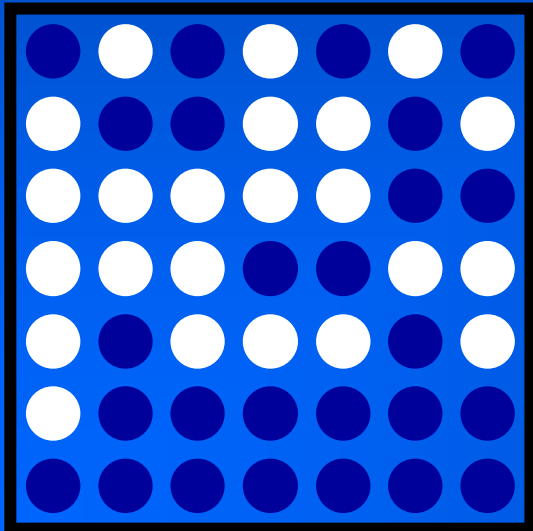


e-Science Curation

Digital Preservation Coalition Forum
London, 12th March 2003

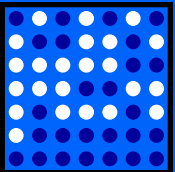


Philip Lord and Alison Macdonald
Digital Archiving Consultancy

alison@d-archive.co.uk
philip@phiplord.com

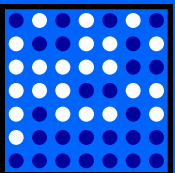
Presentation structure

1. Background, report brief
2. Findings
3. Core issues
4. Objectives
5. Recommendations



Background

- JISC Preservation Focus, DPC
- Denise Lievesley & Simon Jones 1998 report
- JISC 2002-2005 Preservation Strategy
- e-Science and the Grid – research is increasingly data-intensive, computing-intensive
- NSF Report, 2003



A rainy day in Edinburgh

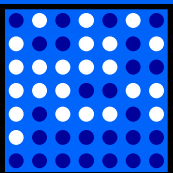


<http://umbriel.dcs.gla.ac.uk/NeSC/general/>

Report brief and methodology

- Report on current provision and future requirements:
- Primary research data, e-Science projects
- Across the seven Research Councils* and HE sector
- Methodology
 - Literature review
 - Questionnaires
 - Interviews
 - Task Force.

* BBSRC, CCLRC, EPSRC, ESRC, MRC, NERC, PPARC

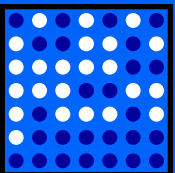


Defining “curation”

- Term adopted from biologists
- Used by John Taylor, DG of RCUK
- Digital Archiving Consultancy definition of digital curation:
 - “the continuing intellectual and content maintenance of data”

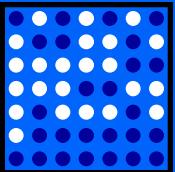
=

digital preservation + archiving + a bit more



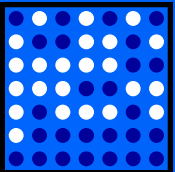
Questionnaires and interviews

- 275 data generators contacted
 - Two-stage process
 - Universities and Research Council units
- 62 librarians and data centre managers
- Response rates 17.5% and 21%
- 35+ interviews with a wide range of people



Some questionnaire findings: Data of value after project end

- 79% of data generator responses said their primary data would be of value after project end
- 90% said their summary/derived data would be of value after project end
- 94% said their published data would be of value after project end.

