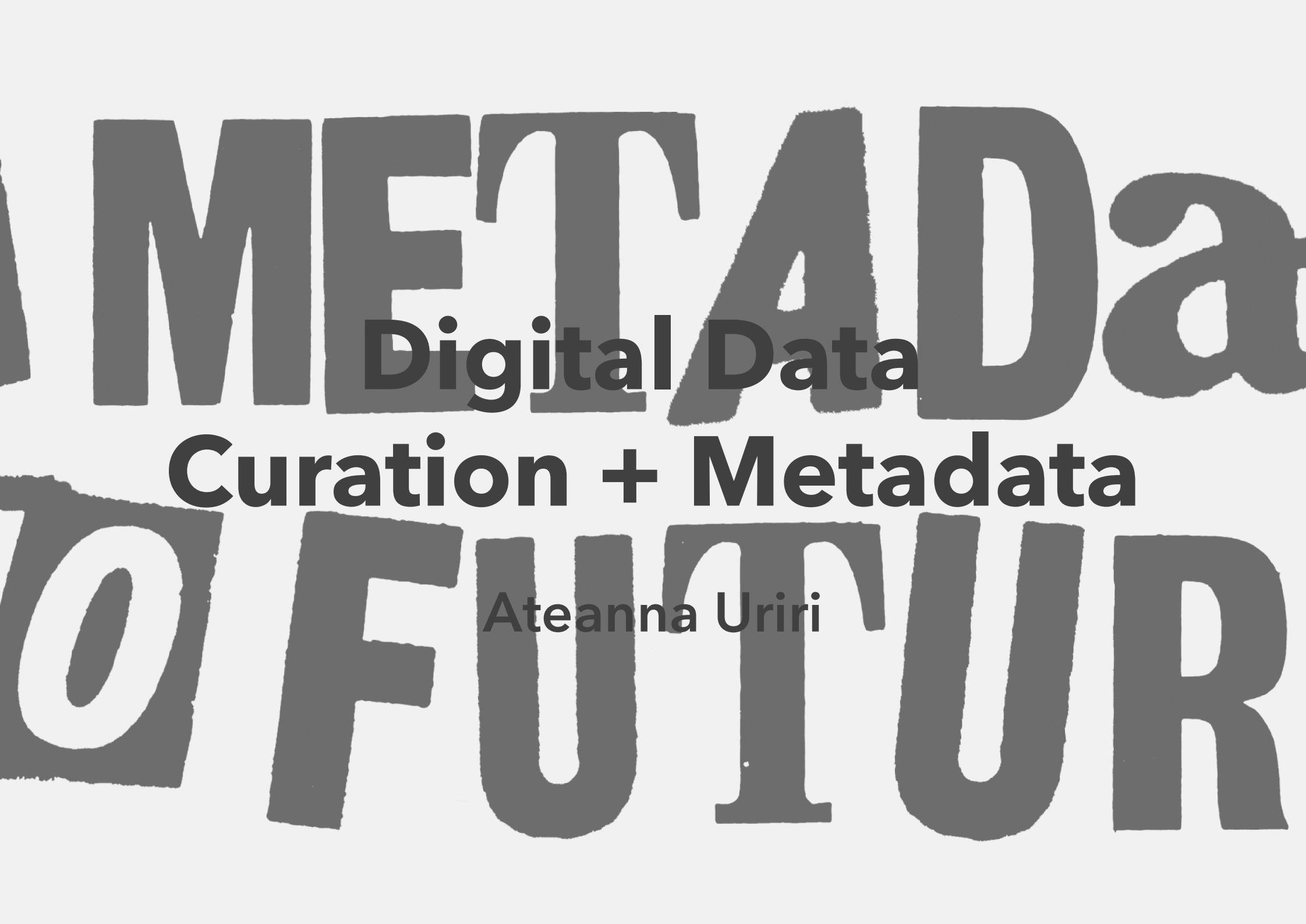


Digital Data Curation + Metadata

Ateanna Uriri

INTRODUCTION

This presentation will cover what *digital data curation* is, what *metadata* is, and how those topics converge. That convergence will be explored by discussing the *importance*, *use*, and *implications* of these topics. *Examples* are provided to assist with understanding both digital data curation and metadata.

