because good research needs good data

Data selection & licensing

NTU workshop March 2017

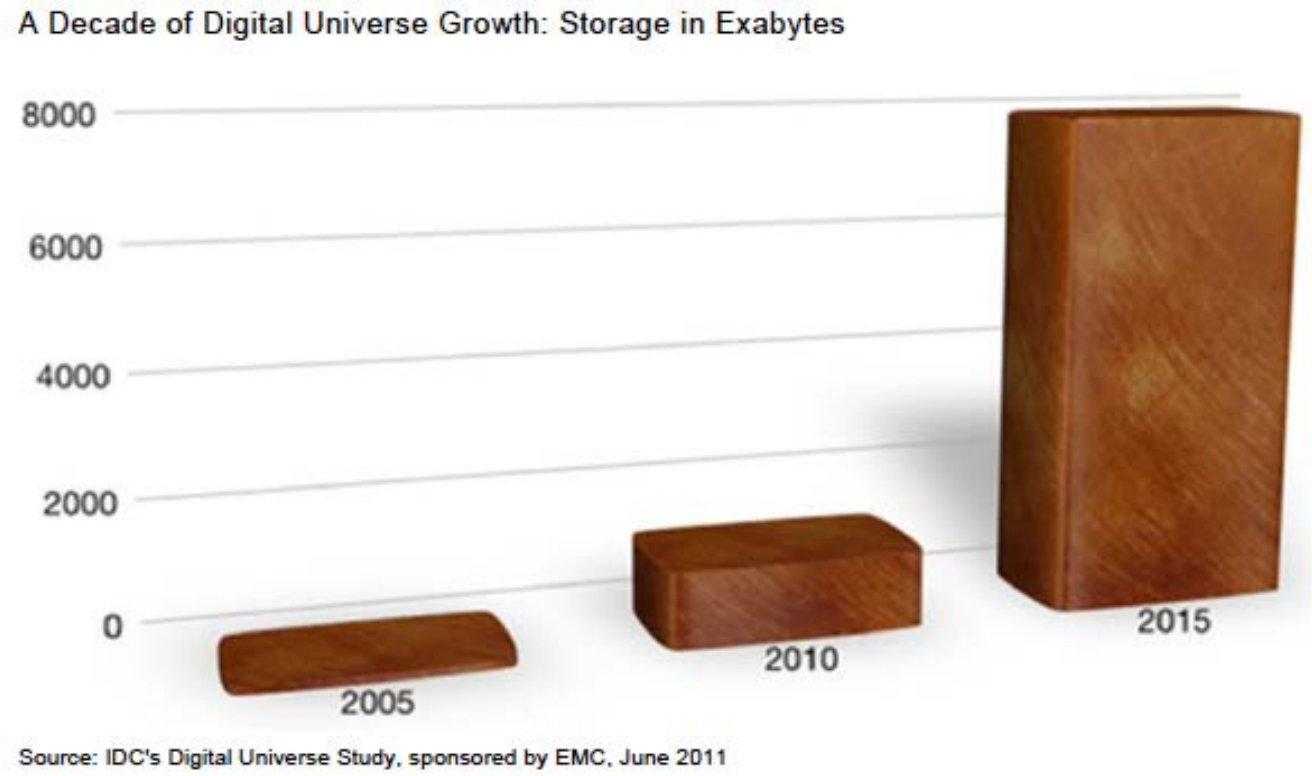
Digital Curation Centre

Outline

- Why selection is important
- How to select data
- Licensing data – options and implications

Why not keep it all?

