Data Curation: a Gentle Introduction

LEADING

Alex H. Poole

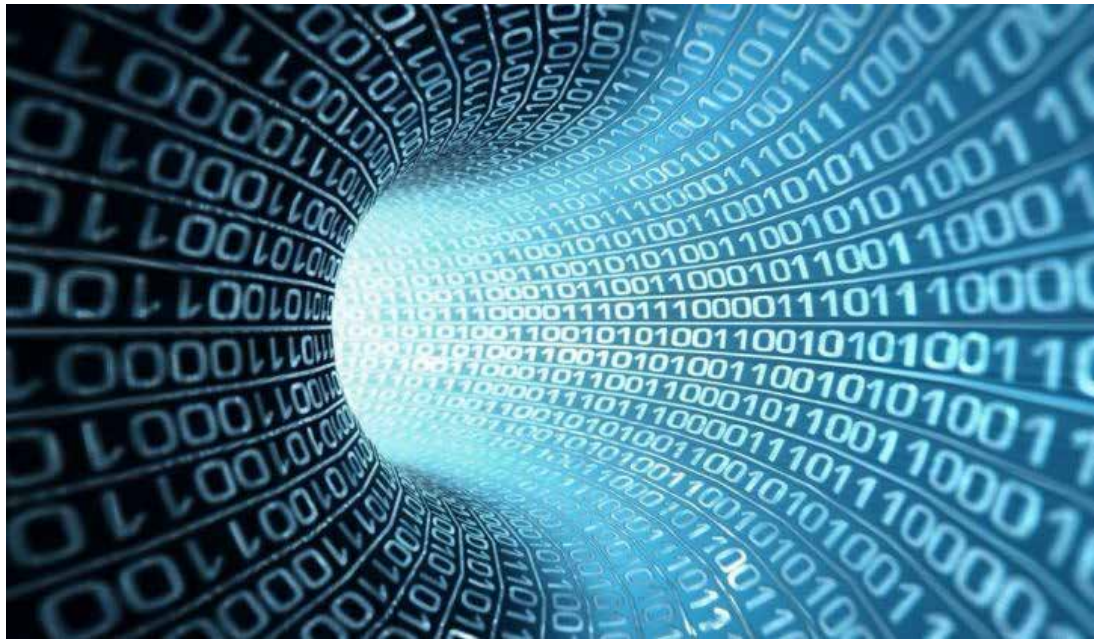
June 14, 2021

Agenda

- Data
- Big Data
- Curation
  - Lifecycle models
  - Metadata
  - Planning and policy
  - Governance
  - Sharing and reuse

Data

- “A reinterpretable **representation** of information in a formalized manner suitable for communication, interpretation, or processing.”
  - Examples: bit sequence; a table of numbers; characters on a page; recording of sounds; rock specimen (DCC)
- 1646 (*OED*)
  - Argument/fact
  - Theological
- Many things to many people!
  - Assets or liabilities!
  - “neither truth nor reality” (Borgman, 2015, p. 17)
  - “means to an end” (Borgman, 2015, p. xviii)
  - Value lays in use!



Data value

- Key questions:
  - 1) What will be useful in future?
  - 2) To whom?
  - 3) How long should data be kept?
  - 4) Should data stay usable?
    - To what extent?
    - for how long?
- Possibilities:
  - 1) Immediate
  - 2) Over time
  - 3) Transient
  - 4) Recreate

Generating data

- Observation
  - Monitoring events
    - Species counts
    - Weather
- Experiment
  - Controlled environments
    - Chemical reactions
- Simulation
  - Computer models of scientific systems
    - Global warming
- Compiled
  - Aggregate from other sources
    - Meta-analysis
    - Databases

Born-digital content

- Mobile devices
- 3D modeling and animation
  - Historic sites; video games
- Audio
- Relational databases
  - FB; *NYT*; JSTOR
- Digital documents
  - Word; PDF
- E-books
- Electronic journals
- E-mail
- Digital images
- Linked data
- Metadata
- Feed lists/playlists
  - RSS, Spotify
- Social media hubs
  - FB, Twitter, SnapChat, YouTube, Instagram
- Software
- Spreadsheets
  - Excel, CSV
- Video
- Computer and video games
- Virtual spaces
  - *Second Life*, *WoW*
- Visualization
- Websites

