LIBER Webinar: Research Data – What To Keep?

WEBINAR HOSTS

Dr Birgit Schmidt

Head of Knowledge Commons

Göttingen State and University Library

bschmidt@sub.uni-goettingen.de

Rob Grim

Economics (Data) Librarian

Erasmus University Rotterdam

rob.grim@eur.nl

SPEAKER

Neil Beagrie

Director of Consultancy

Charles Beagrie

neil@beagrie.com

NOTES

- **The webinar is being recorded.** All participants will receive a link to the recording later today.
- **Slides are on Zenodo:** See the chat box for the link.
- **Questions?** Put them in the chat box. We'll put questions to the speakers at the end of the webinar.

Research Data - What To Keep?

Neil Beagrie
(Charles Beagrie Ltd)

LIBER Webinar March 2020
(Note these slides and recording available via LIBER!)

Webinar Aims

Attendees will learn about:

- Existing practice and guidance for appraisal and review for research data and related materials;
- Generic/disciplinary/sub-disciplinary differences in defining research data e.g. practice-based research in the arts and humanities;
- Research integrity and data sharing as strategic drivers for research data management;
- Differences in levels of curation and considerations of value and cost;
- The What to Keep report recommendations.

Scope

- **Based on UK Jisc-Funded *What* to Keep Study**
- **Definition of Research**
  - Taken from UK Research Excellence Framework (REF) Assessment framework and guidance on submissions
  - Includes Arts Practice (e.g. performance), "scholarship" [intellectual infrastructure of disciplines such as dictionaries or research databases]
- **Definition of Research Data**
  - Concordat on Open Research Data 2016
  - "the evidence that underpins the answer to the research question" mostly digital. RD term absent in some domains.
  - Study recognises significance of related material such as software, metadata and documentation, and physical samples

What to Keep
(May 2018-February 2019)

Project Aims

"provide new insights that will be useful to researchers, institutions, funders, publishers, and Jisc, on what research data to keep and why, the current position, and suggestions for improvement"

Methodology

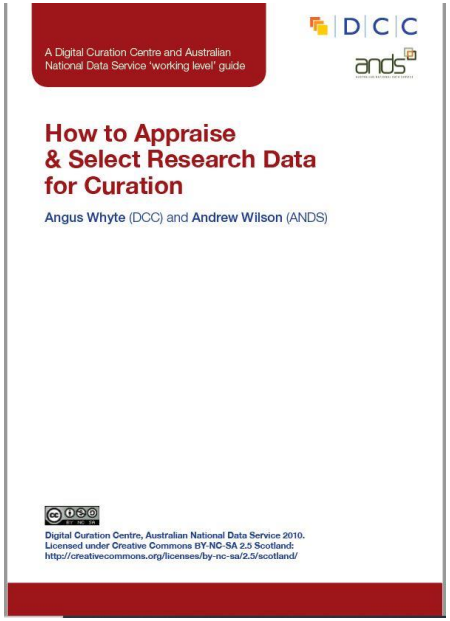
desk research, case studies, interviews, analysis, and a stakeholder workshop. Covering:

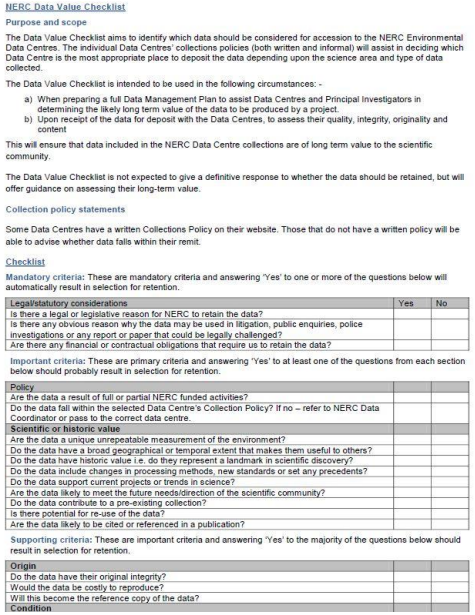
- the main academic research areas in the UK
- a range of disciplinary maturity in terms of policy and practice for what to keep;
- a range of outputs potentially including not only data but software and documentation required for its future use;
- differing requirements for location of data such as domain (international/national disciplinary) repositories, institutional repositories, or instrument data repositories; and
- different reasons for what to keep.

