RESEARCH DATA: THE ETHICAL DIMENSIONS

Paul Sturges
Professor Extraordinary
University of Pretoria
South Africa

Information and Library Science (LIS) as we have known it

- The preoccupations of Information Science have been largely technical, eg:
  - Metadata
  - Retrieval Systems
  - Automation of processes.
- The human dimension has been closer to the forefront of Librarianship as, eg:
  - Selection and presentation of resources
  - Service to users
  - Information Literacy programmes.

LIS as an ethical domain

- It can be argued that a fresh ethical focus is now found across the range of LIS
- This is driven by concerns including:
  - Internet security
  - Privacy and confidentiality
  - Official secrecy and Censorship
  - Equitable services
  - Intellectual property

Data as a new focus for ethical concerns

- Massive computing power and storage capacity, plus ubiquitous networking, means that:
- The twenty first century is emerging as an age of data
  - Personal data,
  - Research data,
  - Big data (described as ‘the sexiest job of the 21st century).
- All of this is heavy with ethical significance and ethical issues are now arguably more important than the technology. But -

What is Ethics?

- Ethics is formally defined as –
  - The science of duty, or
  - The branch of knowledge that deals with moral obligations
- In other words, ethics is the thinking and argumentation that we need to do about morality.
- Morality is the ideas and principles that we put into practice in our lives (including our professional or business lives).
- It is more and more obvious that information professionals need to think ethically and act morally towards individuals, groups, all of humanity and, indeed the natural environment.
- This concerns data as much or more than other forms of information.

Research Data in Context

- Research Data is commonly defined as
  - The recorded and organised factual material commonly accepted in the scientific community as necessary to validate research findings
- But this depends on Raw Data:
  - Information in unorganized form that refers to, or represents, conditions, ideas, or objects
- It also relates to Big Data, which is
  - Generated by communications technology, surveillance systems and commerce,
  - Stored and mined for administrative, technical and business purposes.

Ethics of Big Data

- Although Big Data is not the main concern here it has urgent ethical dimensions:
  - The volumes of data are almost unimaginable, and
  - The technologies used to mine it increasingly sophisticated.
- This presents dilemmas concerning
  - Human autonomy and choice
  - Privacy
  - Political Freedom.
- Research both creates some Big Data and can use Big Data resources.

But what exactly are we talking about when we talk about data?

- The nature and significance of data is essentially debateable.
  - If a tree falls in a forest and there is no one there to hear it, does it make a sound?
- We can infer that there would be a sound, if we believe that ‘a sound’ has an existence independent of human observation.
- But we can only be sure of a phenomenon observed (for which we have data).

The observer effect

- The idea that the observation and the phenomenon are congruent is open to question.
  - Eg The instrumentation used to measure something might actually alter it (eg. wild animals with tracker collars)
- There is also the related, but not identical, uncertainty principle (associated with Heisenberg)
  - The observation of one aspect of a phenomenon makes the observation of other aspects less certain.
- All of this suggests that data is a product of the observer, rather than simply a recording of the essential character of a phenomenon.

The (un)reliability of data

- In science the variations in data about a phenomenon, because of the intervention of different observers, may be minuscule (but significant)
- In social sciences (and LIS) the variations can be enormous and there is a great deal of thoroughly bad data collected and published.
- When research of high public interest (eg medical) is subject to transparency, the practitioners are likely to change what they do, in response to the data.
- It is clear that with data we are dealing with a human product to which ethical values need to be applied.

How the research community has dealt with data

- Research data has traditionally been presented in highly structured and selective forms as part of the reporting and publication process.
- Concern with data has tended to take the form of its availability for transparency purposes and re-use, typically in meta analysis (analysis of analyses).
- Human subjects have been a chief focus of concern:
  - Eg UK Warnock Commission of 1984's finding that *in vitro* culture of human embryos should halt at gastrulation (day 14).
- University Ethics Committees today would be unlikely to permit key LIS research like Hauptmann's 1976 'Mad Bomber' experiment on librarians' neutrality.