Paradigms

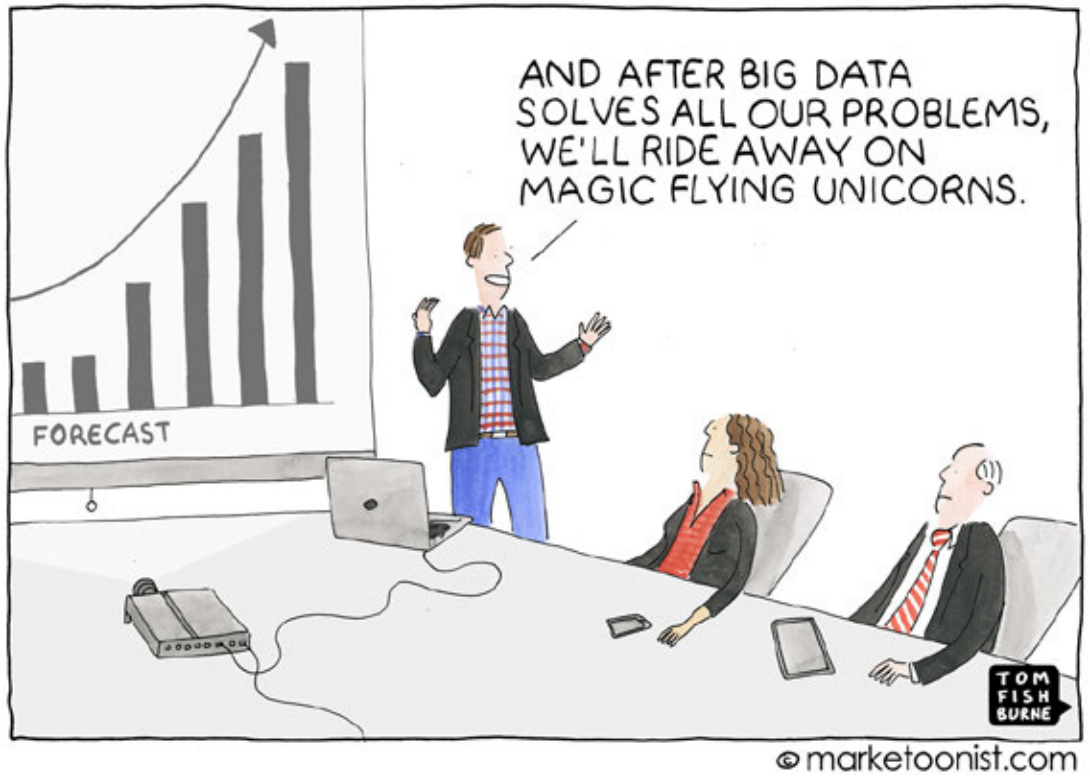
- Post-Newtonian paradigms: 1) experimental; 2) theoretical; 3) computational simulation
- Now is fourth!
  - Techniques & technologies necessary to perform data-intensive science
  - “Data is at the heart of this new paradigm, and it sits alongside empiricism, theory, and simulation, which together form the continuum we think of as the modern scientific method” (Wilbanks, 2009, p. 210)

Data at risk

- “Digital dark age” (Bollacker, 2010)
  - Physical media
  - Data’s comprehensibility
- “Shameful neglect”
  - “Scandalous shortfall” in research data sharing (“Data’s Shameful Neglect,” p. 145)
  - “Data management should be woven into every course in science, as one of the foundations of knowledge” (ibid, p. 145)
- Data as competitive asset
  - Quality as business issue
    - Customer relationship management (CRM)
    - Business intelligence (BI)

Big Data: context

- Historical context
- Term “Big Data” coined ~2008
  - New uses and insights
  - Orwellian side
- Big Data: datasets too big for conventional database tools to capture, store, manage, and analyze
  - Size is always a fast-moving target!
- Big Data as property; public resource; personal identity
- “In God we trust—all others bring data”



Bigness

- Bigness
  - What can be done with them
  - Insights they reveal
  - Scale of analysis required
- Data variety: scale, size, scope
  - What becomes data
  - Practices associated with them
  - How represented
  - How released and used
- Institutional factors
  - Norms/symbols
  - Intermediaries
  - Routines
  - Standards
  - Materials
- Project goals
  - Events/phenomena vs. systems
    - E.g., local weather vs. climate
- Collection
  - Big (machine)
    - E.g., water quality by sensors
  - Small (artisanal)
    - E.g. local water/soil samples
    - Domain knowledge key
- Analysis
  - E.g., cleaning, calibration