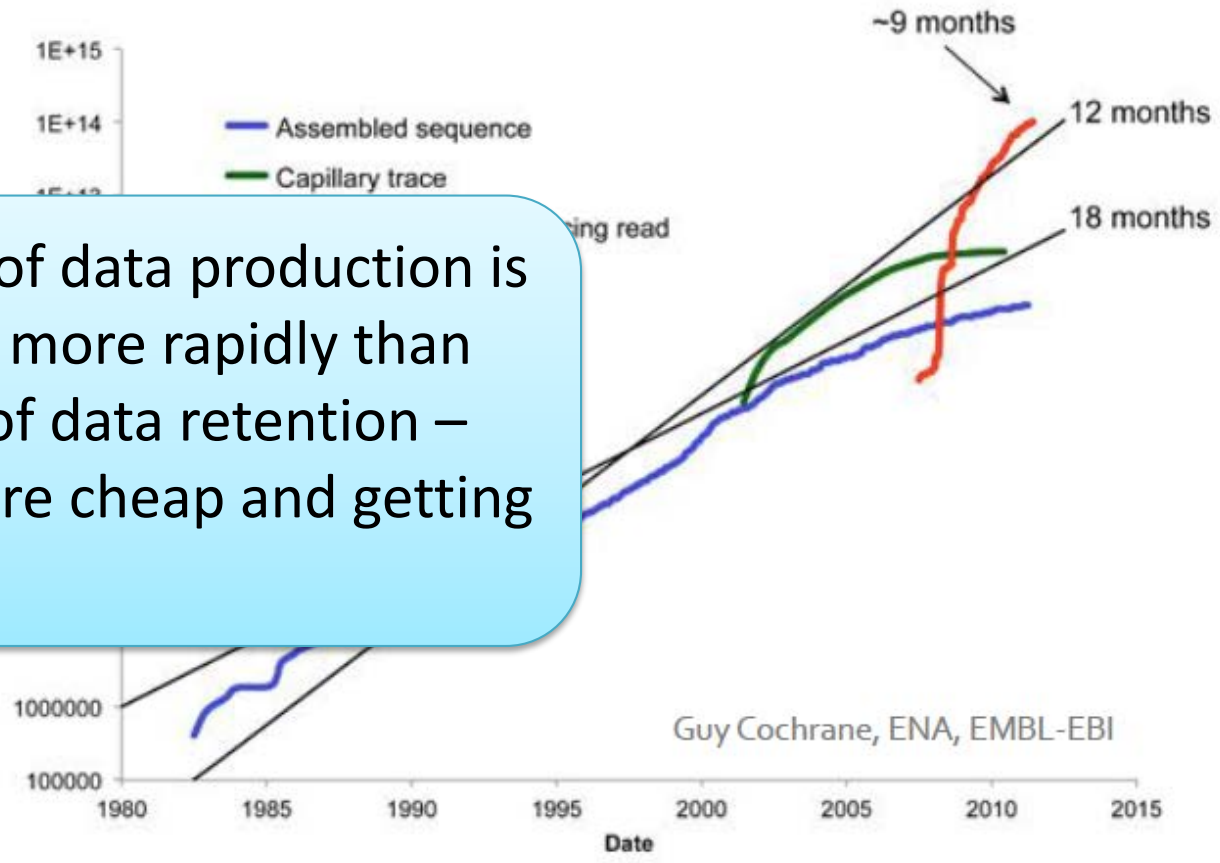
Globally, data volumes are doubling every two years



John Gantz and David Reinsel 2011 ***Extracting Value from Chaos*** *www.emc.com/digital_universe.*

Data volumes escalate

Volumes rising faster in data-intensive research domains
e.g. DNA sequence data is doubling every 6-8 months