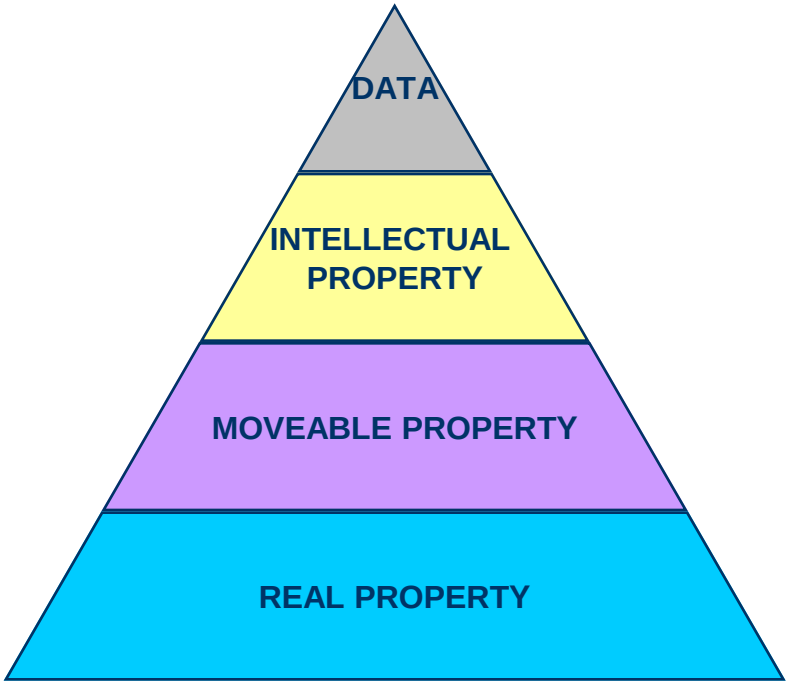
Peer Review as an ethical exercise

- Research (and the data) is also subject to peer review before publication.
- This is intended to ensure (amongst other things) that the data is convincing.
- But there is a fear that pressures for conformity, the demands of commercial funders, etc. skew the process.
  - J. Ioannidis, Why Most Published Research Findings Are False. PLoS Med 2005.
- So, even data published in ‘respectable’ journals requires ethical attention for accuracy, selection and relevance.

How LIS has dealt with data

- There is a case for saying that until recently LIS has not dealt with data at all.
- For example Suzanne Briet's powerful *Qu'est-ce que la documentation?* (1951) discusses the documents which a phenomenon generates, but not the data.
- The role of data is implicit in her analysis and it has remained that way in the literature until the 2010s.
- Arguably of most interest to LIS are issues revolving around the origins, ownership and curation of data.

A hierarchy of property



Ownership: Intellectual Property

- How far can we truly own Intellectual Property?
- Case Study – *Stairway to Heaven*, 1971.
  - A band called Spirit claim Led Zeppelin stole the opening chords from their tune *Taurus*, 1968.
  - It is an A-minor arpeggio played in a descending pattern.
  - Similar patterns in works by Pachelbel, Vivaldi and Chopin, other folk and popular songs.
  - We all borrow, often unconsciously (a process called cryptomnesia).
- Isn't most (all) IP more like a shared human resource then an individual possession?
- Case Study – the 'discovery' of Marginal Utility by Jevons, Menger and Walras in 1870.

Ownership: Data

- Ownership of data
  - Do we own data we collect?
  - Do we even own data about ourselves?
  - If data is not owned what is its status as information or documentation?
- The idea that data can be owned is an unlikely one in many ways,
- But despite that, various kinds of ownership are being claimed.

Ownership of personal data

- The law has recognised that people have a claim over data about themselves, but
- Government and business collect data about people in vast quantities.
- Some adjudication is offered by Data Protection Laws
  - DP laws do not so much as recognise ownership as grant a say in the accuracy, treatment and use of data
  - The concept is gradually being extended by the courts, for instance a Right to be Forgotten.
  - DP is not a privacy right, or a right of data ownership.
- But gradually concepts like Image Rights and restriction of visual images for safety reasons are being expanded.