5 Vs

- Volume
  - Size of data terabytes, petabytes, etc.)
- Velocity
  - Speed of creating/processing/analyzing/storing data
- Variety
  - Various data types, sources, and modes
- Veracity
  - quality, reliability, and uncertainty
- Value
  - discovering actionable knowledge, return on investment, increased relevancy to customers or products, or innovations in business operations/processes (Song & Zhu, 2017)

Bigness

- “Datafication”
  - Quantify to enable tabulation and analysis
- Internet of Things (IoT)
- Exhaust/trace data
  - E.g., clicks
  - “lifestream” of documents
  - Data’s persistence
- “Option value”
  - Immense reuse potential
  - “data fusion”

Big Data’s promise: international development

- 1) Financial services
  - Spending and saving habits
  - Build credit histories
- 2) Education
  - Understand needs and gaps
- 3) Health
  - Trends and outbreaks
- 4) Agriculture
  - Food production trends
  - Storage, waste, and spoilage
  - Distressed regions (World Economic Forum, 2012)

McKinsey's take

- 1) Transparency
  - Reduce search/processing time
  - Concurrent engineering
- 2) Experimentation
  - Performance data: investigate variability
    - From product inventories to sick days!
- 3) Segmenting population
  - Tailor products and services
- 4) Supplant human with automated
  - Decisions, risk, and insight
- 5) New business models, products, services
  - E.g. real-time location data and new navigation services

A management cultural revolution?

- “Actions have far greater consequences, at a more accelerated pace, and direct repercussions for a company’s brand quality, customer relationships, and revenue” (Patterson & Kord, 2012)
- What gets measured gets managed!
- Structured databases no match for unstructured data
- Data-intensive work increasingly feasible given decreasing costs
- What do the data say?
  - Where did they come from?
  - What analyses were undertaken?
  - Level of confidence in results?
- “Data-driven decisions tends to be better decisions” (McAfee & Brynjolfsson, 2012, p. 68).
- Still need human vision and insight!
- “Each of us is now a walking data generator” (McAfee & Brynjolfsson, 2012, p. 63)

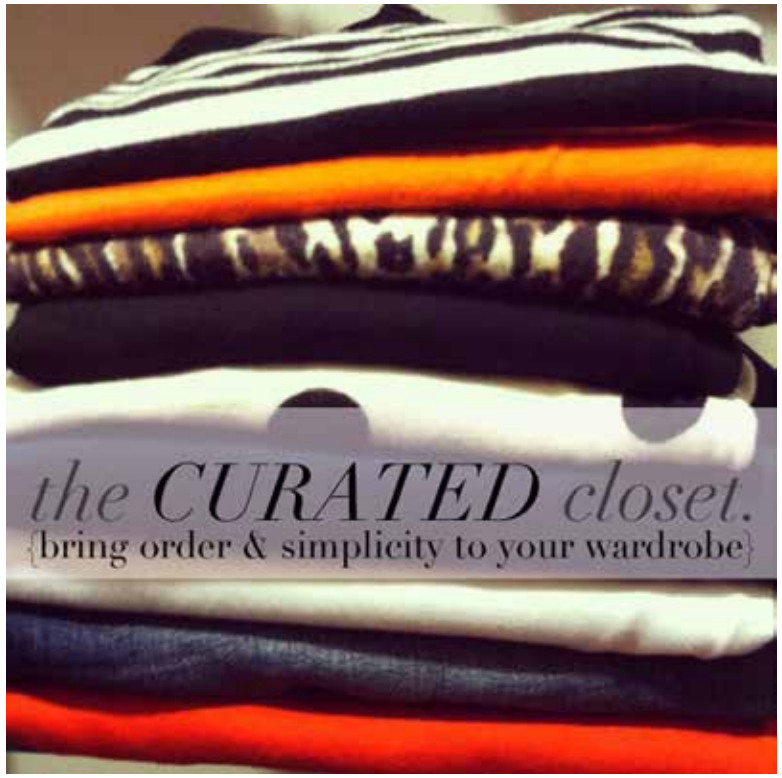
Federal Big Data Research and Development Strategic Plan (2016)