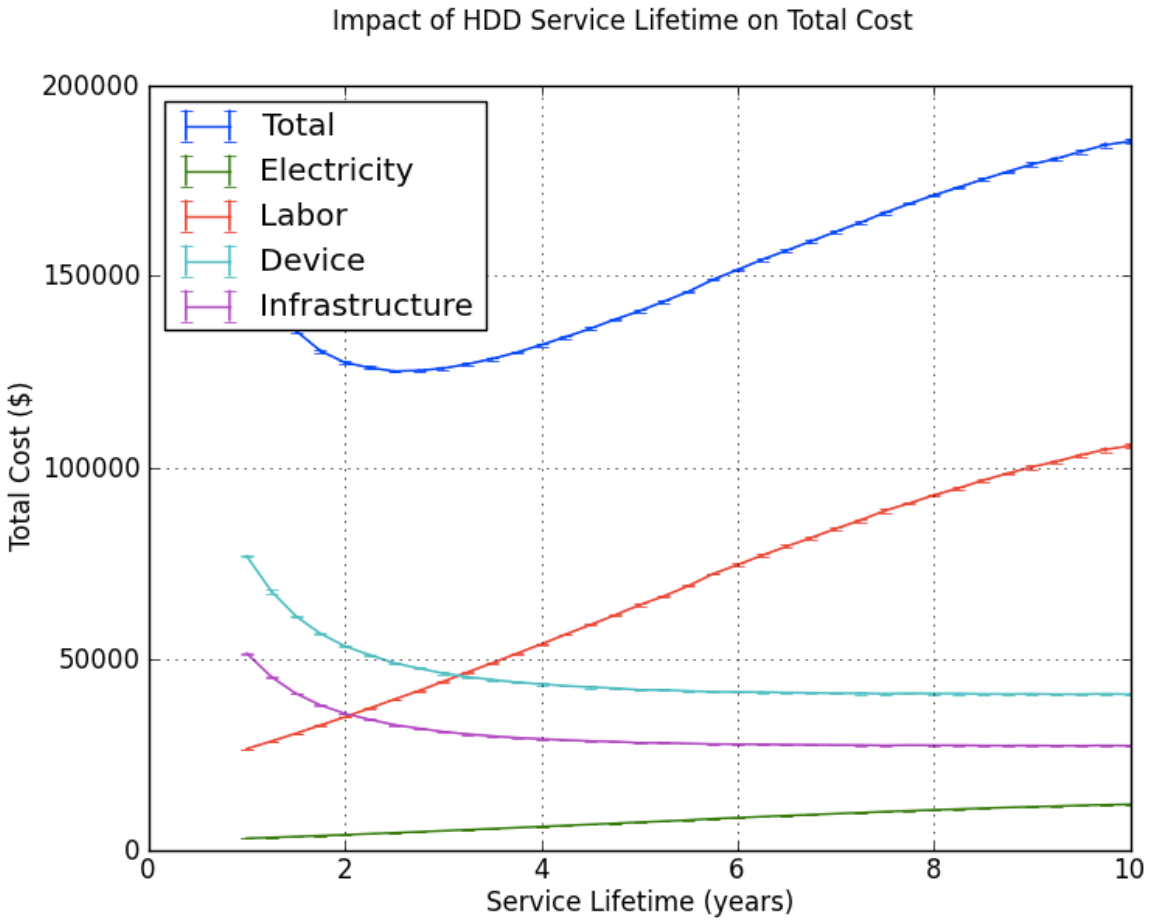


The cost of data production is dropping more rapidly than the cost of data retention – sensors are cheap and getting cheaper

"ELIXIR and Open Data" View from an ELIXIR Node" Barend Mons, ELIXIR Launch event, 18th Dec 2013

Storage mgmt costs rise long-term

Hardware costs decline, but power and staff costs keep rising



David Rosenthal blog.dshr.org/2012/05/lets-just-keep-everything-forever-in.html

The storage is cheap fallacy

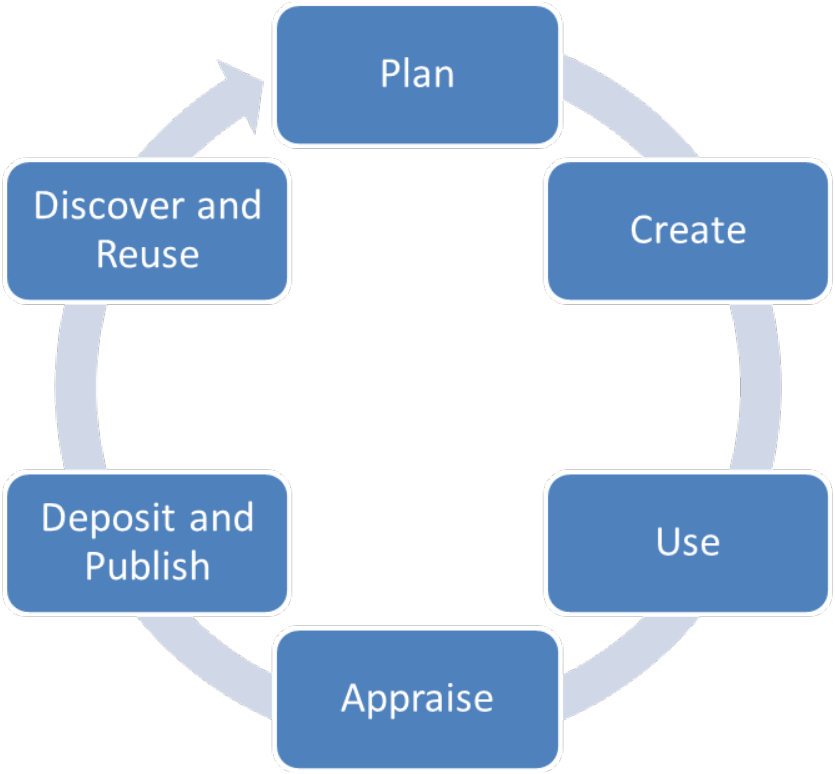
- Decreasing hardware costs offs volume growth
- Backup and mirroring multiplies ost of preserved data
- Discovery becomes harder as chaff outweighs the wheat
- Curation of unused data is a waste of resources