FORESHADOWING.

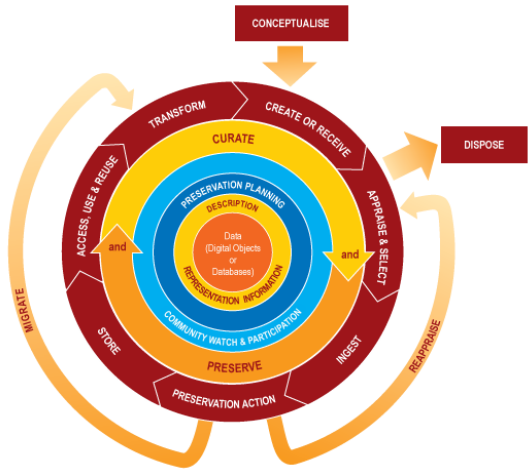
"data must be nurtured, not buried to be disinterred in the future" (Poole, 2015, p. 103).

"As our intellectual heritage moves more deeply into online research and teaching environments, new modes of inquiry emerge..." (Jahnke, Asher, & Keralis, 2012, p. 1).

A need for "trusted and enduring organizations to assume the stewardship for scientific data" and that "Stewardship includes ongoing creation and improvement of the metadata... by people cross-trained in scientific domains and knowledge management" (Atkins *et al.*, as cited in Ray, 2009).

DEFINITIONS

- **digital data curation** - or known as *digital curation* or *data curation*; the process of "maintaining and adding value to a trusted body of digital information for current and future use, specifically, [...] the **active management and appraisal of data over the life-cycle of scholarly and scientific materials**" (Digital Curation Centre, as cited in Harris-Pierce & Quan Liu, 2012).
- **metadata** - "data about data," the "who, what, when, where, and why" of data or more formally, "structured and standardised information that is crafted *specifically* [emphasis added] to describe a digital resource, in order to aid the intelligent and efficient discovery and retrieval of that source, accurate verification of its integrity (e.g. provenance tracking) as well as its apposite use and effective preservation over time" (Shaon & Woolf, 2008, p. 3).



WHAT IS THE LINK?

"The most *important* [emphasis added] reason to invest time and energy in developing metadata is that human memory is short" (Michener *et al.*, as cited in Bird, Willoughby, Coles, & Frey, 2013).

As the popularity or importance of digital data curation grows, the adoption and implementation of metadata is *necessary*, especially in the context of the future use of data. **How can future users or audiences understand/replicate/interpret the data is there is no context or instructions?**

DIGITAL CURATION

As you can see, the concept of digital curation has evolved over time to include or revise the what data are curated or the person(s) involved in curating.

What has remained is the importance of managing data not only for present or current use but for future use to be re-used. Also, the data curated actively moves through a **cycle** that includes ingestion, description, and preservation.

Screen grabs (Dallas, 2016, pp. 429-430).