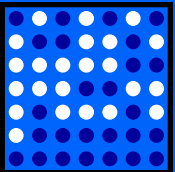


Some questionnaire findings (2)

Table percentages (over all responses)		Has financial provision been made for keeping data?		
		Yes	No	Don't know
Do the terms of your funding require data to be archived/preserved?	Yes	17%	21%	8%
	Some cases	-	8%	-
	No	4%	23%	4%
	Don't know	-	4%	10%

Some questionnaire findings (3):

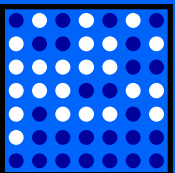
- Question: “Will future users need any of the following to use the data?”
 - Software: 42%
 - Hardware: 15%
 - Explanatory documentation: 75%



Virtual fossils need archiving too

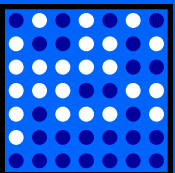
“Taxonomic convention requires that 'type' specimens of newly described species are curated in perpetuity. My work will inevitably result in 'virtual' type specimens of this sort.”

“...The Museum in [...] will take on curation of data for specimens figured in papers and for type specimens, but has no experience of digital curation and will probably expect to be able to just put the data on a CD and leave it in a cupboard”



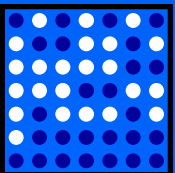
Several types of continuity problems

“....I have data files from projects from years ago which are on disks I no longer have a drive for on computers I no longer have access to or are no longer made or the software/operating system changes would make it extremely difficult to access any more. There are also problems that the nature of research work means a lot of short-term researchers over the years and a difficulty for a principal investigator to always keep definitive copies of all data plus backups. Also as PIs move around and collaborate with many people in other organisations it is pretty difficult to go back more than a few years with confidence that data will be adequately archived.”



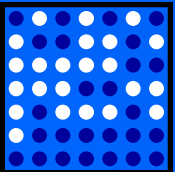
Fundamental requirements for data retention and curation

- Underlying continuity and stability
- Data of value needs to be captured in the first place
- The descriptive information needed for preservation needs to be provided in a timely fashion
- Repositories need to have adequate and appropriate skills and resources.



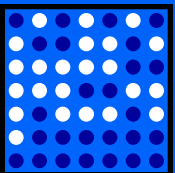
Core issues

- Criteria for keeping data?
- What is the value of keeping data?
- Who decides to keep data?
- Who keeps it where?
- How do you get descriptive information from the data creators?



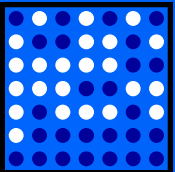
Some objectives

- Accessibility of data:
 - Curated holdings should be visible, as appropriate
 - Links between underlying datasets and publications
- Trust in data:
 - Provenance
- Curation should become part of the process of “doing science”.



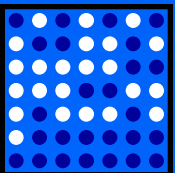
Some recommendations (1)

- Curation needs recognition at all levels – including the highest
- Overall responsibility needs to be assigned and a task force established
- The Research Councils should make curation an article of policy
- The Digital Curation Centre is urgently needed to provide generic services
- Discipline-specific centres are needed to support the user communities



Some recommendations (2)

- Exemplar research projects are needed to identify measurable benefits and rescue valuable data at risk
- Work should be co-ordinated:
 - With e-Library initiatives
 - Between and within the Research Councils
 - With supra-national bodies, specifically in Europe
- Institutions should be encouraged to adopt OAIS compliant repositories
 - The OAIS model needs examining in respect of dynamic datasets
- Training and awareness-raising are key



Organisational components for curation

(i)

Generic support services

(Tools, formats, technology watches, preservation planning)

(ii)

Discipline-related services and management

A

B

C

....

