Data Lifecycle Management: What Has Got to Give

Will Hires
Louisiana State University, will.hires@howard.edu

Will Hires, "Data Lifecycle Management: What Has Got to Give" (2011). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284314934

Data Lifecycle Management: What Has Got to Give

Will Hires, Assistant Librarian, Louisiana State University

Abstract:
The presentation will review the definition and component stages of data lifecycle management and discuss relevant and associated issues. In addition, there will be a focus on elements of workflow management that has to accommodate data use, manipulation, sharing, and preservation. The objective is to provide an outline for understanding data management (DM) and explore opportunities for effective DM implementation. The audience will be asked to help, based on individual experiences, identify issues and obstacles to DM implementation and associated remedies to apply towards eliminating problems and bottlenecks. An attempt will be made to capture lessons from actual experiences with data that can be effectively applied against circumstances that may be known or routine. Additional information will be provided from existing case studies that will be able to illustrate methods of addressing DM needs under similar circumstances as well as identifying considerations applicable to unique situations. The learning expectations and objectives include the development of defining terminology, expressing ways to structure the issues relevant to DM, developing methods for effective implementation DM principles, and efficiently integrating DM in the normal workflow of a professional librarian.

Hello and good afternoon. My name is Will Hires and I'm the Engineering and Scholarly Communication Librarian at Louisiana State University in Baton Rouge, Louisiana. This presentation will review the definition of data lifecycle management (DLM) and discuss the component stages of DLM along with some relevant and associated issues. Hopefully, you will be able to gain an understanding and appreciation for data management and why it is important.

To start with, data is defined as "Unorganized pieces or segments of information extending from observed phenomena or structured activity". So, data are the basic forms of collected information gathered from an activity involving detailed scrutiny of something or some occurrence. What things make up data? Data can be "facts, measurements, or descriptions of observations". Data can result from "sampling" of something that may be too large or complex to be distinctively considered, or a "comparison" of many things taken together as a single event or entity. Data can consist of "numbers" such as when things are counted; "characters" such as labels used to conveniently represent things; or "markings" such as special indicators used to tag an item. "Photographs sequenced over time" can represent data, especially when they are "focused on a single object or thing". Data can be represented many ways, but it can be clearly seen that the term, "data", describes pieces or segments of information. Later on, an analysis of these "pieces" will endeavor to make sense of them and to ascribe meaning to them.

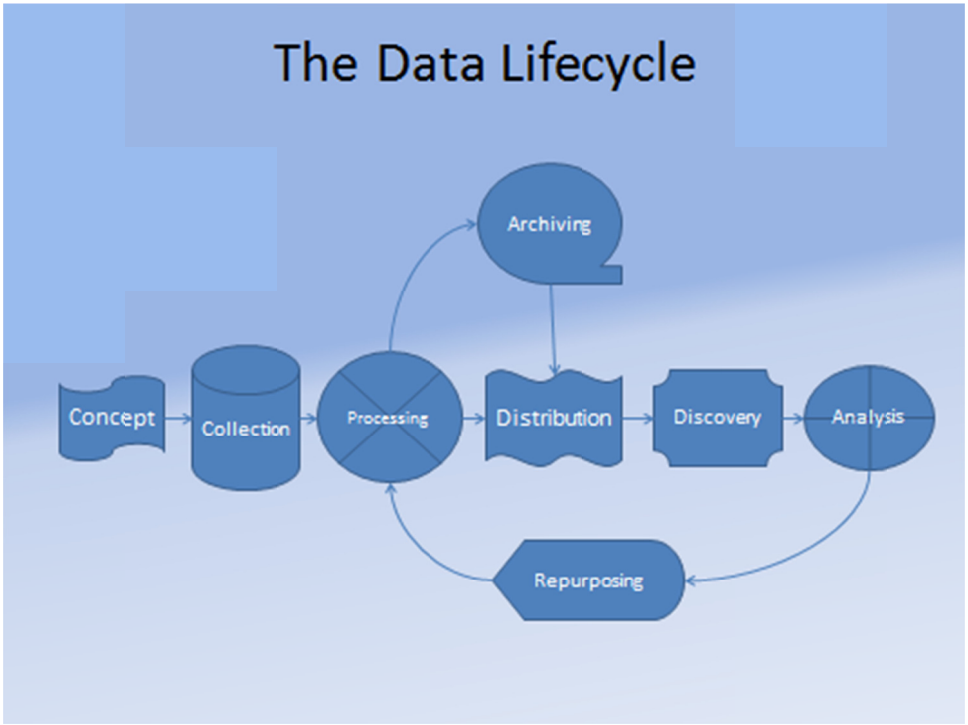
Data can be fundamentally or two types [from slide 4], "discrete" or "continuous". Things such as a "counted number", a "single character", or a "single part/piece of information" can be used to describe data discretely. "Measurements, such as determined from weight, dimensions, or states" are an example of data that cannot usually stand alone and exists within the context of some additional element or thing. This is also true of data derived from "position" or "condition" of an observed phenomena or thing.

