Definitions of *Dataset* in the Scientific and Technical Literature

Allen H. Renear[1], Simone Sacchi[2], Karen M. Wickett[3]
Center for Informatics Research In Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street, MC-493, Champaign, IL 61820-6211
[1]renear@illinois.edu, [2]sacchi1@illinois.edu, [3]wickett2@illinois.edu

ABSTRACT
The integration of heterogeneous data in varying formats and from diverse communities requires an improved understanding of the concept of a *dataset,* and of key related concepts, such as format, encoding, and version. Ultimately, a normative formal framework of such concepts will be needed to support the effective curation, integration, and use of shared multi-disciplinary scientific data. To prepare for the development of this framework we reviewed the definitions of dataset found in technical documentation and the scientific literature. Four basic features can be identified as common to most definitions: *grouping, content, relatedness*, and *purpose*. In this summary of our results we describe each of these features, indicating the directions a more formal analysis might take.

Keywords
Dataset, Data Curation, Information Organization.
Data Conservancy.

INTRODUCTION

> "There needs to be an explicit statement of what the intended preservation of a dataset will imply. There is ambiguity in what type of object a dataset is; with different groups of users applying different connotations." (Pepler, 2008)

The concept of a *dataset* is common to almost every scientific discipline where data provide the empirical basis for research activities. Yet there has been little analysis of this central concept. Although the term occurs routinely in articles, papers and reports, as well as informal conversation among scientists, there is no precisely specified established definition. Nevertheless the term dataset appears to be unproblematic in its common use, suggesting that there is at least a general shared understanding — and in fact our examination of the literature, summarized below, clearly identifies a set of re-occurring themes.

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

At the same time this examination also reveals that usage varies considerably, raising some uncertainties about whether a single shared concept can be precisely defined. This is not surprising: thoroughgoing precision in definitions is often not practical. Rather than insist on agreement on subtleties and distinctions that are difficult to make precise, it is usually more efficient to rely on the informal and general understanding of a disciplinary community. Additional distinctions can then be negotiated as needed. Not only can individual disciplinary communities operate easily with an informal understanding of basic concepts, cross-disciplinary communication as well is facilitated by loosely defined notions.

Nevertheless the lack of a precise common definition that is shared across disciplines can create curatorial problems for multi-disciplinary digital data repositories. These repositories are intended to integrate data from many sources in order to solve real-world multi-disciplinary problems and must present their resources in a uniform common context.

In what follows we summarize some preliminary results of a project examining definitions of dataset in the scientific literature, technical documentation, and information processing standards (Sacchi, 2010). Our examination has already revealed a core set of common features. This project is a first step in preparation for the development of a normative formal framework of precisely defined and related definitions that is both intrinsically coherent and that serves the needs of practicing scientists and the institutions and services which support data-intensive scientific research.

This research is being conducted as part of the Data Conservancy, a multi-institutional NSF funded initiative hosted at Johns Hopkins University Sheridan Libraries. The Data Conservancy is building infrastructure for the management of digital research data. The Data Concepts team, located at the Center for Research in Science and Scholarship, at the University of Illinois Urbana-Champaign, is developing a formal framework of fundamental data concepts. This work will provide the foundation for standardizing how Data Conservancy datasets are identified, described, related, and organized.

COMMON FEATURES OF DATASET DEFINITIONS

We found that most definitions of dataset have four features: *grouping, content, relatedness, purpose*.

Grouping

Grouping terms like *set, aggregation, container,* and *collection* are routinely used to indicate that datasets are data treated collectively, as a unit. Importantly, these terms often occur as the nouns in categorical expressions that suggest that this feature identifies the fundamental kind of thing a dataset is (e.g. in a phrase such as "a dataset is a collection of…"). However exactly what sort of entity is intended by the grouping terms is often not clear, and can vary. In any ontology the assignment of datasets to some particular class of entity will be critical for subsequent modeling and reasoning. Several different meanings seem possible:

[Mathematical] Set: In some cases, particularly when the word "set" is used, it seems likely that set in the mathematical sense is intended. For instance, a "set of RDF triples […]" (Alexander, 2009), a "set of records" (Purchase, 2008), or a "set of numbers" (DAS2 Glossary, 2009). The common assumption that different data mean different datasets would be consistent with this interpretation: mathematical sets cannot, strictly speaking, lose or gain members. If mathematical set is an appropriate characterization, then we can expect set theoretic notions (*element of*, *subset of*, etc) to apply.