G

(iii)

Discipline-specific
curatorial support

(iv)

Access/discovery
tools

(iii)

Discipline-specific
curatorial support

Scientists /
researchers
(in producer role)

SIPs

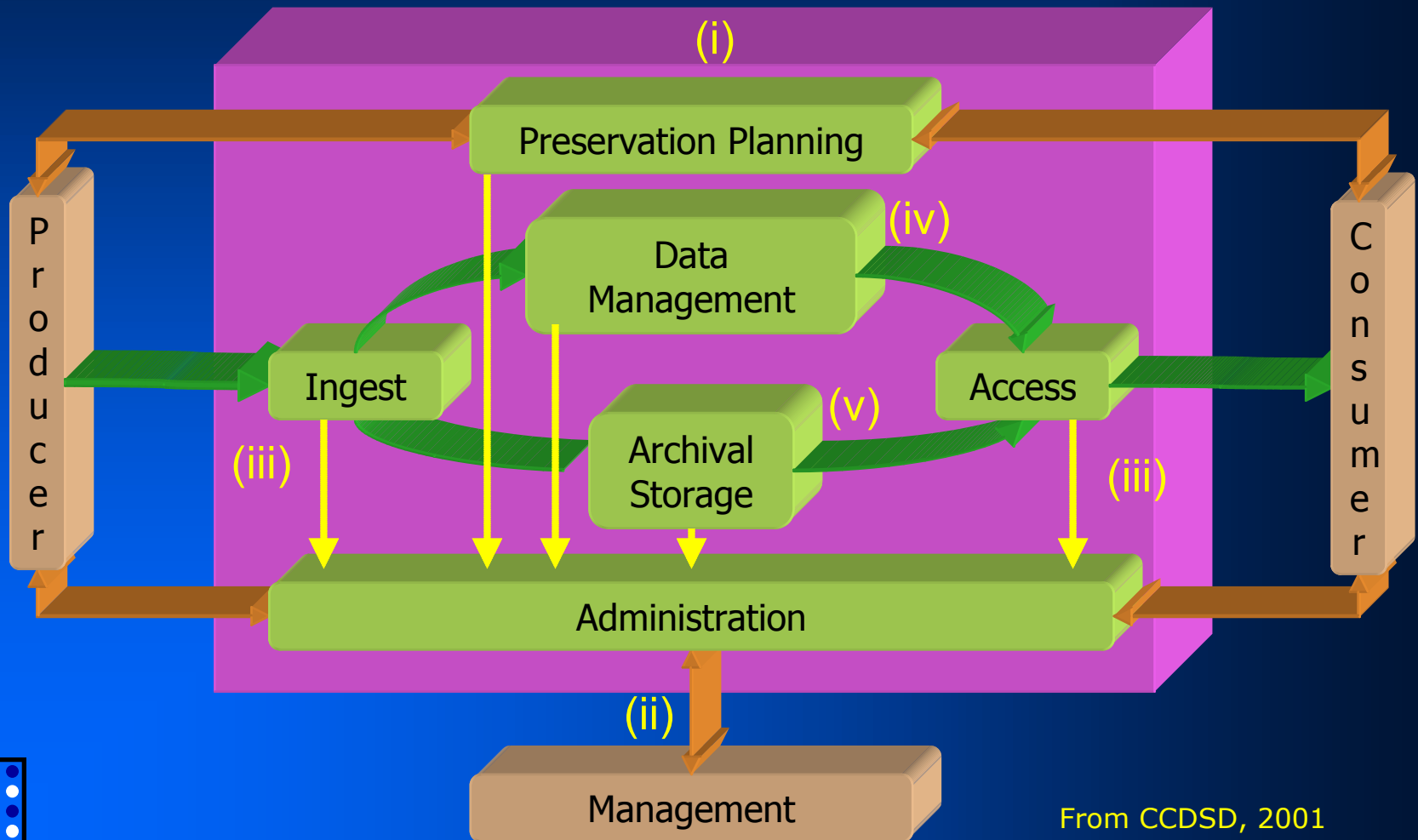
(v)

Storage
(AIPs)