Previous Work

- FAIR (Findable, Accessible, Interoperable, and Reusable) initiatives (e.g. Turning FAIR into reality 2018)
- Research Integrity and Reproducibility initiatives (e.g. the Transparency and Openness Promotion (TOP) guidelines for Journals 2015)
- Detailed acquisition guidelines in many domain repositories
- Previous influential guidance documents include the NERC Data Value Checklist; the DCC/ANDS Guide to How to Appraise and Select Research Data for Curation

Previous Guidance





NERC Data Value Checklist

Purpose and scope

The Data Value Checklist aims to identify which data should be considered for accession to the NERC Environmental Data Centres. The individual Data Centres' collections policies (both written and informal) will assist in deciding which Data Centre is the most appropriate place to deposit the data depending upon the science area and type of data collected.

The Data Value Checklist is intended to be used in the following circumstances: -

a) When preparing a full Data Management Plan to assist Data Centres and Principal Investigators in determining the likely long term value of the data to be produced by a project.
b) Upon receipt of the data for deposit with the Data Centres, to assess their quality, integrity, originality and content

This will ensure that data included in the NERC Data Centre collections are of long term value to the scientific community.

The Data Value Checklist is not expected to give a definitive response to whether the data should be retained, but will offer guidance on assessing their long-term value.

Collection policy statements

Some Data Centres have a written Collections Policy on their website. Those that do not have a written policy will be able to advise whether data falls within their remit.

Checklist

Mandatory criteria: These are mandatory criteria and answering 'Yes' to one or more of the questions below will automatically result in selection for retention.

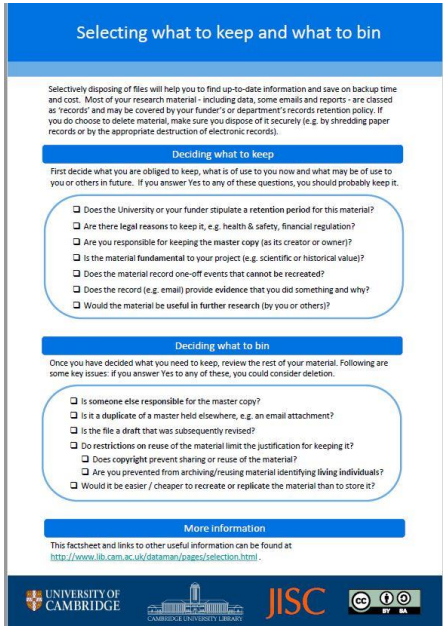
Legal/statutory considerations	Yes	No
Is there a legal or legislative reason for NERC to retain the data?		
Is there any obvious reason why the data may be used in litigation, public enquiries, police investigations or any report or paper that could be legally challenged?		
Are there any financial or contractual obligations that require us to retain the data?		

Important criteria: These are primary criteria and answering 'Yes' to at least one of the questions from each section below should probably result in selection for retention.

Policy		
Are the data a result of full or partial NERC funded activities?		
Do the data fall within the selected Data Centre's Collection Policy? If no – refer to NERC Data Coordinator or pass to the correct data centre.		
Scientific or historic value		
Are the data a unique unrepeatable measurement of the environment?		
Do the data have a broad geographical or temporal extent that makes them useful to others?		
Do the data have historic value i.e. do they represent a landmark in scientific discovery?		
Do the data include changes in processing methods, new standards or set any precedents?		
Do the data support current projects or trends in science?		
Are the data likely to meet the future needs/direction of the scientific community?		
Do the data contribute to a pre-existing collection?		
Is there potential for re-use of the data?		
Are the data likely to be cited or referenced in a publication?		