Table 1 Definitions of digital curation

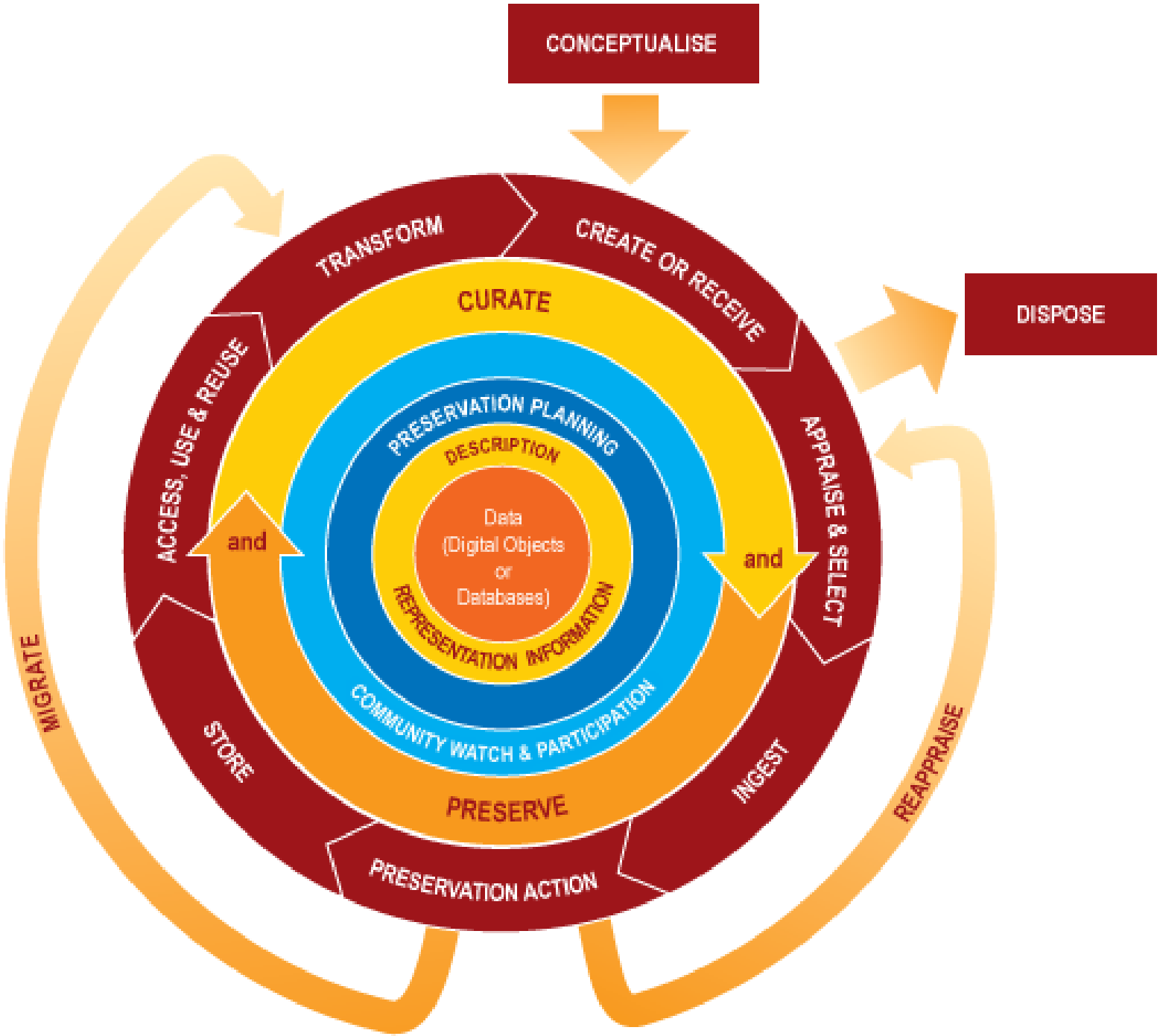
Definition	Source
"Digital curation [is t]he activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials."	Lord and Macdonald (2003)
"The term "digital curation" is increasingly being used for the actions needed to maintain and utilise digital data and research results over their entire life-cycle for current and future generations of users. Data curation in science may be defined as the maintenance of a body of trusted data to represent the current state of knowledge in some area of research. Implicit in this are the processes of digital archiving and digital preservation, but it also includes all the processes needed for good data creation and management, and the capacity to add value to generate new sources of information and knowledge. In most research fields, capturing "knowledge" is more than just the archiving and preservation of source data and associated metadata. It generally involves interaction between creators and providers of data, the archivers of data, and most importantly the consumers of data. Successful curation of data requires data users to be able to utilise the data using their current tools and methodologies."	JISC (2003)
"Digital curation, broadly interpreted, is about maintaining and adding value to a trusted body of digital information for current and future use; specifically, we mean the active management and appraisal of data over the life-cycle of scholarly and scientific material."	Pennock (2007)
"Digital curation involves the management of digital objects over their entire lifecycle, ranging from pre-creation activities wherein systems are designed, and file formats and other data creation standards are established, through ongoing capture of evolving contextual information for digital assets housed in archival repositories. Digital curation involves selection and appraisal by creators and archivists; evolving provision of intellectual access; redundant storage; data transformations; and, for some materials, a commitment to long-term preservation. Digital curation is stewardship that provides for the reproducibility and re-use of authentic digital data and other digital assets."	Lee and Tibbo (2007)
"[D]igital curation activities […] include […] maintaining and adding value to a trusted body of digital information for current and future use, through the active 'questioning', dynamic co-evolution and adequate representation of its epistemic/pragmatic content and context."	Dallas (2007a); Constantopoulos and Dallas (2008)
"Digital curation is the active involvement of information professionals in the management, including the preservation, of digital data for future use."	Yakel (2007)