DIPs

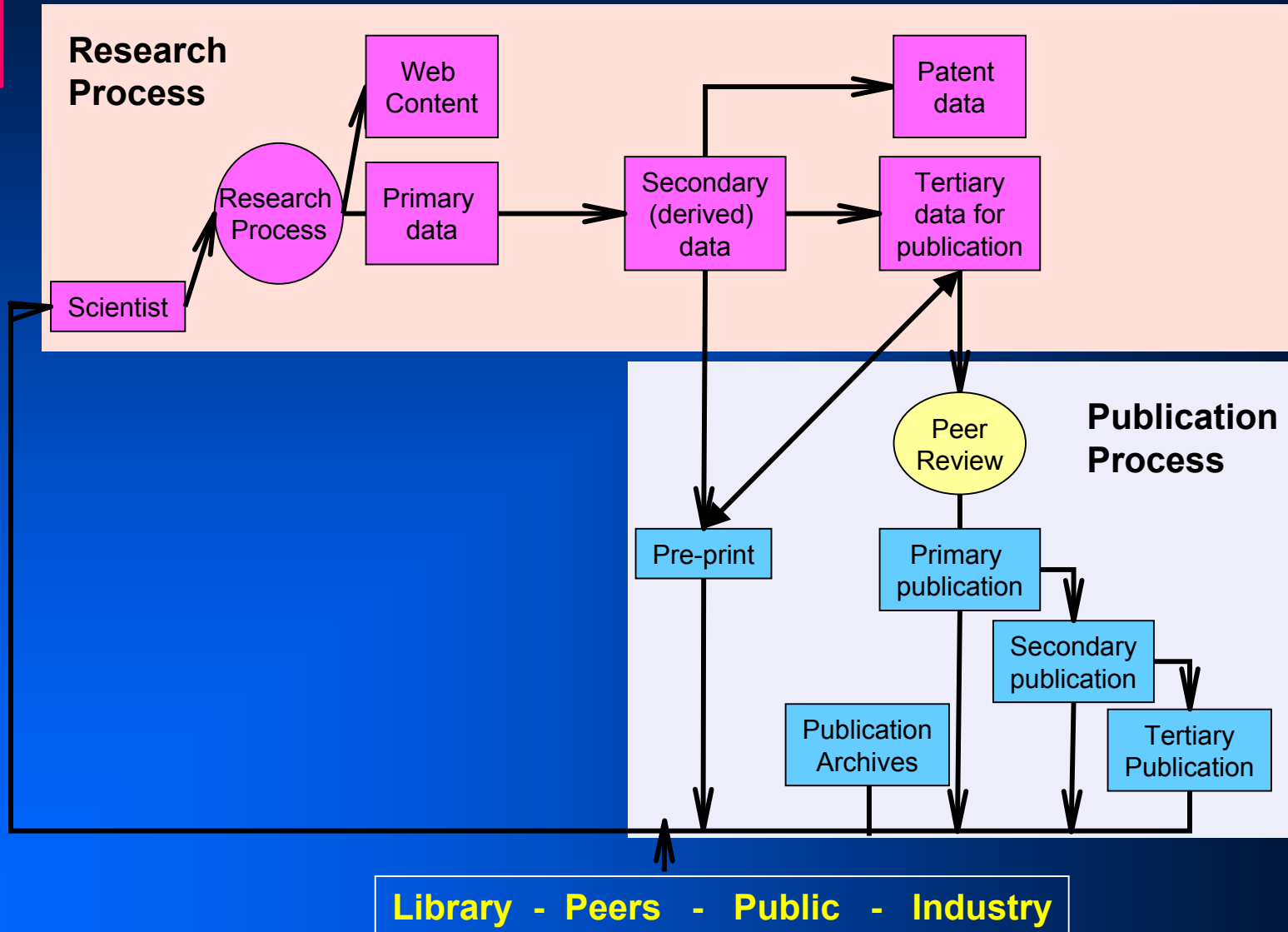
Scientists /
researchers
(in consumer role)