But discovery techniques are getting better – this argument is not clear-cut

This is a significant factor – curation is often manual

When should selection begin?

Appraisal should begin as early as possible!



Periodically for longitudinal and reference datasets

What questions must be answered?

1. What must be kept to manage compliance risk? (and what must NOT be kept)
2. What data has value and should be kept?
3. What data could be re-used?
4. Given costs what will or won't be kept?
5. How will it be kept and shared, on what terms?

What ‘must’ be kept?

Some data may be part of research record, evidence for e.g. …

- Supporting patent applications or IP
- Evidence of investigations or inquiries
- Health & Safety (Lab book)

Compliance also about data that won’t be kept, or may only be shared with approved researchers…

Other drivers for what 'must' be kept

Funder policies:

"Data with acknowledged long-term value "

RCUK Common Principles on Data Policy

Journal policies:

… a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols promptly available to readers… (*http://www.nature.com/authors/policies/availability.html*)

The Law

Appraisal and deposit

How to Appraise & Select Research Data for Curation
Angus Whyte, Digital Curation Centre, and Andrew Wilson, Australian National Data Service (2010)

1. **Relevance to Mission** – including any legal/funder requirement to retain the data beyond its immediate use.
2. **Scientific or Historical Value** – significance and relationship to publications etc.
3. **Uniqueness** – can it be found elsewhere / if we don't preserve it, who will?
4. **Potential for Redistribution** – quality / IP / ethical concerns are addressed.
5. **Non-Replicability** – either impossible to replicate (e.g. atmospheric or social science data) or not financially viable.
6. **Economic Case** – costs of managing and preserving the resource stack up well against potential future benefits.
7. **Full Documentation** – surrounding / contextual information necessary to facilitate future discovery, access, and reuse is adequate.

Step 2 What data *should* have value

Indicators that data have value

1. Quality of the data and its description

   complete, accurate, reliable, valid, representative etc

2. Demand high

   known users, integration potential, reputation, recommendation, appeal

3. Replication difficulty

   difficult, costly, or impossible to reproduce

4. Low barriers

   legal/ ethical, copyright non-restrictive terms and conditions

5. Rarity

   unique copy or other copies at risk

Which related material does data depend on for its value?

Step 3 What *could* it be reused for?