Table 1 continued

Definition	Source
"Digital curation is the management and preservation of digital data over the long-term. All activities involved in managing data from planning its creation, best practice in digitisation and documentation, and ensuring its availability and suitability for discovery and re-use in the future are part of digital curation. Digital curation can also include managing vast data sets for daily use, for example ensuring that they can be searched and continue to be readable. Digital curation is therefore applicable to a large range of professional situations from the beginning of the information life-cycle to the end; digitisers, metadata creators, funders, policy-makers, and repository managers to name a few examples."	Abbott (2008)
"Digital curation is the curation, preservation, maintenance, and collection and archiving of digital assets. Digital curation is the process of establishing and developing long term repositories of digital assets for current and future reference by researchers, scientists, and historians, and scholars generally."	Wikipedia (2008)
"Digital curation is concerned with actively managing data for as long as it continues to be of scholarly, scientific, research, administrative, and/or personal interest, with the aims of supporting reproducibility, reuse of, and adding value to that data, managing it from its point of creation until it is determined not to be useful, and ensuring its long-term accessibility, preservation, authenticity, and integrity."	Harvey (2010a)
"Digital curation is a more inclusive concept than either digital archiving or digital preservation. It addresses the whole range of processes applied to data over their life-cycle. Digital curation begins before data are created by setting standards for planning data collection that results in "curation-ready" data–data that are in the best possible condition to ensure they can be maintained and used in the future. Digital curation emphasizes adding value to data sets, through things such as additional metadata or annotations, so they can be re-used."	Harvey (2010b)
"Digital curation is the selection, preservation, maintenance, collection and archiving of digital assets. Digital curation establishes, maintains and adds value to repositories of digital data for present and future use. This is often accomplished by archivists, librarians, scientists, historians, and scholars. Enterprises are starting to utilize digital curation to improve the quality of information and data within their operational and strategic processes. Successful digital curation will mitigate digital obsolescence, keeping the information accessible to users indefinitely. […] The term curation in the past commonly referred to museum and library professionals. It has since been applied to interaction with social media including compiling digital images, web links and movie files."	Wikipedia (2013)
"Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle. The active management of research data reduces threats to their long-term research value and mitigates the risk of digital obsolescence. Meanwhile, curated data in trusted digital repositories may be shared among the wider UK research community. As well as reducing duplication of effort in research data creation, curation enhances the long-term value of existing data by making it available for further high quality research."	DCC (2014)

"Data curation is a complete life-cycle **process** rather than a **singular activity** performed after an item has been created" (Brophy, 2007; Cervone, 2010, as cited in Harris-Pierce & Quan Liu, 2012).

"The model identifies: curation actions which are applicable across the whole digital lifecycle; those which need to be undertaken sequentially if curation is to be successful; and those which are undertaken occasionally, as circumstances dictate" (Higgins, 2008, p. 135).



How can curators make their curation life-cycle successful and on a larger scale, preserve data for posterity?

METADATA.

METADATA

Metadata "drives virtually all of the…steps in the curation life cycle…from preservation actions to access and reuse", but much of the scientific community lacks awareness of its importance (Riley 2014; Tenopir et al. 2011, as cited in Poole, 2015).

Metadata can tell users a lot about the data it is attached to and is necessary for researchers and other users who want to employ data that had been created and curated in the past. In the case of the image on the right (Greenfield, 2013), metadata can tell the viewer about a sent email that they haven't and may never see.

Crabtree and Cal (2014) describe objectives of data curation:

- Preserve research data,
- Enable possibility for secondary use,
- Understand the research context where data was created,
- Help next generation researchers discover the data,
- Help researchers understand their appropriate uses, and
- Understand collaboration points with research teams.

