ASPE
ASSISTANT SECRETARY FOR PLANNING AND EVALUATION
OFFICE OF SCIENCE & DATA POLICY

Working Paper
September 2022

Glossary of Common Data-Related Terms

The Department of Health and Human Services (HHS) has several different policy groups such as the HHS Data Council, Data Governance Board, Evidence and Evaluation Policy Council, and the HHS AI Council that frequently use many terms related to data, but likely with inconsistent understanding of their definitions and how these terms should be used. This brief provides a list of key data terms and their definitions to facilitate communication and coordination across these Councils and HHS more broadly.

Whenever possible, the definitions presented below are taken from Federal statutes, Federal regulatory guidance, or other Federal sources, in that order of preference. In cases when a Federal source could not be identified, definitions are taken from intergovernmental or non-government not-for-profit organizations. Definitions are provided with slight modifications from source definitions when needed for clarity, brevity, or applicability. Some terms may have alternate definitions; those presented below do not reflect consensus across HHS but are relevant definitions from authoritative sources – such as OMB statistical Directives and Public Laws – for reference and consideration. In some cases, more specific definitions may apply to particular circumstances; widely applicable definitions for HHS are provided here. As work in these areas progresses, HHS may choose to refine or select preferred definitions for selected terms.

KEY TERMS AND DEFINITIONS

Administrative data: Programmatic, regulatory, law enforcement, adjudicatory, financial, or other data held by agencies and offices of the government or their contractors or grantees (including States and other units of government) and collected for other than statistical purposes. Administrative data are typically collected to carry out the basic administration of a program, such as processing benefit applications or tracking services received.[1]

Artificial intelligence: An artificial system that (1) performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to datasets, (2) solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action, (3) thinks or acts like a human, including cognitive architectures and neural networks, (4) is designed to approximate a cognitive task through a set of techniques, including machine learning, and (5) acts rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision making, and acting. Related terms: "Machine Learning," "Natural Language Processing" and "Robotic Process Automation."[2]

Availability: The timeliness, reliability and equitable access to and use of information.[3, 4]

Blended data: Data merged or combined in some fashion from two or more data sources, such as survey data from a statistical survey and administrative data from a program. Blended data are also known as integrated, hybrid, and multiple source data. See also: "Record linkage or data linkage."[5, 6]

Breach: The loss of control, compromise, unauthorized disclosure, unauthorized acquisition, or any similar occurrence where: (1) a person other than an authorized user accesses or potentially accesses protected data, including protected health information (PHI) or business identifiable information (BII), or (2) an authorized user accesses protected data for an other than authorized purpose.[7]

Business identifiable information: Trade secrets, commercial or financial information that is privileged or confidential.[8]

Confidentiality: A quality or condition accorded to information as an obligation not to disclose that information to an unauthorized party.[9]

Data: Recorded information, regardless of form or the media on which it is recorded.[9]

Data analysis: A process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.[10]

Data asset: A collection of data elements or datasets that may be grouped together.[9]

Data element: A unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes.[11]

Data governance: The process of setting and enforcing priorities for managing and using data as a strategic asset throughout the data life cycle, including the availability, usability, integrity, quality, and security of data. It is both an organizational process and a structure. It establishes responsibility for data, organizing program area/agency staff to collaboratively and continuously improve data quality through the systematic creation and enforcement of policies, roles, responsibilities, and procedures.[12, 13, 14]

Data integrity: The property that data have not been altered in an unauthorized manner. Data integrity covers data in storage, during processing, and while in transit. "Integrity" refers to the security of information—protection of the information from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification. Threats in the domain of integrity include lack of scientific integrity, political interference, and data security failures.[15, 16, 17]

Data policy: A set of broad, high-level principles which form the guiding framework in which data management can operate.[18]

Data protection: The application of techniques to ensure that confidential or sensitive information is not disclosed.[19]