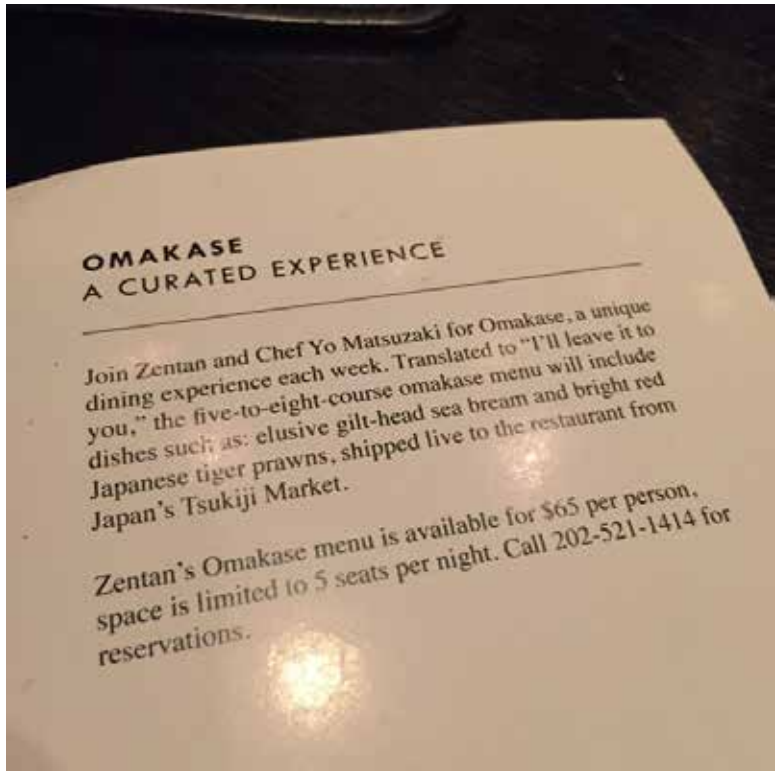
- “Big Data has the potential to radically improve the lives of all Americans. It is now possible to combine disparate, dynamic, and distributed datasets and enable everything from predicting the future behavior of complex systems to precise medical treatments, smart energy usage, and focused educational curricula” (p. 1)
- “a Big Data innovation ecosystem” (p. 1)
  - scientific discovery and innovation
  - educates the next generation of scientists and engineers
  - economic growth
- Takeaways
  - Utopianism
  - Democratization

Curation and preservation

- “Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle” (DCC)
- Digital preservation and data management
  - Digitized and born-digital
  - Add value
  - Trust/quality
  - Entire lifecycle
  - Fit-for-purpose
  - Sharing and reuse
- Never completed!

Curation in everyday life





Curators

- Duties
  - 1) interoperability
  - 2) metadata and annotation
  - 3) persistently linking
  - 4) persistent identifiers
  - 5) citation formats
  - 6) long term
  - 7) storage devices
  - 8) validate/authenticate
- Non-technical skills/human factors
  - 1) PM
  - 2) negotiation
  - 3) team-building
  - 4) problem-solving

Curation’s goals

- Ensure
  - Longevity
  - Integrity
  - Accessibility
  - Quality
  - Protection

- Stakeholders
  - Funders, disciplinary groups, creators, users, and curators

Curation’s payoffs

- 1) Direct to creator/owner/user
  - data quality/protection
  - access
  - sharing and reuse
  - scale/scope of research enabled
  - visibility of research/researcher
  - better records management
- 2) Public good
  - open access
- 3) Regulation, compliance, accountability
  - funding bodies
  - publishers
  - legal requirements
  - E.g., Banking, pharmaceuticals, medicine, aerospace

Big Data curation

- Key issues
  - 1) how much data to store
  - 2) storage cost
  - 3) security
  - 4) how long must it be maintained
  - 5) how curation process is automated
  - 6) how to remove redundancy
  - 7) how to curate software and applications
  - 8) determining necessary scope of curation
  - 9) how to capture metadata
  - 10) how to annotate metadata for extraction and management
  - 11) quality assurance (Song & Zhu, 2017)