Step back and reflect – typical reuse purposes

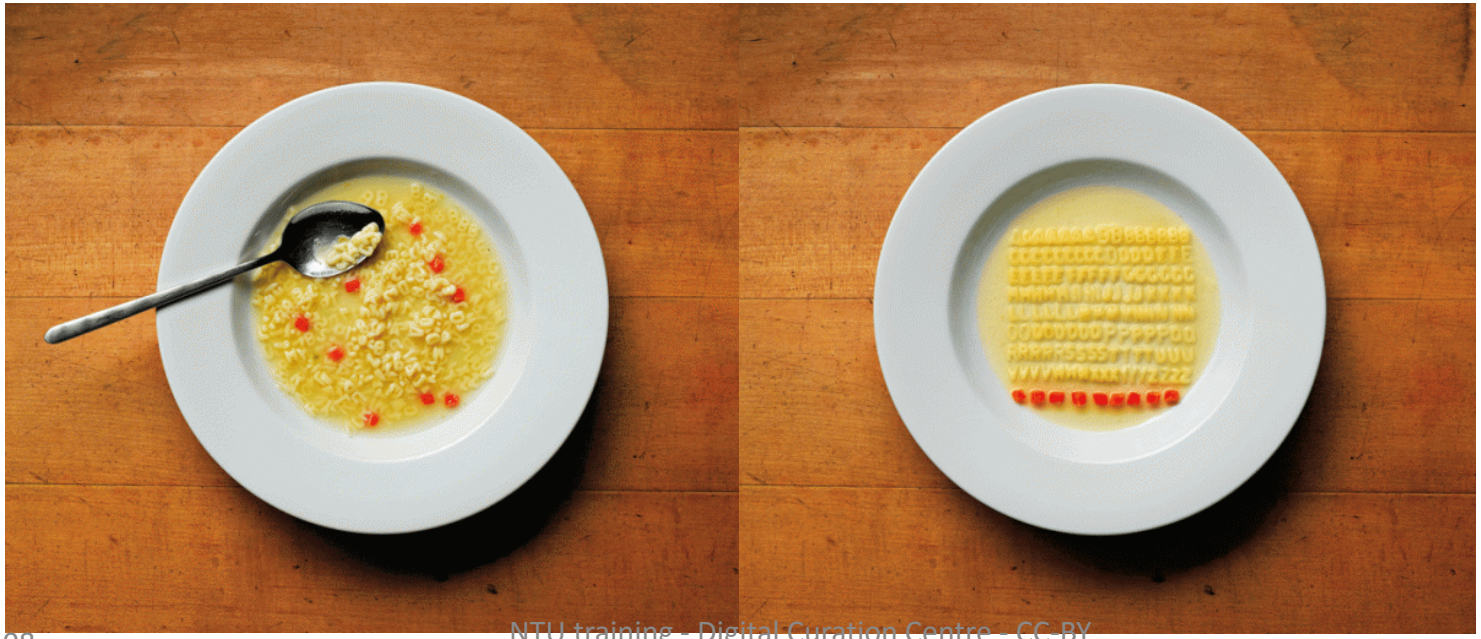
1. Verification
2. Further analysis
3. Reputation building
4. Resource development
5. Further publications inc. data articles
6. Learning and teaching materials
7. Private reference

Then relative to these, which data **must** be kept and which data and related materials will have significant value?

Step 4 Cost factors

Consider these when deciding what to keep

- Costs incurred during project may add to the data's value
- Cost of curation – to make data reusable
- Cost of long-term retention can be calculated



costed?

Who should help appraise?

RLUK 'skills gaps' survey of Subject Librarians & Managers

" …nine key areas where future involvement by Subject Librarians is considered to be important now and is also expected to grow sharply…

1. Ability to advise on preserving research outputs (49% see as essential in 2-5 years; 10% now)

2. Knowledge to advise on data management and curation, (48% essential in 2-5 years; 16% now)…"

Mary Auckland 2012 Reskilling Libraries for Research

What data to keep

Roles and Responsibilities

A Digital Curation Centre and Australian National Data Service 'working level' guide

DCC

ands

How to Appraise & Select Research Data for Curation

Angus Whyte (DCC) and Andrew Wilson (ANDS)

Researcher ('data creator')

- Provide enough information for others to assess the research data's scientific and scholarly quality and compliance with disciplinary or ethical norms.
- Provide relevant information for the repository to identify who will use the data and how i.e. the 'designated community', and any specific access requirements or constraints.
- Provide the research data in formats recommended by the data repository.
- Provide the metadata requested by the repository.

Data centre or repository

- Make explicit its mission in the area of digital archiving, and its selection policy for digital objects.
- Ensure compliance with legal regulations and contracts.
- Ensure the authenticity and integrity of the digital objects and the metadata.
- Assume responsibility from the data producer for ensuring the digital objects are accessible and available to a defined 'designated community'.
- Plan for long-term preservation of the digital assets.

Who else?

Others who may be involved in appraising research data…

- Domain specialists
- Archives
- Research Office- Business development
- IT Support/ Research Computing
- Research Ethics Committee
- Records Management/ FOI Compliance
- Facilities Managers (if physical samples involved)

LEGAL ISSUES & LICENCING

Two types of legal issue

Things that the law requires you to consider

Things that the law allows you to do

Consequences for researchers - requirements

- Your organisation must know what data it possesses
- It must know whether exceptions to access may apply
- It must know if some of the data belongs to others
- It must know what data once existed, but has now been deleted – and why
- These are difficult questions for most of us!