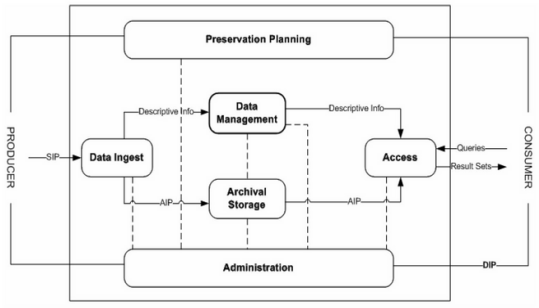
Again, in order to satisfy these objectives, **metadata** must not only be employed from beginning to end of the lifecycle (if not, some time before or after), it must be robust enough to provide data about the data to assist researchers and other users in re-using it.

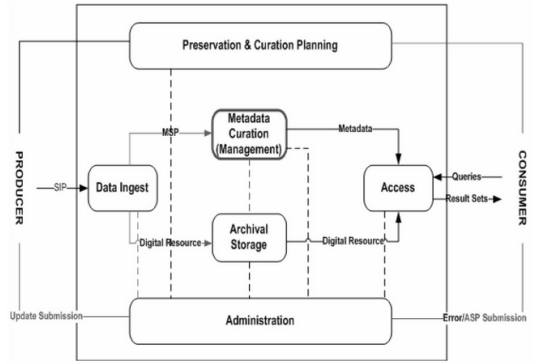
Scholars and researchers are now coming to understand the value of metadata in both digital curation and publication. Before, metadata had been an afterthought - a task that is completed to wrap up the data project. Now, there is a push to weave or embed the metadata process in the steps/aspects of curation that are necessary for the data.

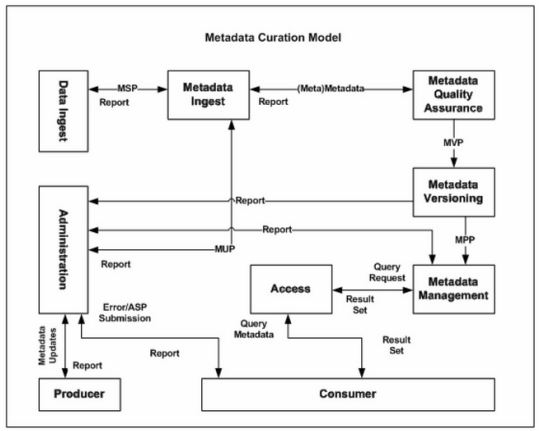
The push for robust metadata has led to the restructuring of the metadata process and how it fits or is supported by digital curation.

Shaon and Woolf (2008) presented the "Metadata Curation Model" in order to provide a better, long-term approach to metadata needed for the Open Archival Information System (OAIS) Reference Model, a popular framework for long-term preservation.

The **top diagram** shows the entities of the OAIS Reference Model. The **middle diagram** displays the proposed Metadata Curation Model overlaid/embedded on the current OAIS Reference Model, metadata curation is now in step with preservation. The **bottom diagram** shows a detailed view of the Metadata Curation Model.







A refresher (& breather):

ADVANTAGES:

- Pairing digital curation with metadata can essentially provide a manual or guide to how the data had been used and could potentially be used.
- Robust metadata can assist with replicating or reproducing data or project results.

LIMITATIONS/CHALLENGES:

- While there are no negatives when applying metadata to a digital curation project, applying metadata at the wrong time (as opposed to ALL the time) or not at all can limit the usefulness of the data and other related information compiled through digital curation.
- Researchers and scientists have felt that creating and maintaining metadata is tedious and not as important as the research itself.

"Data Curation Profiling of Biocollections" (Bishop & Hank, 2016)

"No matter the data curation role, effective data curation planning and implementation benefits from a proactive, approach to understanding the **'data story'** in its active use environment; that is, at the onset and during research activities producing data, rather than as post-script" (p. 1).

"Biocollections not only include specimens and/or images of those specimens, but locality data in various information types, complex data formats, ancillary files, dynamism, overall voluminous amounts, and most biota requires **domain-specific metadata** for appraisal and subsequent use" (p. 2).

"Issues related to provenance in digital biocollections relate to its sheer volume and immense scope and depth; failure to capture provenance at the time of the data's creation (legacy data) or during its manipulation in scientific models hamper scientific advancement" (p. 2).