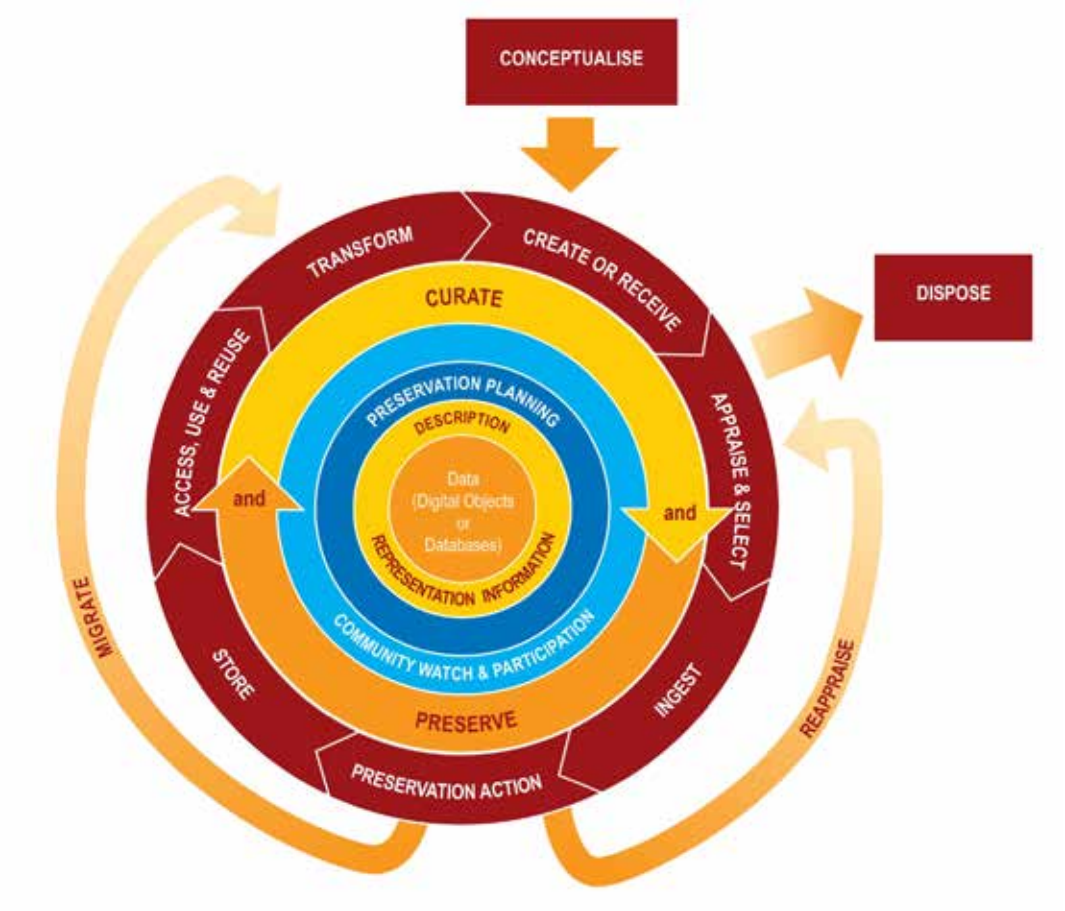
Lifecycle models

- Along with standards, models are pivotal
  - Heuristic value
  - Communication and accountability value
    - Roles and responsibilities
    - Policies and procedures
    - Technical infrastructure
- Payoff: target services and set forth priorities
  - Help researchers to care for and feed their data
  - Overall goal = keep or make data ready for reuse

Lifecycle models

- Usually progressive and circular
  - Helpful in forming a larger/holistic continuum
  - Visual aspect helps clarify/engage consumers (Carlson, 2014)
- Types: 1) individual (e.g., DCP); 2) organizational (e.g., ICPSR); 3) community (e.g., DCC)
- Limitations: 1) Idealized; 2) Downplay complexity; 3) Reflect biases of those who created them
- Key questions:
  - 1) Scope
    - Level of services to be offered?
    - Audience(s)?
  - 2) Integrating best practices/community standards
  - 3) Representing real world activities
    - Current practices? Rationale for them?
    - Any gaps between current and ideal practice?

DCC LM vs. DALM (Song & Zhu, 2017)



Metadata (description and representation information)

- “Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource” (NISO, 2004, p. 1)
- Metadata drives **all** of the steps in the data curation lifecycle! (Riley, 2014)
- “Metadata is like interest: it **accrues** over time” (Gilliland, 2008, p.18).
- “It’s metadata’s **world**, and you’re just living in it” (Pomerantz, 2015, p. 4).
- Reducing complexity!