Data provenance: The origins, custody, and ownership of research data. Because datasets are used and reformulated or reworked to create new data, provenance is important to trace newly designed or repurposed data back to their original datasets. The concept of provenance guarantees that data creators are held accountable for their work, and provides a chain of information where data can be tracked as researchers use other researchers' data and adapt it for their own purposes.[20]

Data quality: The degree to which data capture the desired information using appropriate technology in a manner that sustains public trust. This applies to the components of data files (e.g., variables, data fields), as well as to the entire data file.[17]

Data use agreement (DUA): Establishes who is permitted to use and receive restricted data and the permitted uses and disclosures of such information by the recipient. Typical DUAs may stipulate the following conditions: the recipient will (1) not use or disclose the information other than as permitted by the DUA or as otherwise required by law, (2) use appropriate safeguards to prevent uses or disclosures of the information that are inconsistent with the DUA, (3) report to the covered entity uses or disclosures that are in violation of the DUA, of which it becomes aware, (4) ensure that any agents to whom it provides the restricted data agree to the same restrictions and conditions that apply to the recipient, with respect to such information, and (5) not re-identify the information or contact the individual.[21]

Data standards: An agreed upon set of rules, requirements, or characteristics for governing architectural components of data. Data standards define the approach and practices for data representation, access, and distribution, and provide a common language to promote efficiency and comparability.[22]

Dataset: Any stored collection of information usually containing either individual reporting units, aggregated unit level data, or statistical manipulations of either individual level or aggregated data.[23]

Disclosure risk: The risk of identifying individual reporting units (e.g., persons, facilities, establishments) and information about them.[24]

Evaluation: (1) An assessment using systematic data collection and analysis of one of more programs, policies, and organizations intended to assess their effectiveness and efficiency. (2) Individual, systematic studies to assess how well an entire program or some specific strategy or an aspect of a program is working to achieve intended results or outcomes.[9]

Evidence: Information produced as a result of statistical activities conducted for a statistical purpose. This can include foundational research and analysis such as aggregate indicators, exploratory studies, descriptive statistics, and basic research (foundational fact finding); systematic analysis of a program, policy, organization or a component of these to assess effectiveness and efficiency (program evaluation); development, analysis, and reporting of performance measures and milestones related to the achievements of strategic goals and objectives (performance measurement); and analysis of data, such as general purpose survey or program-specific data, to generate and inform policy, e.g., estimating regulatory impact and other relevant effects (policy analysis).[9, 25]

Federal information: Information created, collected, processed, maintained, disseminated, disclosed, or disposed of by or for the Federal government, in any medium or form.[26]

Identifiable form: Any representation of information that permits the personally identifiable information to whom the information applies to be reasonably inferred by either direct or indirect means. See also, "Personally Identifiable Information."[9]

Information collection: The Federal Government collects a wide range of information from the public to carry out its day-to-day functions. Information collection can be in any format, including but not limited to verbal requests, general regulatory requirements, administrative forms, questionnaires, surveys, and other instruments, and also include record-keeping and reporting requirements.[27]

Information/data owner: An official with statutory or operational authority for specified information and responsibility for establishing the controls for its generation, collection, processing, dissemination, and disposal.[28]

Information security: The protection of information from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability.[29]

Interoperability: The ability of two or more systems or components to exchange information and to use the information that has been exchanged.[30]

Machine learning: The ability of computers to learn from provided data without being explicitly programmed for a particular task. Three types of learning can occur: supervised learning in which the machine analyzes past high-quality data and makes decisions about future data with the learned knowledge; unsupervised learning in which the machine makes inferences about future data based on patterns it finds within past data; and a combination of the two.[31, 32]

Machine-readable: Data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost.[9]

Metadata: Structural or descriptive information about data such as content, format, source, rights, accuracy, provenance, frequency, periodicity, granularity, publisher or responsible party, contact information, method of collection, and other descriptions.[9]

Nonstatistical purpose: The use of data in identifiable form for any purpose that is not a statistical purpose, including any administrative, regulatory, law enforcement, adjudicatory, or other purpose that affects the rights, privileges, or benefits of a particular identifiable respondent.[9]