Over a lifecycle, data can exist in various forms. "Raw data" are "products of observation, experimentation, or computation". This is the most basic form of data, especially before anything has been done to it. "Intermediate data" are "extractions from analyses and processing, such as from simulations or explorations", for example. "Final data" are the "results of research". But data can also take on another form: "obsolete data" happens when data has been "compromised" having "lost validity or relevance"; it can also be data that has a "superseded basis". Now what is meant by a "lifecycle" is just that: the familiar bell-shaped curve that signals a beginning point and an end point with a span of existence that rises and falls within those endpoints. Within any lifecycle, there are distinct phases of growth, maturity, and decline. This is also true

DOI: http://dx.doi.org/10.5703/1288284314934

of data: there is a beginning point or phase and, then a period of utilization which is typically followed by a period of disuse (or neglect). Data can then actually be destroyed because it is no longer useful, shelved or stored for possible later applicability or forgotten about (because it has been superseded). Data can also exist in a preliminary status when there is a planned intention to add to it or further manipulate it as part of the collection effort.

The definition of data lifecycle, then, is "the span of existence of data from the time of its creation to the time of its transformation (destruction, evolution, or repurpose)". This is important because data is essentially not as valuable without an associated plan for its management (and development).



"What is data lifecycle management"? Management will inevitably involve concern about the "format specification" of the data. What format is it in? What other hardware and/or software is associated with being able to accommodate this format? What future issues are likely to develop because of this format? Can this format be easily converted, if necessary? These kinds of questions associate with concerns related to data formats. These concerns may evolve and, very likely, may become complex or additionally complicated by technological developments or other conditions. One of these conditions might be "storage". Will climate control be necessary? Where will the data be stored? How long is storage anticipated? Will access (frequent and/or by whom) be a matter of concern? It may be especially apropos to define the elements of "control/access/dissemination" with respect to data management. Matters of authorization may figure crucially into the equation of management. And finally, "preservation" should be a matter of concern because future intentions with respect to data might be an unknown. Preservation increases in importance according to the type, format, and value of the data. Of course, these are simply management challenges and appropriate attention and planning will mitigate problems. There will be challenges associated "intended purpose", "actual use and/or development", "practical size limits", "required bandwidth", "formats and related derivatives", and "envisioned future". A lot depends on what "resources" are available; whether "network access" is available and/or needed, and any known or "expected applications" that may be associated with the data. Challenges are matters for manage-

ment to overcome and proper and thorough consideration will enhance preparedness for them.

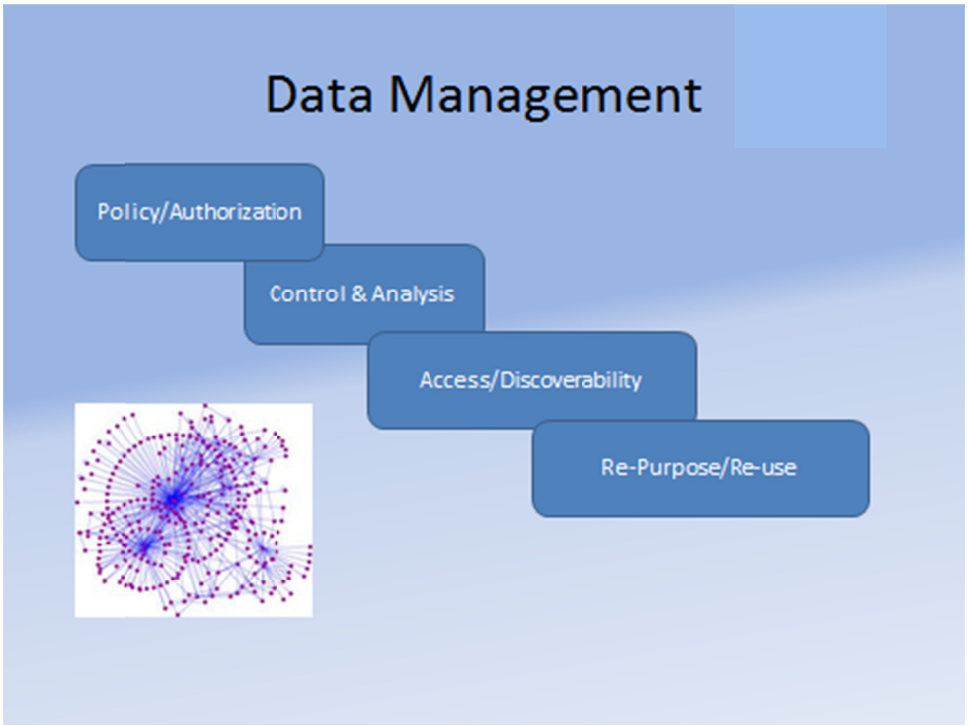
The "data management plan" requirement recently established by the National Science Foundation (NSF) goes a long way towards helping data managers prepare for handling data. Describing data as "products of research", the NSF plan asks that "data formats" be described and recorded; that there be specific consideration made for "data access and data sharing"; that researchers expect and describe any pertinent conditions for "data re-use, redistribution, and dissemination"; that researchers establish a policy that will pertain to the "production of derivatives"; and that there be a plan for "archiving and/or preservation" for the data. The NSF data management plan provides for the a priori development of guidelines which will apply to data developed through [NSF] funded research. These guidelines essentially divide between "data collection" and "storage/preservation". The data collection concerns will enforce consistency in describing "what kinds of data" are created; what "standards of production and/or use" pertain to the data; and what "limitations and/or restrictions" pertain, whether they be "legal or ethical" in nature. The storage/preservation concerns will highlight the "control" of data, how data "sharing and/or dissemination" will be accomplished, and what specific "long-term considerations" pertain (especially with respect to "format and ownership").