Collection: The term "collection" is widely used in defining datasets (Alexander, 2009; Liguang, 2007; OECD Glossary of Statistical Terms, 2006; Pepler, 2008; Toupikov, 2009; United Nations Statistical Division, 2000). *Collection* seems to suggest that the addition or deletion of data does not imply a change in dataset identity, which would be inconsistent with the mathematical concept of a set. If datasets are collections in this sense, there is no accepted system of logical notions that can be applied, as the conceptual nature of collection as an ontological category remains unclear. The concept of collection also suggests that there is an intentional collecting of the constituents of a dataset.

Containment: Terms of *containment* are also used to indicate grouping. For instance "A DataSet contains the acquired or derived data" (Lohrey, 2009). It is possible that this suggests a distinctive entity, different from sets or collections. However words related to "containment" are used very generally (both sets and collections are sometimes said to "contain" their members) and so from the use of containment terms alone, without further evidence, one cannot infer that an entity other than set or collection is intended.

Plural Reference: In some cases it is actually not entirely clear whether the intention is to treat data collectively, or to make a "plural reference" (McKay, 2006) to the data. For instance, "Data Set — Digital data and its metadata derived from any research activity" (LTER General Data Use Agreement, 2005) might be interpreted as referring to the data and not anything apart from the data, such as a dataset that contains the data.

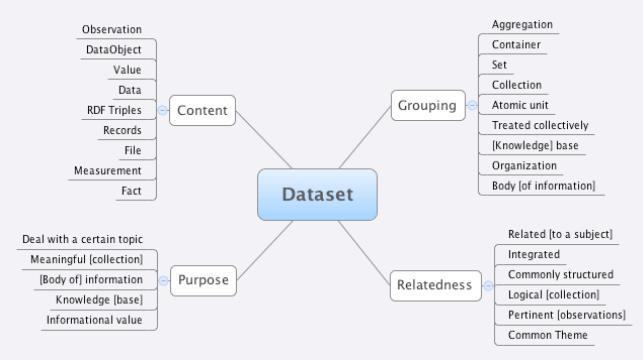

Figure 1: Conceptual map of dataset features indicated by words and phrases in definitions in the literature.

Content

Although the term "data" is sometimes used without any additional qualification to indicate the contents of datasets, most definitions imply that the constituents of a dataset are things of some particular kind. The data in datasets are variously described with terms such as "observations" (Feeley, 2004; Purchase, 2008), "facts" (McDermott, 2001), "values" (United Nations Statistical Division, 2000), and "records of values" (Purchase, 2008). Descriptions such as "observations" and "facts" indicate content of a propositional sort, while "values" or "records of values" might be understood as being representations of propositional content.

Datasets sometimes have parts that are themselves conceptualized as datasets. Although it is not clear whether the nested datasets are strictly speaking, components of datasets, or whether nesting simply means that the components of the nested dataset are components of the including dataset. If the former, then dataset nesting corresponds to a transitive relation (like *subset of*, or *part of*), but if the latter then the nesting relationship will be intransitive (like *member of*).

Typically the content of a dataset is intended to reflect the results of certain sorts of activities, such as measuring or observing. In particular what is recorded are observations (Feeley, 2004) or "the results of" observations (Purchase, 2008).

Of obvious importance is a considerable variation in the level of abstraction at which dataset contents are conceived. In some places these contents appear to be abstract conceptual entities (observations, property values), and in other places particular representations of those entities (records of values, XML elements), or even lower level entities (files for instance).

These distinctions are of course often explicit in discussions of the nature of datasets: "we adopt notations from which we can derive methods to read the physical representation of a dataset into an abstract one, and vice versa to write the abstract representation into a physical one" (Moreau, 2005). Nevertheless ambiguity and variation with respect to levels of abstraction make it clear that a formal framework will not be making explicit an existing univocal concept of dataset, or even recommending revisions in an existing concept, but will need to develop a family of related notions to replace a term that is used in a variety of senses.

Relatedness

It is evident from these definitions that datasets are thought of as grouping together constituents (data) that are related to each other in some way that goes beyond both the grouping itself, and the identification of the grouped things as all being of the same general kind of entity. We refer to this further commonality as the *relatedness* condition. Several kinds of relatedness can be identified.