Open Data: Publicly-available data structured in a way that enables the data to be fully discoverable and usable by end users, which means making the data available to the widest range of users for the widest range of purposes, often by providing the data in multiple formats for consumption.[33]

Open License: A legal guarantee that a data asset is made available at no cost to the public; and with no restrictions on copying, publishing, distributing, transmitting, citing, or adapting such asset.[9]

Personally Identifiable Information (PII): Information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual. See also: “Identifiable form.”[34]

Privacy: An individual’s right to control the acquisition, uses, or disclosures of his or her identifiable data.[35]

Privacy-Enhancing Technology: (1) any software solution, technical processes, or other technological means of enhancing the privacy and confidentiality of an individual’s personal data in data or sets of data; and (2) includes anonymization and pseudonymization techniques, filtering tools, anti-tracking technology, differential privacy tools, synthetic data, and secure multi-party computation.[36]

Privacy Impact Assessment: An analysis of how information is handled: (1) to ensure handling conforms to applicable legal, regulatory, and policy requirements regarding privacy, (2) to determine the risks and effects of collecting, maintaining and disseminating information in identifiable form in an electronic information system, and (3) to examine and evaluate protections and alternative processes for handling information to mitigate potential privacy risks.[26]

Public use file: A subset of data that have been coded, aggregated, or otherwise altered to mask individually identifiable information, and thus is available to all external users. Unique identifiers, geographic detail, and other variables that cannot be suitably altered are not included in public use data files.[19]

Protected Health Information (PHI): All "individually identifiable health information" held or transmitted by a covered entity or its business associate, in any form or media, whether electronic, paper, or oral, as defined by the HIPAA Privacy Rule.[37]

Record linkage or data linkage: A process, usually computer-based, that brings together information from two or more data files into a new combined file containing selected information about individual reporting units that were not available in the separate records.[38]

Records management: The planning, controlling, directing, organizing, training, promoting, and other managerial activities related to the creation, maintenance and use, and disposition of records, carried out in such a way as to achieve adequate and proper documentation of Federal policies and transactions and effective and economical management of agency operations.[39]

Security incident: An occurrence that: (1) actually or imminently jeopardizes, without lawful authority, the integrity, confidentiality, or availability of information or an information system; or (2) constitutes a violation or imminent threat of violation of law, security policies, security procedures, or acceptable use policies.[40]

Statistical activities: The collection, compilation, processing, or analysis of data for the purpose of describing or making estimates concerning the whole, or relevant groups of components within, the economy, society, or the natural environment; and includes the development of methods or resources that support those activities, such as measurement methods, models, statistical classifications, or sampling frames.[9]

Statistical agency: An agency or organizational unit of the executive branch of the Federal government whose activities are predominantly the collection, compilation, processing, or analysis of information for statistical purposes.[9]

Statistical purpose: The description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups; and includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support such purposes.[9]

Survey data: Data collection from a population of interest whose purposes include the description, estimation, or analysis of the characteristics of groups, organizations, segments, activities, or geographic areas. A survey may be a census or may collect information from a sample of the target population.[19]

Surveillance: The ongoing systematic collection, analysis, and interpretation of data tracked over time to detect patterns, disparities, and changes.[41]

Survey: An investigation about the characteristics of a given population by means of collecting data from a sample of that population.[42]

Synthetic data: Data that are artificially created to mimic real-world data. Data constructed from existing records on the basis of statistical models that induce noise in statistics relative to those from the original data, but whose approximate means and variances (and other pre-specified quantities) closely match those of the original dataset.[43, 44]

References

[1] **M-14-06 Memorandum for the heads of Executive Departments and Agencies**. Guidance for Providing and Using Administrative Data for Statistical Purposes: Office of Management and Budget 2014. https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf.

[2] **Artificial Intelligence (AI) Strategy**. U.S Department of Health and Human Services 2021. https://www.hhs.gov/sites/default/files/final-hhs-ai-strategy.pdf.

