Data Management Glossary

ACTI Data Management Working Group (ACTI-DM)
www.educause.edu/acti

February 2013

This data management glossary was created by the Advanced Core Technologies Initiative Data Management Working Group and pulls from existing definitions for commonly used terms, as well as includes definitions that have been created by the working group. It is intended to be regularly added to and adjusted. Feedback and suggestions are welcome, and can be sent to acti@educause.edu.

Administrative Data: The university administrative data is shared data, managed within a conceptual framework. It is likely that the university administrative data will be distributed across processing units both within and outside the university. Specific types of data, such as research data and electronic mail “boxes,” may be covered by specially tailored policies. …Some examples of administrative data include student course grades, patient records, employee salary information, and vendor payments.
http://its.virginia.edu/policy/admindataaccess.html

Anonymization: The act of permanently and completely removing personal identifiers from data, such as converting personally identifiable information into aggregated data. Anonymized data is data that can no longer be associated with an individual in any manner. Once this data is stripped of personally identifying elements, those elements can never be re-associated with the data or the underlying individual.
https://wiki.internet2.edu/confluence/display/itsg2/Guidelines+for+Data+De-Identification+or+Anonymization

Archives: A place or collection containing static records, documents, or other materials for long-term preservation.
Defined by the ACTI-DM Working Group
**Authentication:** The process of confirming the identity of a principal. Since computer identification cannot be absolute (e.g., passwords can be stolen), authentication relies on a related concept of level of trust, in which an institution relies on good identity management practice (so that the institution believes they have correctly identified an individual) and secures mechanisms for sharing identity. This is sometimes referred to as AuthN (authentication), in contrast to AuthZ (authorization).

[https://spaces.internet2.edu/display/macepaccman/MACE-paccman-glossary](https://spaces.internet2.edu/display/macepaccman/MACE-paccman-glossary)

**Authorization:** The process of deciding if a subject (person, program, device, group, role, etc.) is allowed to have access to or take an action against a resource. Authorization relies on a trusted identity (authentication) and the ability to test the privileges held by the subject against the policies or rules governing that resource to determine if an action is permitted for a subject.

[https://spaces.internet2.edu/display/macepaccman/MACE-paccman-glossary](https://spaces.internet2.edu/display/macepaccman/MACE-paccman-glossary)

**Big Data:** A loosely defined term used to describe data sets so large and complex that they become difficult to process and store using data management tools.

*Defined by the ACTI-DM Working Group*

**Cloud Computing:** Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet).


**Cyberinfrastructure:** United States federal research funders use the term cyberinfrastructure to describe research environments that support advanced data acquisition, data storage, data management, data integration, data mining, data visualization and other computing and information processing services distributed over the Internet beyond the scope of a single institution. In scientific usage, cyberinfrastructure is a technological and sociological solution to the problem of efficiently connecting laboratories, data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge.


**Dark Archive:** Archived data that focuses on preservation and not accessibility.

*Defined by the ACTI-DM Working Group*

**Dark Data:** “Dark data” is not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost.


[https://www.ideals.illinois.edu/bitstream/handle/2142/9127/Heidorn_LongTail_PreprintwEdits.doc](https://www.ideals.illinois.edu/bitstream/handle/2142/9127/Heidorn_LongTail_PreprintwEdits.doc)
Data (See also “Research Data”): [V]alues of qualitative or quantitative variables, belonging to a set of items. … Data … can be viewed as the lowest level of abstraction from which information and then knowledge are derived.

http://en.wikipedia.org/wiki/Data

Data Analysis: [A] process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

http://en.wikipedia.org/wiki/Data_analysis

Data Citation: Data citation offers proper recognition to authors as well as permanent identification through the use of global, persistent identifiers in place of URLs, which can change frequently. Use of universal numerical fingerprints (UNFs) guarantees to the scholarly community that future researchers will be able to verify that data retrieved is identical to that used in a publication decades earlier, even if it has changed storage media, operating systems, hardware, and statistical program format.

http://thedata.org/citation/standard

Data Curation: The selection, preservation, maintenance, collection and archiving of digital assets.


Data Governance: See “Governance”

Data Lifecycle¹: [A data lifecycle] takes into consideration the value of the information over time, how quickly and at what cost it must be made available for user queries and how long it must be retained before being deleted.

http://encyclopedia2.thefreedictionary.com/data+life+cycle+management

Data Management: Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.