Supporting criteria: These are important criteria and answering 'Yes' to the majority of the questions below should result in selection for retention.

Origin		
Do the data have their original integrity?		
Would the data be costly to reproduce?		
Will this become the reference copy of the data?		
Condition		



6 guidance documents: the DCC/ANDS Guide to How to Appraise and Select Research Data for Curation (2010); Cambridge PrePare Selecting what to keep and what to bin (2012); the NERC Data Value Checklist (2013-2015); Five steps to decide what data to keep: DCC Checklist for Appraising Research Data (2014); UK Data Service Collections Development Selection and Appraisal Criteria (2018); University of Bristol Research Data Evaluation Guide (2018)

The WTK Analysis

Charles Beagrie

Table 3 – Data appraisal and selection criteria mapped from *what to keep* and other sources

Mapped criteria	WTK Optimal Data to Keep	NERC checklist	DCC 5 steps	Bristol checklist	PrePARe checklist
1. Relevant to mission					
1.1 Funder requirement	✓ (IIIIIIIII)	✓ (mandatory)	✓ (step 2)	✓ (mandatory)	✓ (probably keep)
1.2 Potential for reuse	✓ (IIIIIIII)	✓ (important)	✓ (step 1)	✓ (important)	✓ (probably keep)
1.3 Legal requirement	✓ (II)	✓ (mandatory)	✓ (step 2)	✓ (mandatory)	✓ (probably keep)
1.4 Publisher requirement	✓ (III)		✓ (step 2)	✓ (mandatory)	
1.5 Data to substantiate research publications & findings	✓ (IIIIIIII)	✓ (important)	✓ (step 1)	✓ (mandatory)	✓ (probably keep)
1.6 Degree of Openness	✓ (IIIIII) ✓ (supporting)	✓ (supporting)	✓ (step 2) ✓ (step 3)		
1.7 Uniqueness	✓ (IIIIIIII)	✓ (important)	✓ (step 3)	✓ (important)	✓ (probably keep)
1.8 Time Series/ Aggregate collection	✓ (III)	✓ (important)		✓ (important)	
1.9 From specified creator	✓ (IIIIII)	✓ (important)			
1.10 Intangible Cultural Heritage	✓ (IIIII)				

Appendices – Table 3 (extract only!)

The 7 WTK Case Studies

- 8.1. FUNDER ALL DISCIPLINES (UK RESEARCH AND INNOVATION)
- 8.2. DOMAIN REPOSITORY ARCHAEOLOGY (ARCHAEOLOGY DATA SERVICE)
- 8.3. DOMAIN REPOSITORY SOCIAL SCIENCES (UK DATA SERVICE)
- 8.4. INSTRUMENT REPOSITORY SCIENCE AND TECHNOLOGY (SCIENCE AND TECHNOLOGY FACILITIES COUNCIL)
- 8.5. UNIVERSITY RESEARCH DATA SERVICE (UNIVERSITY OF BATH)
- 8.6. RESEARCHER PRACTICE-BASED ARTS AND HUMANITIES (PRACTICE RESEARCH ADVISORY GROUP)
- 8.7. PUBLISHER SCIENCE TECHNOLOGY MEDICINE (INTERNATIONAL ASSOCIATION OF STM PUBLISHERS)

The WTK Case Studies

8.1. FUNDER ALL DISCIPLINES (UK RESEARCH AND INNOVATION)

- The optimal research data to keep:
  - Data which support primary research findings, e.g. are necessary to reproduce or query those findings…
  - Data that is of obvious long-term value e.g. longitudinal studies...
  - Data which is subject to legal requirements...
  - Some data with short term value for one purpose or set of users can also have long term value for other purposes or users…
  - Quality of data...
  - Data suitable for reuse…
- Some questions remain around:
  - Instrumentation data..., Outputs from models and simulations…, Serendipity…,"Curated Databases"…
- Regarding supplementary data and materials, we should keep:

  Metadata...,Some software/algorithms/codes support data reproduction or interpretation...,Physical materials...