[3] **Title 44—Public Printing And Documents**. United States Code 2006. https://www.govinfo.gov/content/pkg/USCODE-2008-title44/pdf/USCODE-2008-title44-chap35-subchapIII-sec3542.pdf.

[4] **Statistical Policy Directive No. 4: Release and Dissemination of Statistical Products Produced by Federal Statistical Agencies**. Federal Register 2008. https://www.govinfo.gov/content/pkg/FR-2016-10-17/pdf/2016-25049.pdf.

[5] **Principles for Modernizing Production of Federal Statistics**. Office of Management and Budget. https://nces.ed.gov/fcsm/pdf/Principles.pdf.

[6] **Transparent Quality Reporting in the Integration of Multiple Data Sources: A Progress Report**: Federal Committee on Statistical Methodology 2017-2018. https://nces.ed.gov/fcsm/pdf/Quality_Integrated_Data.pdf.

[7] **M-17-12 Memorandum for Heads of Executive Departments and Agencies: Preparing for and Responding to a Breach of Personally Identifiable Information**. Office of Management and Budget 2017. https://osec.doc.gov/opog/privacy/Memorandums/OMB_M-17-12.pdf.

[8] **Privacy Program Plan**. Department of Commerce 2021. https://www.osec.doc.gov/opog/Privacy/Documents/PRIVACY_PROGRAM_PLAN.pdf

[9] **Public Law 115-435 - Foundations for Evidence-Based Policymaking Act of 2018**. 115th Congress 2018. https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf.

[10] **Data Analysis | Responsible Conduct in Data Management.** The Office of Research Integrity, U.S Department of Health and Human Services. https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html

[11] **Data Element Definition**. UNData Glossary 2008. http://data.un.org/Glossary.aspx?q=data+element

[12] **President's Management Agenda**. Data Governance Playbook: Federal Data Strategy 2020. https://resources.data.gov/assets/documents/fds-data-governance-playbook.pdf.

[13] **Data Governance Overview.** Institute of Education Sciences, U.S Department of Education. https://slds.grads360.org/#program/data-governance-overview.

[14] **Traveling Through Time: The Forum Guide to Longitudinal Data Systems.** National Center for Education Statistics 2011. https://nces.ed.gov/pubs2011/2011805.pdf.

[15] **Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems**. National Institute of Standards and Technology 2016. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160.pdf.

[16] **Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies; Republication**. Federal Register 2002. https://www.govinfo.gov/content/pkg/FR-2002-02-22/pdf/R2-59.pdf.

[17] **FCSM-20-04 A Framework for Data Quality**. Federal Committee on Statistical Methodology 2020. https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf.

[18] **Organization for Economic Co-operation and Development.** OECD Glossary of Statistical Terms: OECD Publishing 2008. https://stats.oecd.org/glossary/detail.asp?ID=4454 .

[19] **Standards and Guidelines for Statistical Surveys Appendix** Office of Management and Budget 2017. https://www.samhsa.gov/data/sites/default/files/standards_stat_surveys.pdf.

[20] **Data Provenance Definition**. Data Thesaurus: National Libraries of Medicine. https://old.nnlm.gov/data/thesaurus/data-provenance

[21] **Disclosures for Emergency Preparedness - A Decision Tool: Data Use Agreement**. U.S Department of Health and Human Services 2017. https://www.hhs.gov/hipaa/for-professionals/special-topics/emergency-preparedness/data-use-agreement/index.html.

[22] **Data Standards**. The Office of the National Coordinator for Health Information Technology. https://www.healthit.gov/playbook/pddq-framework/platform-and-standards/data-standards/.

[23] **Organization for Economic Co-operation and Development.** OECD Glossary of Statistical Terms: OECD Publishing 2006. https://stats.oecd.org/glossary/detail.asp?ID=542.

[24] **Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology**. Office of Management and Budget 2005. https://www.hhs.gov/sites/default/files/spwp22.pdf.

[25] **M-19-23 Memorandum for the Heads of Executive Departments and Agencies**. Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel and Planning Guidance: Office of Management and Budget 2019. https://www.whitehouse.gov/wp-content/uploads/2019/07/M-19-23.pdf.