Circumstantial Relatedness: A dataset is sometimes thought of as consisting of data related by time, place, instrument, or object of observation. These features draw attention to the circumstances around the creation or maintenance of a dataset as opposed to any internal features of the data: "[data] originate from a certain source or process, are hosted on a certain server, or are aggregated by a certain custodian" (Alexander, 2009); "[data] should specify the context in which the observations or measurements were obtained. The context may include, for example, the place and the time of observation or measurement, and the object or group of objects observed" (Purchase, 2008); or again "Data having mostly similar characteristics (source or class of source, processing level and algorithms, etc.)" (NASA Earth Observatory Glossary, n.d.).

Syntactic relatedness: Data in a dataset are typically expected to have the same syntactic structure (records of the same length, field values in the same places, etc). For instance, "All records of a dataset are assumed to have a common structure, with each position having its specific meaning, which is common to all values appearing in it" (Purchase, 2008); data in a dataset are "sharing a structure" (OECD Glossary of Statistical Terms, 2006).

Semantic relatedness: Data in a dataset may be about the same subject or make assertions similar in content. For example datasets are said to "deal with a certain topic" (Alexander, 2009), have constituents that "[relate] to a single subject" (United Nations Statistical Division, 2000), or have "a common theme" (Pepler, 2008). Also Feeley (2004) suggests this relation where he qualifies the term "observations" with the adjective "pertinent". This suggests that some sort of thematic coherence should be expected in the data included in a dataset.

These different kinds of relatedness are not disjoint. In fact they describe – at different levels of abstraction – the peculiar cohesion that characterizes a dataset as a unit and its intrinsic normative adherence to a common circumstantial, semantic, and syntactic pattern.

Purpose

Beyond the immediate objective of recording information datasets have a larger distinctive intended application as well. They are clearly created in order to contribute in some way to scientific activity. This might be by providing evidence to be analyzed, suggesting new hypotheses, providing refutation or confirmation of existing hypotheses, or supplying new phenomena to be explained.

Curiously, although indication of this distinctive scientific purpose is routine in dictionary definitions of dataset, it is not as often explicitly included in definitions in the scientific and technical literature, where, presumably it is implicit in the general context of the article or report. Nevertheless it is indirectly in evidence: "A dataset represents a knowledge base. […] The knowledge base is more than just the sum of its parts: by itself, it is of informational value whether a piece of information belongs to a dataset or not." (Cyganiak, 2008), indicates that datasets are meant to carry information that sustains scientific investigation. Similarly "[a dataset] represents an organization of pertinent experimental observations, their uncertainties, and mechanistic knowledge of a subject of interest" (Feeley, 2004) also anticipates the use of this information in the scientific process.

FINDINGS

Our examination of some explicit definitions of dataset in scientific and technical literature reveals that:

1. There are common themes in dataset definitions, suggesting that there is some degree of agreement and shared understanding, at least at a high level of generality.
2. More specifically, these definitions usually exhibit in some form these four characterizing features: *grouping, content, relatedness,* and *purpose*.
3. Although definitions of dataset do appear to fit a common pattern, with recurring phrases and semantically similar terms, it is clear that there is no single well-defined concept of dataset. The variations in individual terms are significant, the terms themselves are often used in different senses, and critical characteristics are left underdetermined.
4. In particular there is uncertainty as to the ontological status of datasets, and considerable ambiguity and conflation with respect to level of abstraction of dataset contents.
5. It is clear from the forgoing that the notion of "dataset" found in the literature cannot itself be provided with a precise formal definition, but that this general notion must be replaced by an interrelated family of more specific concepts.

NEXT STEPS

The curation and integration of scientific data from multiple sources and disciplinary communities will require a shared framework of dataset concepts. These concepts must make all needed distinctions and be precisely and formally defined. A review of the variety of dataset definitions in the literature, along with empirical studies of data practices, supply necessary first steps towards this framework, revealing the community expectations that must be accommodated.

But these empirical studies are only preliminary. Developing a normative framework of dataset concepts requires a thoroughgoing formal analysis of the modeling and representation issues, confirmed and shaped by iterative testing in real interdisciplinary scientific data repositories. This is the next phase of our project.