"Curating Menus: Digesting Data for Critical Humanistic Inquiry" (Rawson, 2016)

"Defined both as the origin and the record of origin, **provenance** is central to using humanities data in ways that are rigorous—to see the ways that it is situated historically, shaped by the people and societies that formed it" (p. 61).

"For example, understanding Buttolph's catalog cards is critical to
understanding the overall project. Knowledge is structured in many ways, but **metadata is integral** to how people research in the digital humanities. Metadata makes it possible to make claims about the data or to perform comparative or other pattern-seeking analytical processes, be they computational or not" (p. 66).

CONCLUSION

"The most *important* [emphasis added] reason to invest time and energy in developing metadata is that human memory is short."

"A data preservation programme suited to the individual institution must be used to safeguard this huge investment of time and resources. Without good practices in place, the scientific record and documentary heritage created in digital form will remain at risk from digital obsolescence and also from the fragilities inherent to digital media" (Why preserve digital data?).


QUESTIONS?

REFERENCES.

Bird, C. L., Willoughby, C., Coles, S. J., & Frey, J. G. (2013). Data curation issues in the chemical sciences. *Information Standards Quarterly, 25*, 4-12. Retrieved from https://login.ezproxy.uta.edu/login?url=https://search.proquest.com/docview/1545796186?accountid=7117

Bishop, B. W., & Hank, C. (2016). Data curation profiling of biocollections. *Proceedings of the Association for Information Science & Technology*, *53*(1), 1-9. Retrieved from https://libproxy.library.unt.edu:9443/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=lls&AN=120386518&scope=site

REFERENCES.

Dallas, C. (2016). Digital curation beyond the “wild frontier”: a pragmatic approach. *Archival Science*, *16*(4), 421-457.

Gray, J., Szalay, A. S., Thakar, A. R., & Stoughton, C. (2002, December). Online scientific data curation, publication, and archiving. In *Virtual Observatories* (Vol. 4846, pp. 103-108). International Society for Optics and Photonics.

Greenfield, R. (2013). What your email metadata told the NSA about you. Retrieved from https://www.theatlantic.com/technology/archive/2013/06/email-metadata-nsa/313842/

REFERENCES.

Harris-Pierce, R. L., & Quan Liu, Y. (2012). Is data curation education at library and information science schools in North America adequate?. *New Library World*, *113*(11/12), 598-613.

Hawkins, D. T. (2017). Don's conference notes: Charleston seminar: Introduction to data curation. *Against the Grain, 26*(6), 36.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation*, *3*(1).

REFERENCES.

Hudson-Vitale, C., Imker, H., Johnston, L. R., Carlson, J., Kozlowski, W., Olendorf, R., & Stewart, C. (2017). Data curation, SPEC kit 354. Retrieved from http://libproxy.library.unt.edu:3932/Data-Curation-SPEC-Kit-354/

Jahnke, L., Asher, A., & Keralis, S. (2012). The problem of data. *Other Faculty Research and Publications,* Retrieved from https://digitalcommons.bucknell.edu/fac_pubs/52

Poole, A. (2015). How has your science data grown? digital curation and the human factor: A critical literature review. *Archival Science, 15*(2), 101-139. 10.1007/s10502-014-9236-y Retrieved from https://search.proquest.com/docview/1681267215

REFERENCES.

Rawson, K. (2016). Curating menus: Digesting data for critical humanistic inquiry. *Laying the foundation* (pp. 59-72) Purdue University Press.10.2307/j.ctt163t7kq.7 Retrieved from http://www.jstor.org/stable/j.ctt163t7kq.7

Ray, J. (2012). The rise of digital curation and cyberinfrastructure. *Library Hi Tech, 30*(4), 604-622. 10.1108/07378831211285086 Retrieved from http://www.emeraldinsight.com/doi/abs/10.1108/07378831211285086

Shaon, A., & Woolf, A. (2008). An OAIS based approach to effective long-term Digital Metadata Curation. *Computer and Information Science*, *1*(2), 2.

Why preserve digital data? (n.d.). Retrieved from http://www.dcc.ac.uk/digital-curation/why-preserve-digital-data

THANK YOU!