What dataset descriptions *actually* describe

Applying the Systematic Assertion Model to connect theory and practice

Karen Wickett, Andrea Thomer, Simone Sacchi, Karen Baker, Dave Dubin
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

A dataset is a symbol structure that expresses data content. The record identifier picks out the record in the context of the dataset.

The example record appeared originally as a row in a spreadsheet packaged as a Darwin Core Archive.

Data Content is propositional content expressed by a dataset that is justified by an observation or computation.

In the example, data content includes the fact that a specimen of *Mola mola* was collected at a particular time and place. This fact is expressed through the connection between the species identifier, the specimen identifier, and the collection event information.

Understanding this fact from the record depends on common expectations about how fields in a record are used together to encode information.

Methodological Metadata gives information about things like the tools used in collection of data, preservation of a specimen, or software used for processing data. Its presence is crucial for assessing the credibility of a dataset within a community.

Provenance Metadata gives information about the assertion events that result in data content.

In this example, the asserting agents for the collection event and the identification event are recorded. Although they are the same person, these are two distinct events important to understanding the provenance of the data.

Some data content is justified on the basis of a **computational event**.

In this case, decimal longitude and latitude were computed on the basis of the recorded information. Similarly, date information was processed and broken out to allow different views of the dataset.

Access Metadata is included in a dataset to increase the likelihood that scientists will find records that are of interest to them. For species occurrence records it may include broader scale information about locality, such as country or region.

Scientific vocabularies describe different dimensions of scientific data collection and communication processes. We present an analysis of a species occurrence record based on the Systematic Assertion Model [1] *to offer a more precise account of Data Content and Metadata as they appear in scientific datasets*

What the record elements express	Example Record
Metadata (dataset) Record identifier	`id: 1821`

Data Content Species identifier	`scientificName: Mola mola`
	`family: Molidae`
	`order: Tetraodontiformes`
	`class: Actinopterygii`
	`genus: Mola`
	`specificEpithet: mola`

Metadata (provenance) Identification event	`identifiedBy: Wiley, Martin` `basisOfRecord: preservedspecimen`

Data Content: Specimen identifier	`collectionCode: KUI`
	`institutionCode: KU`
	`catalogNumber: 32586`
	`fieldNumber: MLW 34`

Metadata (methodological) Details on the specimen	`preparations: 1 EtOH`

Data Content Collection event identification	`verbatimEventDate: 1/8/65`
	`verbatimLongitude: 75.9000 W`
	`verbatimLatitude: 34.1217 N`
	`locality: Atlantic Ocean, about 100 mi. E of Carolina Beach, North Carolina`
	`minimumDepth: 31`
	`maximumDepth: 31`

Metadata (provenance) Collection event	`recordedBy: Wiley, Martin L`

Data Content (derived) Through computation	`year: 1965`
	`month:1`
	`day: 8`
	`eventDate: 1/8/65`
	`decimalLongitude: -75.9`
	`decimalLatitude: 34.1217`

Metadata (access) Aid finding the record	`continent: Atlantic Ocean`
	`country: USA`
	`StateProvince: North Carolina`

Thanks to Laura Russell and http://vertnet.org for assistance with the example record.

References

[1] Dubin, D., Wickett, K. M., & Sacchi, S. (2011). Content, Format, and Interpretation. In B. T. Usdin (Ed.), *Proceedings of Balisage: the Markup Conference 2011*, Balisage Series on Markup Technologies (Vol. 7). Montréal, Canada. doi:10.4242/BalisageVol7.Dubin01

DataConservancy

Center for Informatics Research in Science & Scholarship CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois