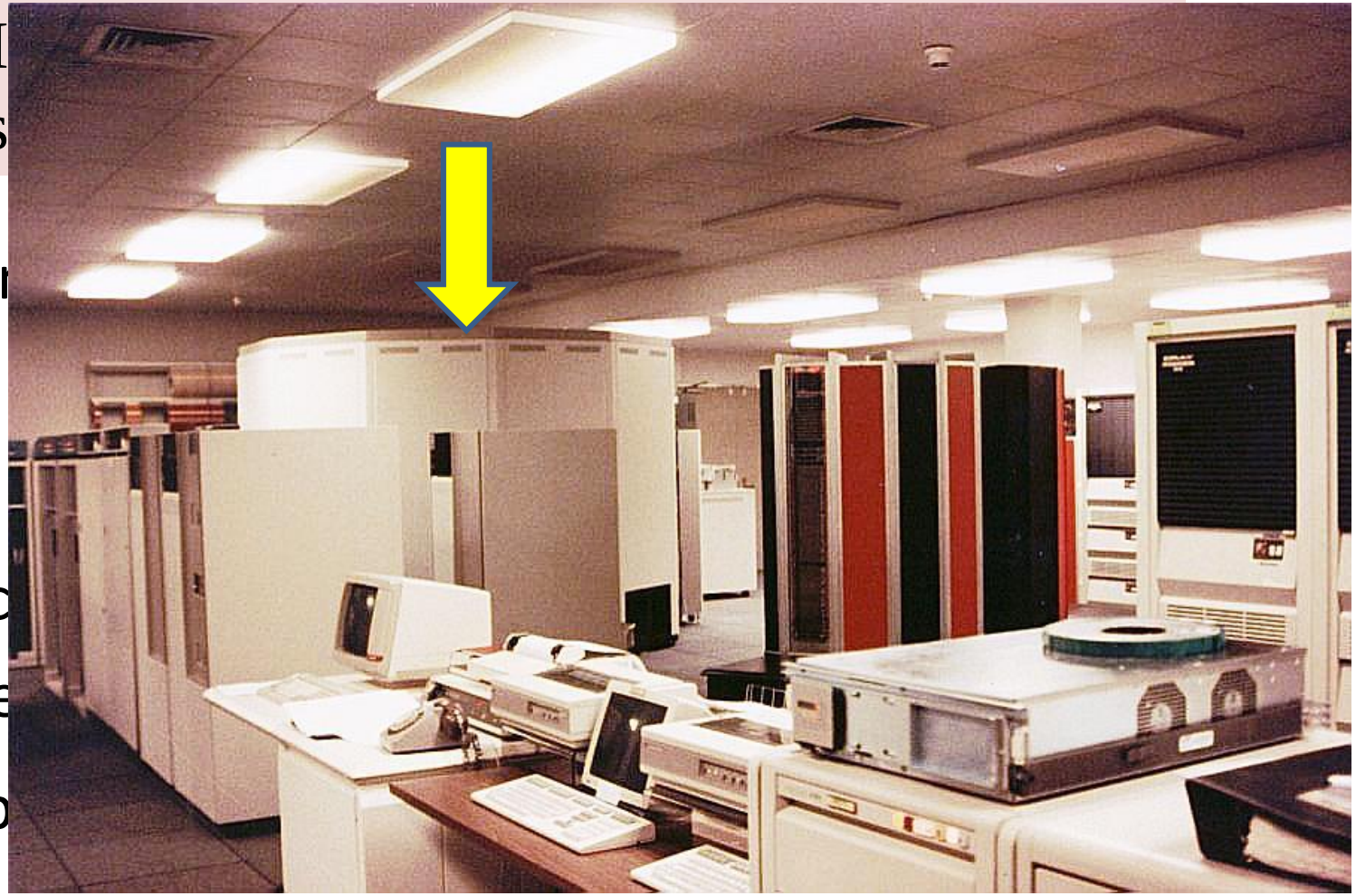
- Active

Less than

- Lifecyc

Sometime

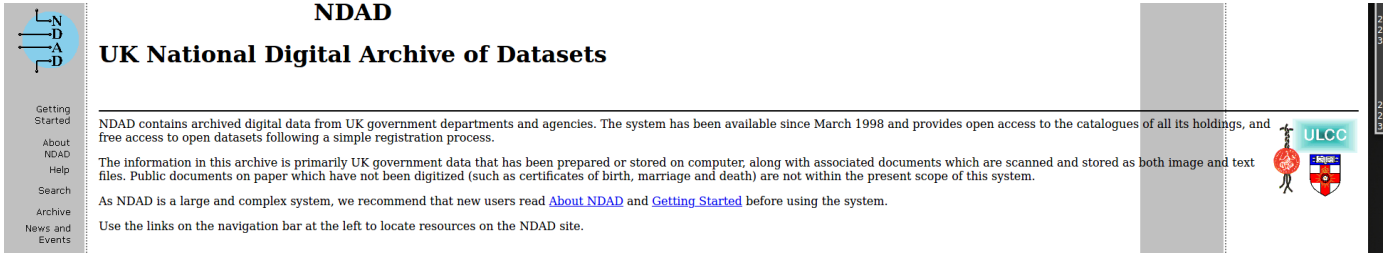
Always ab



# Anecdote time

- Web sites – classic example with complexity of content, constant change & constant use
- Also where what we capture is often distant from the 'record' we see
- Concern when running web-hosting services in mid-1990s for public bodies
- Also a concern of university archivists at that time
- Two big issues (other than the technical ones):
  - Organisational distance of creators from records staff
  - Unwillingness to see long-term utility of content

# Government databases




- From 1998 – 2010, we ran NDAD for the Public Record Office/National Archives
- Much data was non-current – but not all
- Some transfers were easy, and involved regular snapshots
- Some were not and involved challenges around control
- Others had effects on the systems being preserved

# When preservation changes the record

- Our ingest processes involved lots of data validation
- Unlike research data archives, we didn't fix bad data – we noted the errors
- For semi-current records, we told the originator about the problems we found
- … which often caused the original records to be changed
- Not just changed data, but (e.g.) changed database design as in DOMUS

# Scientific data - IUPHAR

- Example due to Peter Buneman – “How to cite curated databases & how to make them citeable” https://doi.org/10.1109/SSDBM.2006.28
- IUPHAR is a typical curated scientific database, aggregating information from many sources over time
- Scientists make assertions depending in part or in whole on its contents (and should cite it when they do)
- But how to validate those assertions when the data itself is changing?

# Examples of assertions

- All of these may change with time
- Should the hosts themselves be responsible for preservation of this varying record
- Or is that the job of a different archive – and if so how do I determine which I should be using?

*1. The IUPHAR database (C 1 ) contains no information about Ginandtonicin.*

*2. The IUPHAR database (C 2 ) lists five ligands for Melatonin receptor MT 1 .*

*3. The IUPHAR database (C 3 ) asserts that luzindole is an antagonist ligand for receptor MT 1 .*

# Funder requirements

● Full Coverage ◐ Partial Coverage ○ No Coverage

| | Policy Coverage | | Policy Stipulations | | | | | Support Provided | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Research Funders | Published outputs | Data | Time limits | Data plan | Access/ sharing | Long-term curation | Monitoring | Guidance | Repository | Data centre | Costs |
| AHRC | ● | ● | ● | ● | ● | ◐ | ○ | ● | ○ | ◐ | ◐ |
| BBSRC | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◐ | ● |
| CRUK | ● | ● | ● | ● | ● | ● | ● | ◐ | ● | ○ | ○ |
| EPSRC | ● | ● | ● | ◐ | ● | ● | ● | ◐ | ○ | ○ | ● |
| ESRC | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◐ |
| MRC | ● | ● | ● | ● | ● | ● | ○ | ◐ | ● | ○ | ◐ |
| NERC | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◐ |
| STFC | ● | ● | ● | ● | ● | ● | ● | ◐ | ● | ◐ | ◐ |
| Wellcome Trust | ● | ● | ● | ● | ● | ● | ● | ● | ● | ◐ | ● |

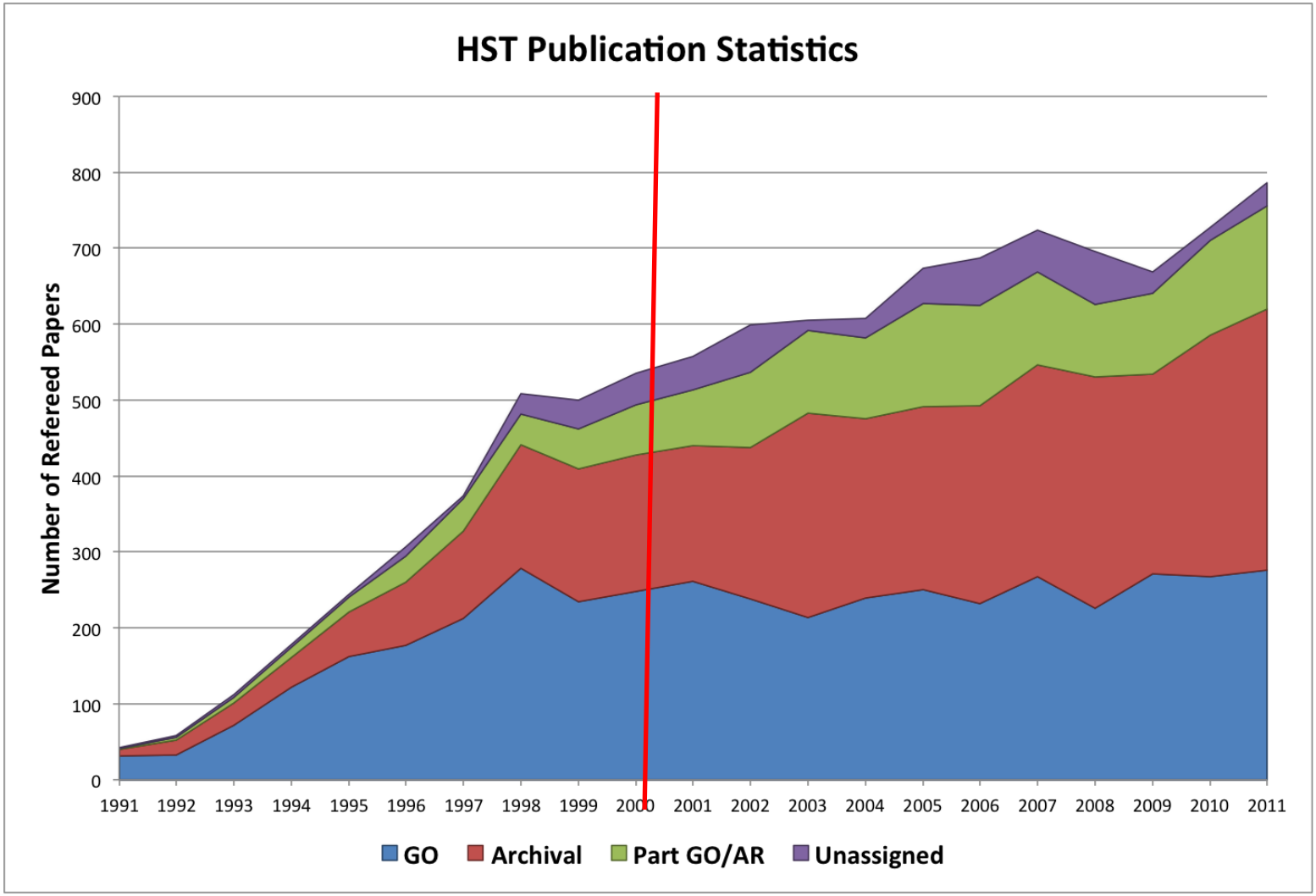
# Research data as semi-current record

- Sometimes simple – data is seen as ‘publication’ – a completed object, sent to somewhere for preservation & access
- Often analagous to more contentious issues NDAD saw in government – unwillingness to relinquish control
- But some examples show how this can be circumvented with appropriate workflows

# The Hubble Space Telescope

- Typical large scientific instrument with infrastructure for data management
- Researchers compete for its use – then get exclusive access to their observations for six months
- After this time, data is made public automatically
- The effect on research is dramatic

# Data reuse from Hubble



# Data management planning

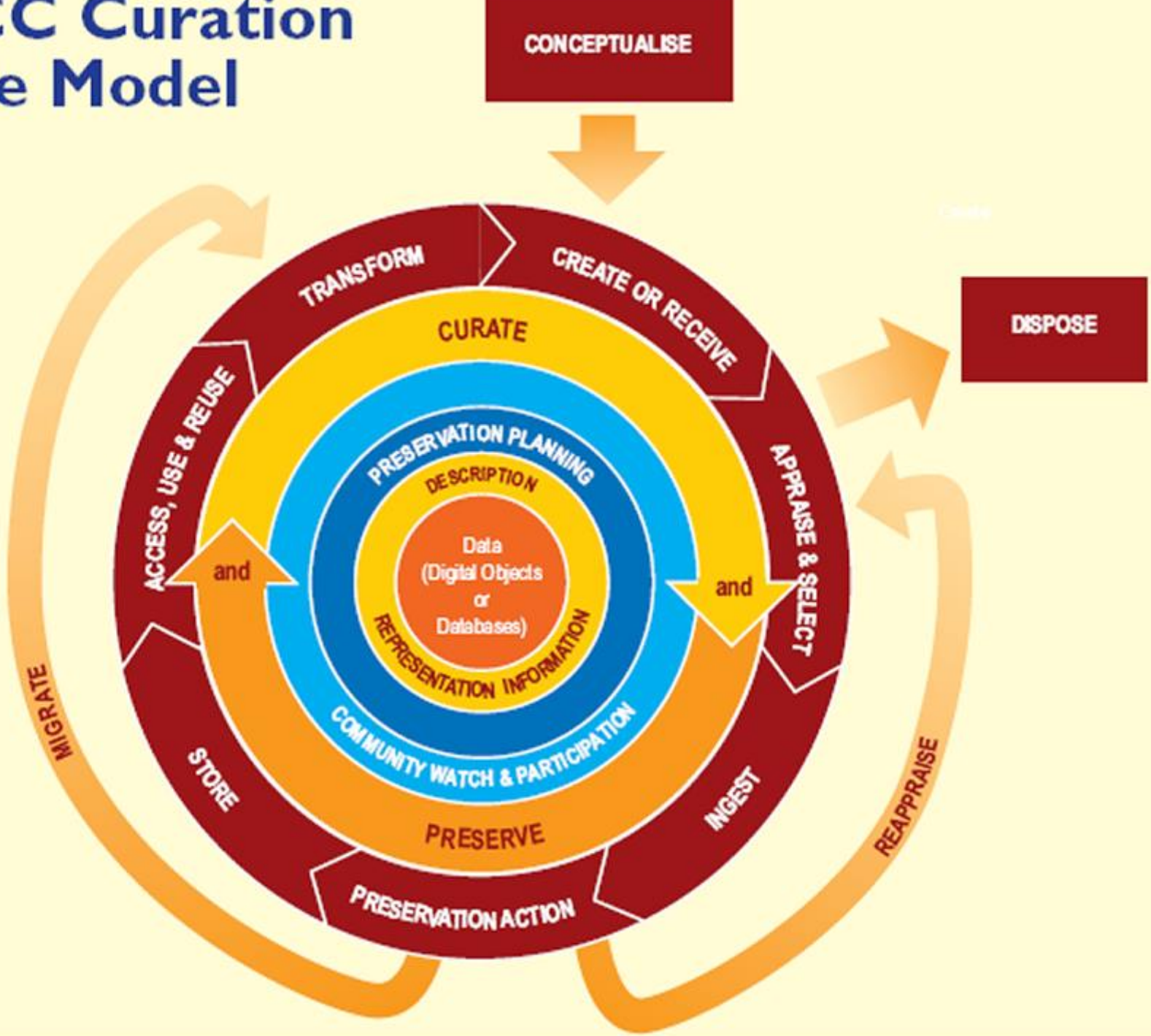
- Funder requirements on preserving data has generated increased use of discipline of data management planning
- Urges thinking about use, custody, governance, sharing before content creation begins

# Possible elements of a plan

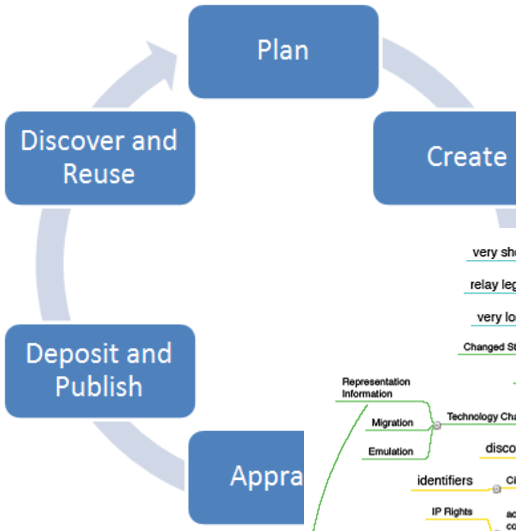
- Introduction & context
- Legal, rights & ethical issues
- Access, data sharing & re-use
- Data collection / development methods
- Data standards
- Short-term data storage & data management
- Deposit & long-term preservation
- Resourcing
- Compliance & review
- Agreement/ratification by stakeholders

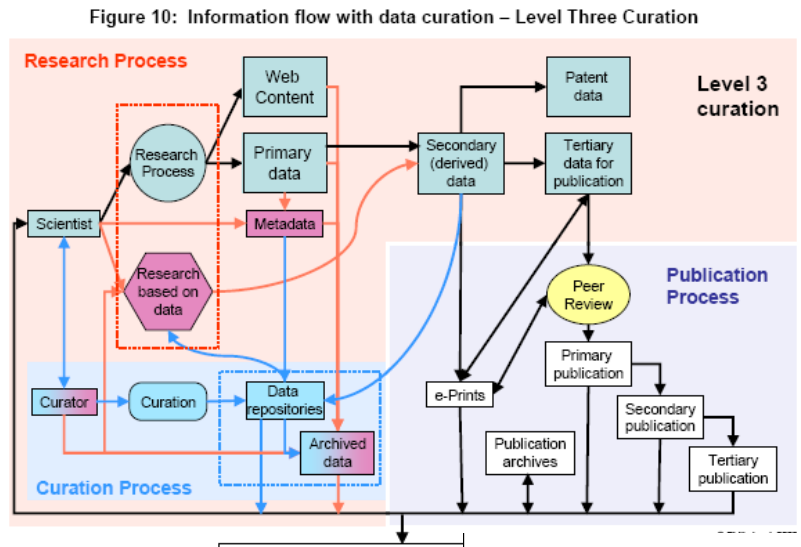
For some areas of research, some of these areas will be empty or irrelevant – that doesn't matter

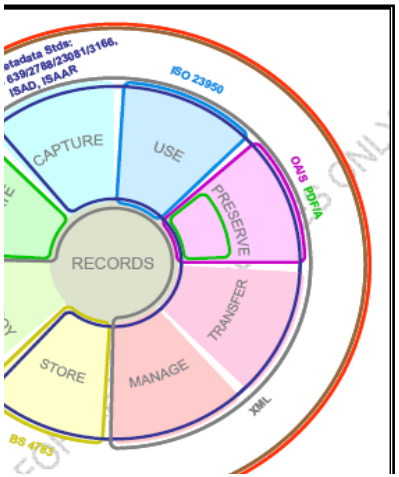
# The DCC Curation Lifecycle Model

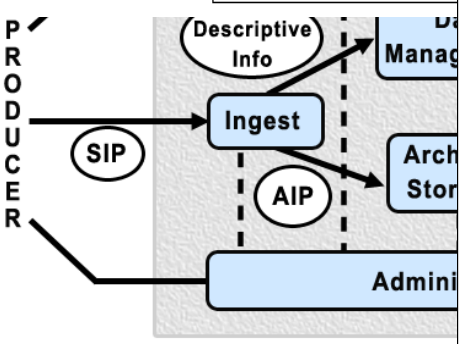


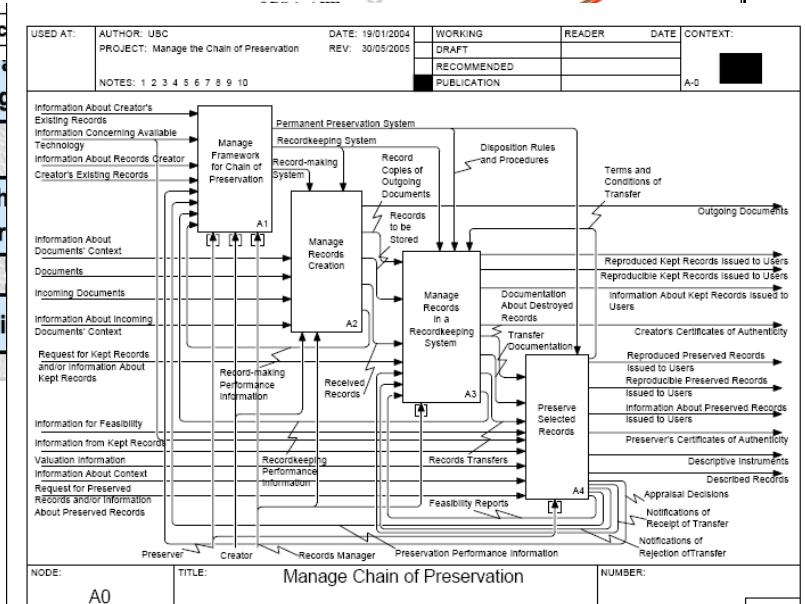
# Other lifecycles exist



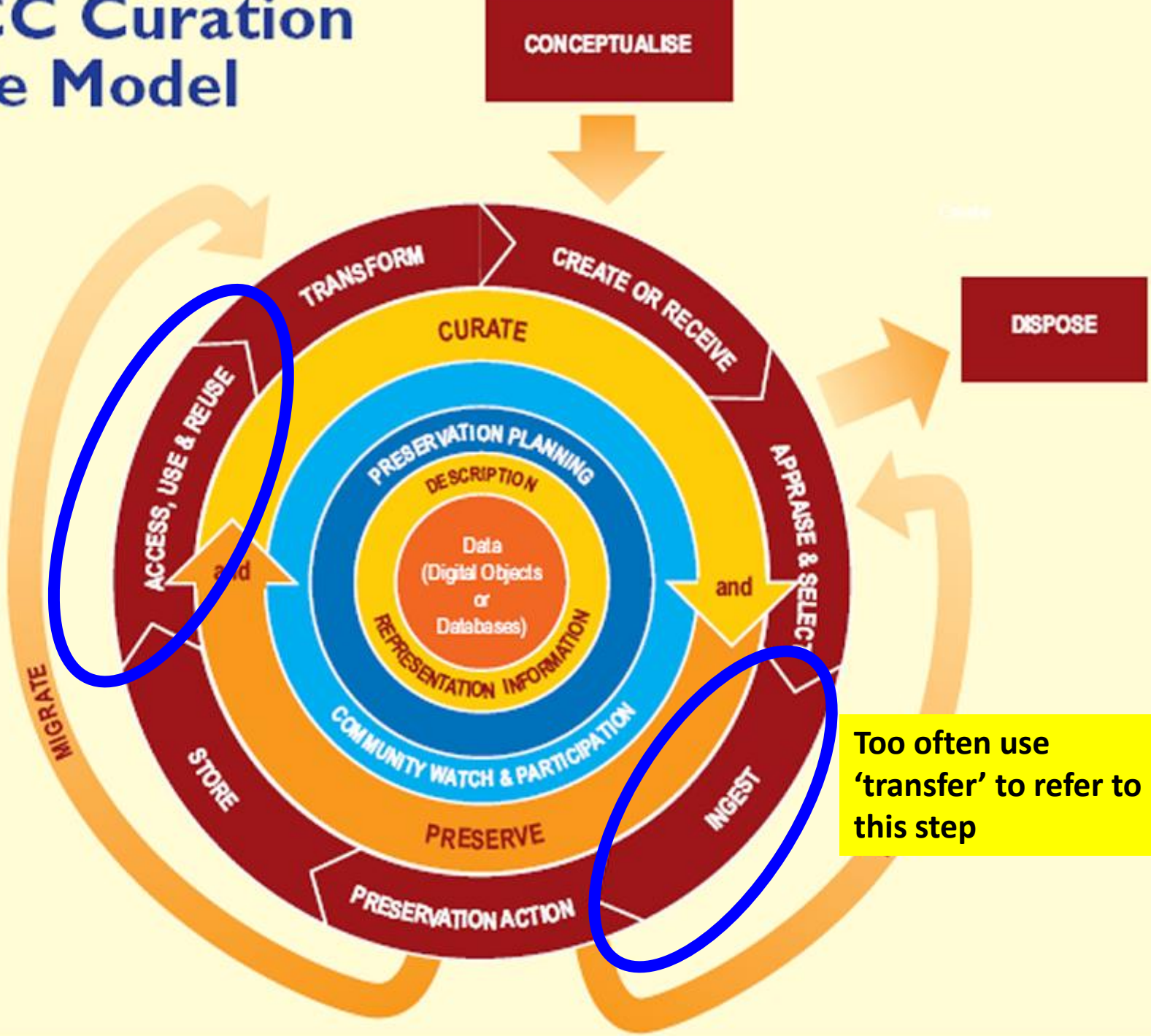



# The DCC Curation Lifecycle Model




Too often use 'transfer' to refer to this step

# Some general reflections

- Despite obvious differences between paper & digital, classifying material as ‘semi-current’ still sometimes used as a control mechanism
- Snapshots – of web sites, databases, etc – a common means of capturing semi-current material
- Change logs or transaction-based methods less common, but provide different insight
- Sometimes all is static; sometimes content is but metadata is not; sometimes all is changing

# Points to consider

- Why are we worried?
  - Loss through neglect/technology change?
  - Loss through inadvertent change?
  - Lack of access?
- What's changing, if anything?
- Is dual governance of copies achievable?
- What do we want to preserve – and how does this affect our actions?
- Are we concerned about preservation actions with real-world effects?