ACKNOWLEDGMENTS

The research reported here is being carried out at the Center for Research in Informatics and Scholarship (CIRSS) at the University of Illinois at Urbana-Champaign, Carole L. Palmer, Director. It is funded by the National Science Foundation as part of the Data Conservancy, a multi-institutional NSF funded project (OCI/ITR-DataNet 0830976) hosted at Johns Hopkins University Sheridan Libraries. Other members of the CIRSS Data Conservancy group contributing to the analysis of dataset concepts are David Dubin, Tiffany Chao, and Melissa Cragin.

REFERENCES

Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2009). Describing Linked Datasets-On the Design and Usage of voiD, the'Vocabulary of Interlinked Datasets'. In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09).

Cyganiak, R., Stenzhorn, H., Delbru, R., Decker, S., & Tummarello, G. (2008). Semantic sitemaps: Efficient and flexible access to datasets on the semantic web. In *Proceedings of the 5th European Semantic Web Conference on the Semantic Web.* (Tenerife, Canary Islands, Spain, June 01-05, 2008). S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds. *Lecture Notes in Computer Science*, 5021, 690.

Data set. (n.d.). In NASA Earth Observatory Glossary. Retrived May 30, 2010 from http://earthobservatory.nasa.gov/Glossary

Data set definition (2006). In OECD Glossary of Statistical Terms. Retrieved May 30, 2010 from http://stats.oecd.org/glossary/detail.asp?ID=542

Dataset. (2009). In DAS2 Glossary. Retrieved May 30, 2010 from http://das2.org/wiki/index.php?title=Das2.glossary

Feeley, R., Seiler, P., Packard, A., & Frenklach, M. (2004). Consistency of a reaction dataset. *J. Phys. Chem. A*, 108(44), 9573–9583.

Liguang, M. A., Yanrong, C. A. O., Jianbang, H. E., & PR, C. (n.d.). Study on Data Management and Sharing Service Based Metadata and Dataset Concept A Case Study in Environment Sciences and Ecology Area.

Lohrey, J. M., Killeen, N. E., & Egan, G. F. (2009). An integrated object model and method framework for subject-centric e-Research applications. *Front. Neuroinform.* 3(19), 1-10 doi:10.3389/neuro.11.019.2009

LTER General Data Use Agreement. (2005) In LTER Network Data Access Policy, Data Access Requirements, and General Data Use Agreement. Retrived from http://www.lternet.edu/data/netpolicy.html

McDermott, D., Burstein, M., & Smith, D. (2001). Overcoming ontology mismatches in transactions with self-describing service agents. In *Proc. Int'l Semantic Web Workshop*.

McKay, T. J. (2006), *Plural Predication*, New York: Oxford University Press.

Moreau, L., Zhao, Y., Foster, I., Voeckler, J., & Wilde, M. (2005). XDTM: The XML data type and mapping for specifying datasets. In *Advances in Grid Computing (European Grid Conference)*. *Lecture Notes in Computer Science*, 3470, 495.

NASA. (1986). *Earth observing system. Data and information system*. Volume 2A: Report of the EOS Data Panel. NASA Technical Memorandum, Document ID: 19860021622. http://ntrs.nasa.gov

Pepler, S. J., & O'Neil, K. (2008). *Preservation intent and collection identifiers*: CLADDIER Project Report II. Retrieved May 30, 2010 from http://epubs.cclrc.ac.uk/work-details?w=43640

Purchase, H. C., Andrienko, N., Jankun-Kelly, T. J., & Ward, M. (2008). Theoretical foundations of information visualization. In *information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, et al, Eds. *Lecture Notes In Computer Science*, vol. 4950.

Sacchi, S., Wickett, K. M., & Renear, A. H. (2010). *Dataset definitions*. Champaign, IL: Center for Informatics Research in Science and Scholarship. (Rep. No. CIRSS/DATACONS--2010/1/VER01+DCDC)

Toupikov, N., Umbrich, J., Delbru, R., Hausenblas, M., & Tummarello, G. (2009). DING! Dataset Ranking using Formal Descriptions. *In WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*, Madrid, Spain.

United Nations Statistical Division. (2000). *Handbook on geographic information systems and digital mapping*. Studies in methods, no. 79. New York: United Nation.