So document those appraisal decisions!

Requirements

- Data protection
  - If human subjects are involved
  - Common European framework – Singapore has similar requirements
  - Informed consent essential
  - Make consent broad to allow reuse
  - Protect data
  - Provide subject access
  - Right of correction

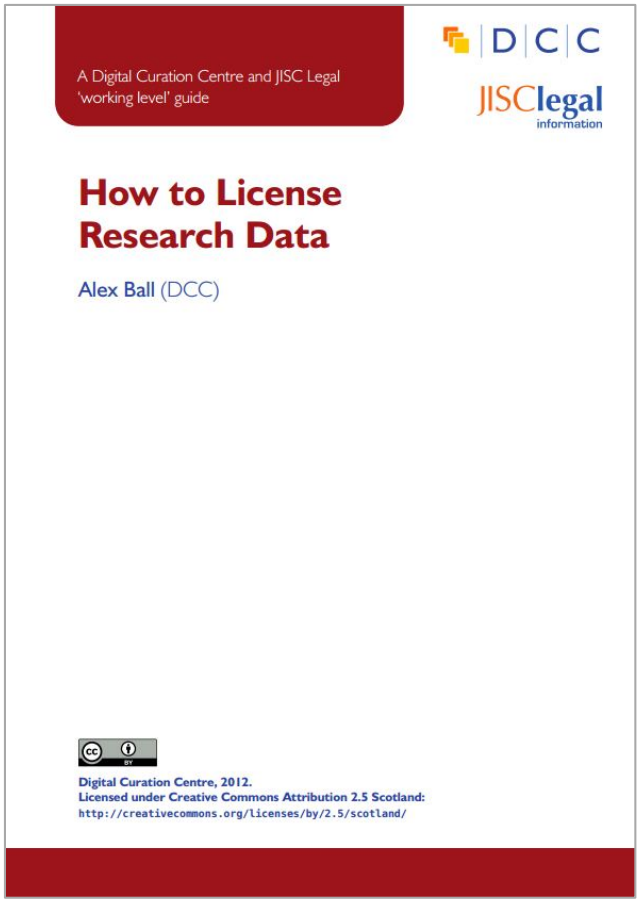
FOI & EIR

- FOI = Freedom of Information
- EIR = Environmental Information Regulations
- First is nation-state specific; second from European regulation
- Both have similar effects, but differ in detail

What the law allows – licensing & trust agreements

- Licences allow you to constrain how others use your data
- They range from very open to very restrictive
- You MUST own the data in order to be able to licence it
- Licences unlikely to help with sensitive data – seek help, use trusted repositories

License your data for reuse



Outlines pros and cons of each approach and gives practical advice on how to implement your licence

CREATIVE COMMONS LIMITATIONS

NC Non-Commercial
What counts as commercial?

SA Share Alike
Reduces interoperability

ND No Derivatives
Severely restricts use

www.dcc.ac.uk/resources/how-guides/license-research-data

Data and copyright

- Ability to copyright data varies throughout the world
- Europe also offers ‘database right’ – applies even if data cannot be copyrighted.
- International licences help avoid this legal minefield
- Standard licences strongly recommended – we are not all legal experts

Nature, Wednesday 3rd August 2016

Legal confusion threatens to slow data science

Researcher who spent months chasing permission to republish online data sets urges others to read up on the law.

Simon Oxenham

03 August 2016

Rights & Permissions

Steve Babuljak

Daniel Himmelstein, pictured at his previous research post at the University of California, San Francisco.

Knowledge from millions of biological studies encoded into one network — that is Daniel Himmelstein's alluring description of Hetionet, a free online resource that melds data from 28 public sources on links between drugs, genes and diseases. But for a product built on public information, obtaining legal permissions has been surprisingly tough.

Types of data licence

- Creative Commons V4.0 CC-BY or CC0 strongly recommended
- Also in existence:
  - Open Data Commons
  - Open Government Licence (UK)
- If your data includes/derived from other data – what licence applied to it?
- Licence ‘stacking’ can be a problem
- See also CODATA/RDA study on legal interoperability

Any questions?

Images

Alphabetti spaghetti from Ursus Wehrli: https://www.kunstaufraeumen.ch/en

Closed data from: http://www.gsma.com