Attendees at the presentation were asked to identify potentially discuss some "issues and/or obstacles" that may be associated with data management. No responses were forthcoming, so the following questions were put forward: "whose responsibility" is it to manage the data? Would the researcher be the best manager of the data he created or helped to create? Alternately, should a special data manager be appointed so that the researcher can be free and to pursue technical issues? "Where is the consistency", especially in the descriptions and actions taken, with respect to the data? If each researcher has the added responsibility of management of the data allow for the necessary consistency (for validity and reliability)? "How much redundancy" is needed? Redundancy improves storage options and response to hazards or disasters; however, there is a concomitant impact on resources for this to be effective and efficient. "How long is long enough"? Is it enough to keep the data until its format is obsolete, it is no longer actively used, or new and better data is developed? Who would be authorized to decide when long enough is sufficient? What about "privacy and/or ownership"? Is the data sufficiently separated from the entity from which it is derived? What happens to the data when the researcher is transferred or leaves the organization or institution under whose authority the research was conducted? Again, these are essentially challenges that management should be expected to overcome.

Data is increasingly voluminous and, as was pointed out during the discussion yesterday by Clifford Lynch, is being produced quicker than ever; so, "what has to give"? These three things, in my determination, will have to be definitively addressed: "data management", "data sharing/dissemination", and "data storage/preservation". "Data management" has to be made "a priority", especially for reasons that will allow people and organizations "to reduce liability", "to substantiate productivity", "for record authenticity", and "to preserve privacy". All of these reasons are relevant and necessary and represent, in my opinion, an absolute imperative with regard to realizing control and responsibility with respect to data. [from slide 16] "Data sharing and dissemination" must be expected in order to facilitate "expanding accessibility", "encouraging collaboration", "stimulating and maximizing the potential of ideas", "supporting the rights of citizen/taxpayers", and eliminating barriers to research". Of course, sharing and disseminating data as much as possible will also "encourage the extending of research" and promote and maximize wide participation in research. Additionally, the appreciation for research and a better understanding of why research is needed are further benefits. "Data storage and preservation" is concerned with ensuring the "availability and usability" of data for some future purpose. Data that is properly stored and preserved will retail properties that sustain its functionality and applicability. Additionally, management will ensure that the data is, appropriately, "functional", "findable", "changeable" and/or "transformable".

No good management plan should be considered complete without a policy. A data lifecycle policy should "identify and prioritize objectives", "achieve

minimalization" (since everything cannot be conveniently kept—nor should it be), establish a plan to "train everyone" involved with data, both technically and administratively, ensure appropriate security" of the data, "establish audits" (to make sure the storage solution is tested, affirmed, and purged when needed), and "document" the (actions, significant events, and transfers) that occur in connection with handling the data to achieve preservation. These specific elements of "data management" should be minimally included in any plan.



The data management cycle identifies several points associated with data starting with the researcher who may be involved with the creation of field data, to the repository that temporarily contain the data, to the consumer who benefits from the analyzed results. Key among these elements is the requirement for systematic management of the data and, ultimately, proper preservation to sustain intended and future purpose that may be associated with that data. "When all is said and done, successful data management depends upon commitment and sufficiently allocated resources". Thank you for your attention and attendance, any "questions"?