Ownership of our own visual image

- In the past, laws have not recognised any ownership over images of oneself, now this is changing
  - Ownership of images for commercial exploitation, Douglas and Zeta Jones v Hello (2005)
- Essentially it was expected that everyone's image would be freely available, but now
  - The right of Muslim women to cover their faces is accepted in many countries
  - Pixillation of facial images of children (and adults) is practiced in media reports.

Compulsory sharing of personal data

- Parallel with the growth of rights over personal data, governments seek and obtain access (at least temporarily) to
  - CCTV images
  - Electronic transaction data
  - DNA databases
  - Other databases (eg.vehicle and driver licensing)
- The rationale is crime prevention and detection (particularly, but not only, terrorism).
- Critics claim it is the basis of a Surveillance State.

Problem areas: Commercialisation

- Data can be monetised.
- Case Study: Fertility Tracking Apps
  - Several companies are working on Apps that collect women's biodata, either
    - To create devices for personal use (for purposes of contraception or to increase chances of impregnation), or
    - To generate data which can be used to offer profitable fertility-related services to women (in partnership with pharmacies, fertility clinics, etc., or collecting data speculatively to seek uses).
- The case study involves access to medical data as well as purely commercial data, but avoids medical confidentiality.

Problem areas: Evidence-based Practice (EBP)

- Represents a break with the tradition of practitioner-based knowledge in medicine that has distant origins in folk medicine.
- It incorporates
  - Research based on ‘systematic collection of data through observation and experiment’
  - Clinical expertise
  - Patient values
- Raises questions over the origins and reliability of medical and pharmaceutical data (including the role of ‘Big Pharma’)
- Cochrane Review system seeks to address these questions based on ethical neutrality

Finally: Research Data

- Should research data be owned?
  - By researchers because of its ties to intellectual property in research publications;
  - By funders because they paid for its collection;
  - By journal publishers because they paid for its publication.
- Should research data be open?
  - Data open – for research transparency purpose
  - Data open for re-use eg. For meta analysis (the analysis of analyses).

Database Right as a model for research data ownership

- Although intended for commercially available databases, there is a clear model in the EU Rights in Databases Directive 1996.
  - Grants 15 years protection to computer databases on grounds of
    - Arrangement
    - Selection
    - Presentation
- In effect the Directive recognises the work that goes into data collection even though the data itself is not capable of being owned.

Benefits of data property rights

- Intellectual property is conventionally expected to recognise –
  - Economic rights
  - Moral rights.
- Moral rights in data are surely indisputable, but economic rights are much more complex and questionable.
- What is needed are intellectual property rights that guarantee *Responsibility* for data in the interests of *Curation*.
- The Open Data idea is a response to this.

Open Data

- The 2012 European Commission recommendation on scientific information calls for open access to data.
- Elements of this sharing include:
  - Data plans, providing for
  - Access and sharing,
  - Long term curation, and
  - Storage.
- This is fine in principle, but who plans, who curates and who pays?

Final Remarks

- Although a coherent set of ethical responses to data are very much a work in progress,
- Strong and helpful statements are emerging, eg
  - The seven core principles of the UK Engineering and Physical Sciences Research Council research data policy framework, 2011, suggest that:
    - Publicly funded research data should be widely and freely available, whilst
    - The research process should not be damaged by inappropriate release.
- Such statements are a strong basis for ethical thinking about research data.

Final Remarks (Continued)

- Furthermore, the LIS literature is beginning to include useful contributions to a debate on Research Data Management (RDM).
  - eg. Cox, A. et al ‘Moving a brick building: UK libraries coping with research data management. *Journal of Librarianship and Information Science*, 2016, 48(1), 3-17.
- A sound basis for ethical thinking about Research Data seems to be in the process of emerging, even in LIS.