The WTK Case Studies

8.6. RESEARCHER PRACTICE-BASED ARTS AND HUMANITIES (PRACTICE RESEARCH ADVISORY GROUP)

We are starting from a position where almost nothing is available...

The term “research data” is anathema to many colleagues working in these areas. They would not recognise the term and are more likely to use colloquial expressions such as “stuff”...

One problem is people see their projects as open-ended. Without the imperative to complete research projects for publication, arts practitioners want to keep exploring and extending the work...

Some Key Findings

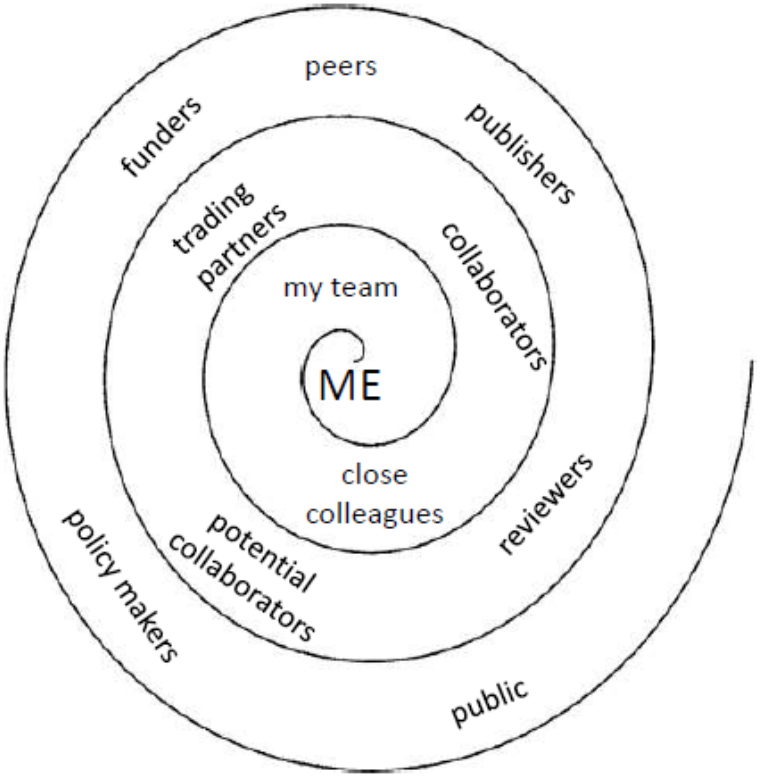
- WTK Study interviewees identified some 33 different factors or elements to consider in appraisal and selection of What to Keep and some 17 reasons for why data may be kept
- The suggested factors or elements for what to keep from interviewees can be mapped to existing check lists
- Researchers role in WTK appraisal and selection within universities - guided by curators (different emphasis in domain repositories)
- It is essential to consider not only What to Keep, but for How Long to keep it, Where to keep it, and increasingly How to keep it [...+ who funds, etc]

Perceptions

Carole Goble - Open Science: how to serve the needs of the researcher? -Jisc/CNI July 2018

Staged Open (Access) Spiral & Benefits

organisation – collaboration - dissemination



The number of assets reduces

The richness of metadata needed increases

As reach of sharing increases and burden of work increases

Staged sharing

Levels of Curation

Charles Beagrie

US National Science Board 2005

Long-lived Data Collections:

Research data collections

Resource/community data collections

Reference data collections

Core Trust Seal Requirements June 2018

Level of Curation Performed:

A. Content distributed as deposited

B. Basic curation – e.g., brief checking, addition of basic metadata or documentation

C. Enhanced curation – e.g., conversion to new formats, enhancement of documentation

D. Data-level curation – as in C above, but with additional editing of deposited data for accuracy

The WTK Case Studies

8.3. DOMAIN REPOSITORY SOCIAL SCIENCES (UK DATA SERVICE)

Policy and appraisal help to determine the appropriate service solutions offered by UKDS: ReShare or the UKDS main catalogue:

1. Data collections selected for long-term curation. These data will have long term secondary analysis potential (the UKDS main catalogue);

2. Data collections selected for “short-term” management. These collections will not initially be retained for long-term preservation (but may be moved to category one in the future). They will be backed-up i.e., bit-level preservation only (ReShare).