Digital archiving defined by OAIS

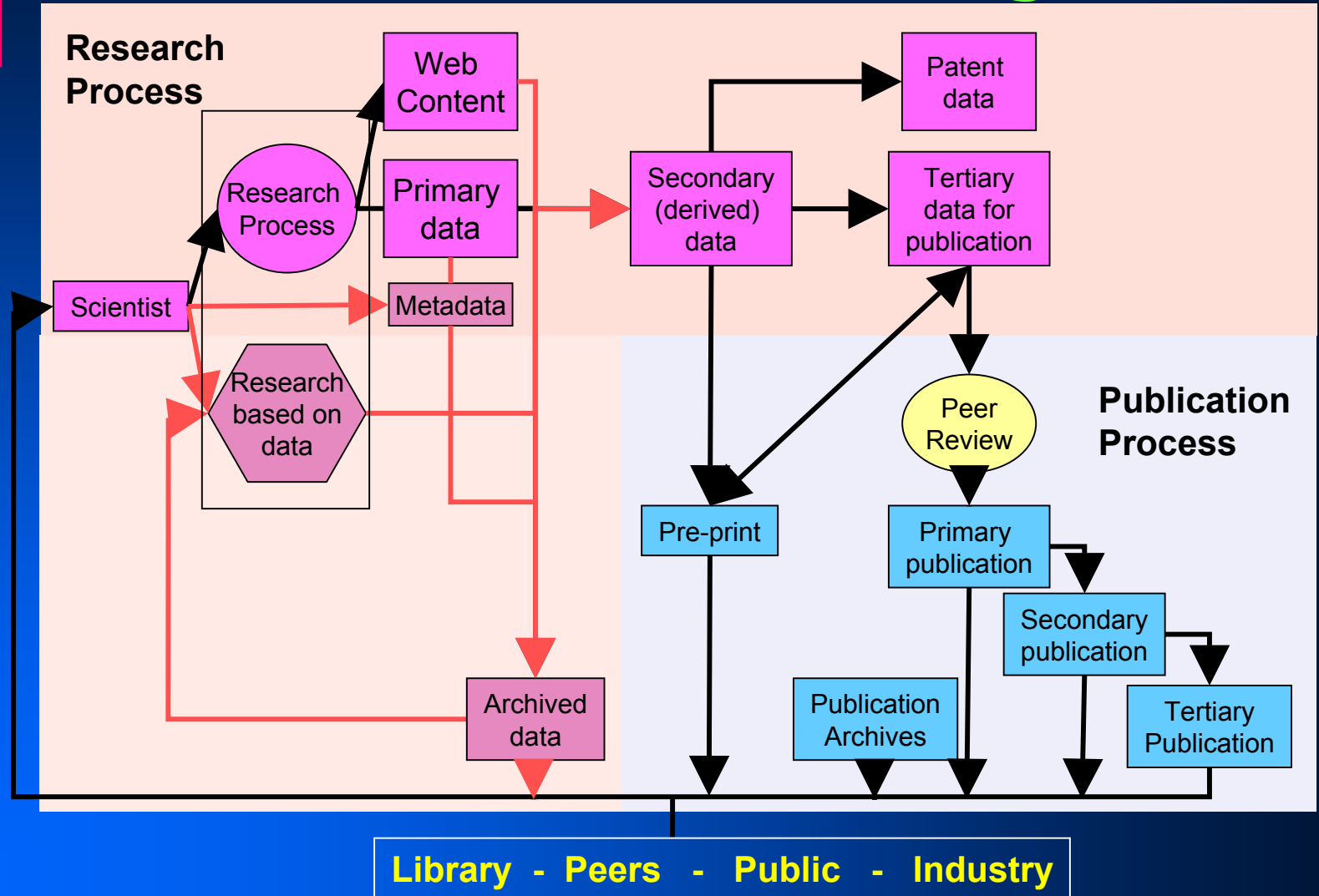


From CCDS, 2001

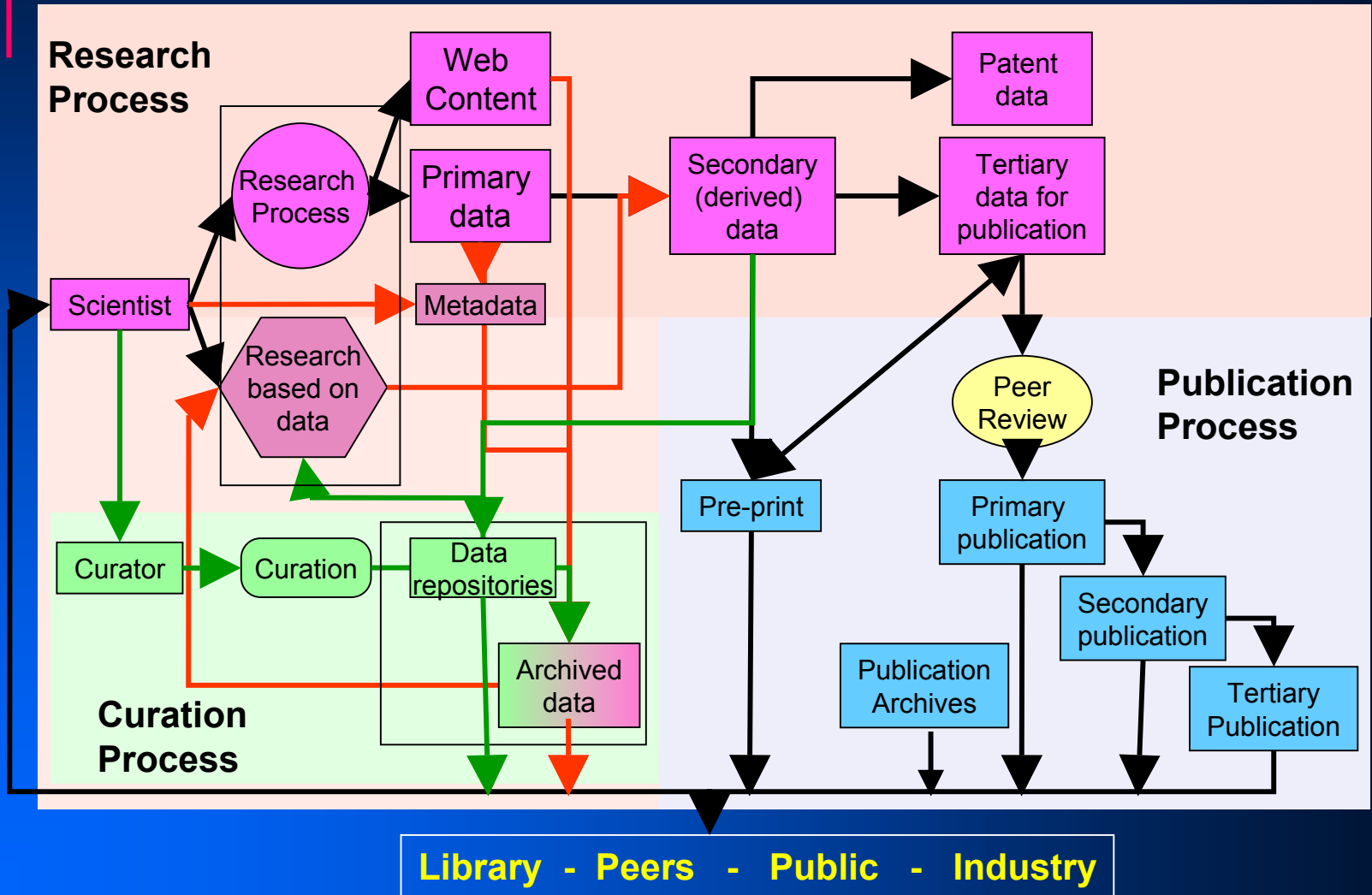
Research/publication process



... + Data archiving



... + curation



Contact

Digital Archiving Consultancy
2 Wayside Court
TWICKENHAM
TW1 2BQ
UK

 +44 (0)20 8607 9102

 +44 (0)70 5067 5010

philip@philiplord.com

www.philiplord.com