Benefits of metadata

- Understand the context of objects and relations to others:
  - Persistently identify objects
  - Preserve reliable/persistent links
  - Describe objects
  - Identify technical characteristics of data
  - Determine who's responsible for management and preservation of objects
  - Detail what can be done to objects
  - Determine what's necessary to re-present objects
  - Record object's history
  - Document relationships, authenticity, integrity, completeness
  - Preservation
- Overall payoff = accessibility and use!
  - Legal and evidential relevance
- "Rosetta Stone"
  - "Love note to the future"

Planning

- Proactive preservation activity to reduce/minimize risk
- Risks:
  - 1) viruses
  - 2) backing up
  - 3) storage media failure
  - 4) equipment obsolescence
  - 5) lack of metadata
  - 6) lack resources
  - 7) physical disaster

- Risk management principles:
  - 1) regular backup
  - 2) multiple copies
  - 3) disaster recovery
  - 4) secure and stable media storage
  - 5) regularly copying data to more stable media
  - 6) ensure data security
  - 7) ongoing access
  - 8) limiting number of file formats
  - 9) community watch activities – new developments
  - 10) collaborative solutions

Policy

- Long term, regularly reviewed and updated:
  - 1) coherent strategy
  - 2) accountability
  - 3) protect organization
  - 4) acceptable practice
  - 5) repository commitment
- Good policy common elements:
  - 1) what's allowed/not
  - 2) how to be monitored and by whom
  - 3) links to relevant policies/materials
  - 4) date and frequency of review
- Kinds of policies necessary:
  - 1) archival storage
  - 2) management
  - 3) disaster recovery
  - 4) security

Governance

- “the system of decision rights and responsibilities covering who can take what actions with what data, when, under what circumstances and using what methods” (Ray, 2014, 45-46)
  - Includes: 1) business processes; 2) risk management
  - Ensures: 1) data trusted; 2) individuals accountable

- Key facets
  - 1) legal/policy
  - 2) attribution/citation
  - 3) archives and preservation
  - 4) discovery and provenance
  - 5) data schema/ontology discovery and sharing
  - 6) access to necessary infrastructure for data interpretation

Sharing and reuse of data

- Perhaps most important contribution!
- Impact
  - 1) climate change
  - 2) protect natural environment
  - 3) apply genomics-proteomics to human health
  - 4) shore up national security
  - 5) master nanotech
  - 6) guard against natural disasters

Possible benefits of sharing/reuse

- 1) compare, reproduce, validate research
  - Reinterpretation and new RQs
- 2) avoid duplication (cost and time)
- 3) make public assets available to the public
  - Transparency
- 4) leverage investment whether public or private
- 5) build on and advance research and innovation,
  - Create new data
  - Apply new methods
  - Link data
  - Mutual learning
- 6) researcher/stakeholder reputation
- 7) inform policy

Sharing and reuse

- Preconditions for sharing
  - Open (no IP restrictions)
  - Interoperability (standards)
  - Trust (authentic)
- Mechanisms of sharing
  - Discrete peer to peer
  - Project website
  - Institutional repository or center or archive

- *Willingness* to share does *not* equate to *actually* sharing
- “my data” is key barrier
  - Right to use
  - For how long
  - Free riders

Challenges

- Time commitment
- Discoverability
  - Identifiable and locatable
    - Metadata
    - Persistent identifiers
- Documentation
- Standards, e.g., Dublin Core, OAIS
- Ethical reuse
- Citing
  - Source data
  - Metadata
  - Annotations
  - Relevant/linked data
- Legal issues
  - Funders
  - Legislative
    - Confidentiality and privacy
  - IP and digital rights management
  - E.g.: researchers' rights; database; database structure; multiple holders
- Licenses (CC)
- Informed consent

References

- Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. The MIT Press.
- Brown, A. (2013). *Practical Digital Preservation: A How-To Guide for Organizations of Any Size*. Neal-Schuman.
- Oliver, G., & Harvey, R. (2016). *Digital Curation*. Neal-Schuman.
- Owens, T. (2018). *The Theory and Craft of Digital Preservation*. Johns Hopkins University Press.
- Poole, A. H. (2016). The conceptual landscape of digital curation. *Journal of Documentation*, *72*(5), 961–986. https://doi.org/10.1108/JD-10-2015-0123
- Ray, J. (2014). *Research Data Management: Practical Strategies for Information Professionals*. Purdue University Press.
- Ryan, H., & Sampson, W. (2018). *No-nonsense guide to born digital content*. Facet.