UKDS broadly focusses on (CTS) curation level C in the UKDS main collection and curation level A in ReShare.

Recommendations

- Recommendation 1: Consider what is transferable in terms of effective practice in what to keep between disciplines. Support adoption of generic effective practice...
- Recommendation 2: Support workshops to bring communities together to evolve disciplinary norms for what to keep for their research data where these are currently absent or evolving.
- Recommendation 3: Seek to harmonise funder requirements for research data where relevant, e.g. where the data type is the same but the funders and their requirements differ.
- Recommendation 4: Investigate the relative costs and benefits of differential curation levels, storage, or appraisal for what to keep...
- Recommendation 5: Apply the FAIR principles and the Open Research Data Concordat principles, as appropriate, to kept data….

Current UK Position
(as of February 2019)

Some Key Findings

- The current state is highly varied: in a few rare cases where there is an end-to-end process and all outputs are in a single repository, the current state is categorically known. In the majority of cases it is known when deposited in a central repository in the organisation but largely unknown when deposited elsewhere.
- It is currently often difficult to establish from sources such as research publications or grant databases which research projects have generated research data and therefore what the total population maybe, or where the data is.
- The broad picture from the interviews is not contradicted by the desk research and a number of surveys at specific points in time in recent years at local level, nationally, or internationally.
- Interviewees expressed different views on the potential utility of ResearchFish, institutional current research information system (CRIS) systems, or Data Management Plans (DMPs), to give a better picture of the current state of research data.

Recommendations

- Recommendation 6: Enhance data discoverability and enable unambiguous identification of what has been kept…
- Recommendation 7: Require Data Access Statements (alternatively referred to as a ‘data availability statement’) in published research articles and encourage adoption of the Transparency and Openness Promotion (TOP) guidelines, created by journals, funders, and societies to align scientific ideals with publication practices.

Shortfalls and Suggestions

Some Key Findings

- Interviewees identified some 15 different areas where there are current shortfalls and made some 26 different suggestions for improvements
- Growing Data Volumes (More selection, More funding, Tiered storage, Costing more accurately, Improve discovery/access)
- Not following grant conditions for sharing (Improve incentives, More publisher and funder collaboration, audit, sanctions, advocacy, training, automation of deposit workflow, one to one guidance)
- "*Maybe funders and publishers should collaborate more. They could encourage the right behaviours If they worked as a team at the only two points in the research lifecycle when researchers are incentivised – when they get the grant and they when they get published*"
- Costs seen as too high (Better demonstration of value, Develop guidance, Sustainability models - budgeting for time and effort needed)

Recommendations

- Recommendation 6: Improve the recording of, and ability to identify, data generated by UK research projects in existing research databases and DMPs…
- Recommendation 7: Require Data Access Statements (alternatively referred to as a 'data availability statement') in published research articles…
- Recommendation 8: Improve incentives for data sharing
- Recommendation 9: Increase publisher and funder collaborations around research data
- Recommendation 10: Improve communication on what research data management costs can be funded and by whom...

Future Value

The WTK Report

https://repository.jisc.ac.uk/7262/

CESSDA-SaW Cost-Benefit Advocacy Toolkit

Valuable tools for thinking about future cost and benefits of research data (particularly the 3 Factsheets: Benefits, Costs, and Return on Investment)

http://dx.doi.org/10.18448/16.0013

Thank you!

neil@beagrie.com

THANKS!

Questions?

Please put them in the chat box.

Slides and a recording will be sent to all registered delegates.