¹ For more information see Appendix A; The Data Lifecycle in Developing an Institutional Research Data Management Plan Service (EDUCAUSE, 2013):

http://www.educause.edu/library/resources/developing-institutional-research-data-management-plan-service
Data Management Plan: See “Research Data Management Plan”

Data Management Services (See also “Research Data Services”): Services provided to researchers and faculty to assist them with the creation of data management plans, metadata, storage planning, and curation requirements. Defined by the ACTI-DM Working Group

Data Migration: The process of transferring data between storage types, formats, or computer systems. http://en.wikipedia.org/wiki/Data_migration

Data Mining: A particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Data mining results in the discovery of new patterns in large data sets. The overall goal of the data mining process is to extract knowledge from an existing data set and transform it into a human-understandable structure for further use. Defined by the ACTI-DM Working Group

Data Provenance: In its most general form, provenance (also sometimes called lineage) captures where data came from, how it was derived, manipulated, and combined, and how it has been updated over time. http://infolab.stanford.edu/panda/

Data Repository (See also “Repository” and “Trusted Digital Repository”):
- A system that provides online access to research data.
- By an entity that is committed to providing access to the data
- That serves as a primary, authoritative source for data
- With a significant collection of like data or data within the scope of a management policy.


Data Stewardship: The formalization of accountability for the management of data resources. http://www.tdan.com/view-articles/5037

Digital Preservation: The active management of digital content over time to ensure ongoing access. http://www.digitalpreservation.gov/about/

Embargo Period: [An] embargo is a period during which access is not allowed to certain types of users. http://en.wikipedia.org/wiki/Embargo_%28academic_publishing%29
**Federated Architecture:** [A] pattern which describes an approach to enterprise architecture that allows interoperability and information sharing between semi-autonomous de-centrally organized lines of business (LOBs), information technology systems and applications.  
http://en.wikipedia.org/wiki/Federated_Architecture

**Federated Storage:** Federated storage is the collection of autonomous storage resources governed by a common management system that provides rules about how data is stored, managed, and migrated throughout the storage network.  
http://wikibon.org/wiki/What_is_Federated_Storage%3F

**Governance:** Specifying the decision rights and accountability framework to encourage desirable behavior in using IT.  

**Infrastructure as a Service (IaaS):** In the most basic cloud-service model, providers of IaaS offer computers—physical or (more often) virtual machines—and other resources.  
http://en.wikipedia.org/wiki/Infrastructure_as_a_service#Infrastructure_as_a_service_.28IaaS.29

**Intellectual Property:** Intellectual property (IP) refers to creations of the mind for which exclusive rights are recognized in law.  

**Metadata:** In general, “data about data;” functionally, “structured data about data.” Metadata includes data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation.  
http://dublincore.org/documents/usageguide/glossary.shtml#M

**Open Access:** Open access (OA) is the unrestricted free access to information and data.  
*Defined by the ACTI-DM Working Group*

**Personally Identifiable Information (PII):** [I]nformation that can be used to uniquely identify, contact, or locate a single person or can be used with other sources to uniquely identify a single individual.  
http://en.wikipedia.org/wiki/Personally_identifiable_information

**Platform as a Service (PaaS):** [P]rovides a toolkit for conveniently developing, deploying, and administering application software that is structured to support large numbers of [users], process very large quantities of data, and potentially be accessed from any point in the Internet.  
http://www.nist.gov/customcf/get_pdf.cfm?pub_id=911075
**Data Management Glossary**

**Policy:** The set of basic principles and associated guidelines, formulated and enforced by the governing body of an organization, to direct and limit its actions in pursuit of long-term goals.

[http://www.businessdictionary.com/definition/policy.html](http://www.businessdictionary.com/definition/policy.html)

**Principal Investigator:** The [person] administratively in charge of an experiment or research project.


**Record:** Anything preserving information and constituting a piece of evidence about past events; esp. an account kept in writing or some other permanent form.


**Research Computing:** IT methods and tools useful for [researchers] in conducting their investigations. By definition, research involves extending the boundaries of subject area knowledge, and IT is often critical to the success of these endeavors. An inability to obtain or invent the right IT tool may mean that research simply cannot occur.


**Research Data** (See also “Data”): Data in any format or medium that relate to or support research, scholarship, or artistic activity.


**Research Data Management Plan:** A data management plan is a formal document that outlines how you will handle your data both during your research, and after the project is completed.


**Research Data Management Plan Tool:** A software tool that provides guidance and resources for the construction of data management plans.

Defined by the ACTI-DM Working Group

**Repository** (See also “Data Repository” and “Trusted Digital Repository”): Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse.

Defined by the ACTI-DM Working Group

**Research Data Services** (See also “Data Management Service”): Services and expertise related to the management of research data. These may include data management plan guidance, metadata creation, intellectual property and copyright consulting, etc.

Adapted from [https://confluence.cornell.edu/display/rdmsgweb/services](https://confluence.cornell.edu/display/rdmsgweb/services)
**Software as a Service (SaaS):** Software deployed as a hosted service and [made available] over the Internet


**Trusted Digital Repository** *(See also “Data Repository” and “Repository”)*: [A repository] whose mission is to provide [its designated community with] reliable, long-term access to managed digital resources.


---

**Acknowledgments**

The [ACTI Data Management Working Group](http://www.acti.org) created this glossary as a result of their efforts on data management. Special thanks go in particular to working group member Tom Dopirak (Consulting Architect, Carnegie Mellon University) for his efforts.