[26] **Circular No. A-130 to the heads of Executive Departments and Agencies.** Office of Management and Budget 2016. https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/OMB/circulars/a130/a130revised.pdf.

[27] **Information Collections/Paperwork Reduction Act (PRA) Requirements.** U.S. Department of the Interior Indian Affairs. https://www.bia.gov/as-ia/raca/information-collections.

[28] **Minimum Security Requirements for Federal Information and Information Systems**: National Institute of Standards and Technology Computer Security Resource Center. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.200.pdf.

[29] **Standards for Security Categorization of Federal Information and Information Systems.** Federal Information Processing Standards Publication 2004. https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.199.pdf.

[30] **Cybersecurity Glossary**. National initiative For Cybersecurity Careers and Studies (NICCS) 2021. https://niccs.cisa.gov/about-niccs/cybersecurity-glossary.

[31] **Machine Learning Paired with Skilled Data Scientists is the Future of Data-Driven Decision Making.** Digital.gov 2017. https://digital.gov/2017/10/03/machine-learning-paired-with-skilled-data-scientists-future-datadriven-decision-making/.

[32] **Artificial Intelligence: Adversarial Machine Learning.** National Cybersecurity Center of Excellence: National Institutes of Standards and Technology. https://www.nccoe.nist.gov/ai/adversarial-machine-learning.

[33] **M-13-13 Memorandum for the Heads of Executive Departments and Agencies**. Open Data Policy—Managing Information as an Asset: Office of Management and Budget 2013. https://project-open-data.cio.gov/policy-memo/.

[34] **M-07-16 Memorandum for the Heads of Executive Departments and Agencies**. Safeguarding Against and Responding to the Breach of Personally Identifiable Information: Office of Management and Budget 2007. https://georgewbush-whitehouse.archives.gov/omb/memoranda/fy2007/m07-16.pdf.

[35] **June 22, 2006 Letter to the Secretary – Recommendations regarding Privacy and Confidentiality in the Nationwide Health Information Network.** National Committee on Vital and Health Statistics 2006. https://ncvhs.hhs.gov/rrp/june-22-2006-letter-to-the-secretary-recommendations-regarding-privacy-and-confidentiality-in-the-nationwide-health-information-network/.

[36] **Congressional Record Senate S7542.** GovInfo 2021. https://www.govinfo.gov/content/pkg/CREC-2021-11-01/pdf/CREC-2021-11-01-pt1-PgS7542.pdf#page=6.

[37] **The HIPAA Privacy Rule**. U.S. Department of Health and Human Services. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html.

[38] **Record Linkage Definition**. UNData Glossary 2008. http://data.un.org/Glossary.aspx?q=record+linkage.

[39] **Annual Records Management Training.** Department of Health and Human Services (HHS). https://humancapital.learning.hhs.gov/courses/2020recordsmanagement/a001_what_is_records_management_what_is_records_management.html.

[40] **Title 44—Public Printing And Documents**. United States Code 2014. https://www.govinfo.gov/content/pkg/USCODE-2014-title44/pdf/USCODE-2014-title44-chap35-subchapII-sec3552.pdf

[41] **Evaluating Obesity Prevention Efforts: A Plan for Measuring Progress.** National Academies Press 2011. https://www.ncbi.nlm.nih.gov/books/NBK202495/.

[42] **Organization for Economic Co-operation and Development.** OECD Glossary of Statistical Terms: OECD Publishing 2008. https://stats.oecd.org/glossary/detail.asp?ID=2620.

[43] **HHS Announces Synthetic Health Data Challenge Winners**. U.S. Department of Health & Human Services 2021. https://www.hhs.gov/about/news/2021/09/21/hhs-announces-synthetic-health-data-challenge-winners.html.

[44] **Federal Statistics, Mulitple Data Sources, and Privacy Protection: Next Steps.** The National Academies Press 2017. https://mitsloan.mit.edu/shared/ods/documents?PublicationDocumentID=4439.