ERIM Terminology V4

Introduction

When the members of the ERIM Project started to investigate (engineering) research data and their better management for re-use, it became apparent that there was no existing terminology that allowed them to discuss with clarity the subject in hand.

To overcome this difficulty, a terminology is in the process of being evolved to provide a means of clearer communication. Many of the terms are defined to increase discrimination of like concepts on the assumption that different management implications will apply at the level of detail chosen. Also, a higher resolution for modelling follows from greater discrimination.

It is intended that the selection and definition of the terms should be validated by the extent to which they prove useful to and are adopted by the community interested in the management of research data for the purpose of making them more re-usable for research.

It is hoped, then, that the community will contribute to the evolution of the terminology such that it might, in due course, become normative. To this end discussion is invited and use and developmentg of the terminology in its current form by interested parties is encouraged.

The terms as they currently exist and are defined are given below. For the purposes of the discussion an explanation and examples are given as necessary for clarification

A useful rubric for lexica has been coined by Felix Cohen; it is in the spirit of this that the terminology is presented:

Once we recognize that a definition is, strictly speaking, neither true nor false but rather a resolution to use language in a certain way, we are able to pass the only judgment that ever needs to be passed on a definition, a judgment of utility or inutility. (Cohen, 1950)

Copyright

Notes

The terms are not strictly in alphabetical order but have been grouped loosely into logically associated clusters; the clusters are demarked by horizontal lines.

The terms and their definitions have come from a number of sources. Where possible the provenance of a term that has been borrowed is given against the entry.

The 'data discovery verbs' coined and defined by the Australian National Data Service (ANDS) are not included here. This is partly because a number of their definitions conflict with those here, and partly because their verbs are task-specific (providing a structure and high-level architectural device to support services necessary to sharing) and our task (managing data for the purposes of re-use and re-purposing) is somewhat different and more diverse.

Amongst the terms is a set of words defined which can be applied in noun form (when related to an act or process being carried out on data – for example association or augmentation) or as a verb (for example, associate or augment). An understanding of the meaning of both forms of the word should be apparent from the accompanying definition.

The Terminology

Data. Reinterpretable representations of information in a formalized manner suitable for communication, interpretation or processing. (CCS02)

Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen. (CCS02 Reference Model)

Information. Any type of knowledge that can be exchanged. In an exchange, it is represented by Data. An example is a string of bits (Data) accompanied by a description of how to interpret a string of bits as numbers representing temperature observations measured in degrees Celsius (Information). (after CCS02 Reference Model)

Note 1: the 'accompanying description' referred to above could be substituted by Knowledge appropriate to achieving the same interpretation.

Note 2: There is a logical tension apparent in the definitions of Data and Information in respect of the enterprise of making Research Data re-usable. Research Data are only re-usable if the Information or Knowledge is (made) available for correct interpretation to follow, in which case strictly it is 'Information' not 'Data'.

Knowledge. The stock of Information, skills, experience, beliefs, and memories existing in the head of the individual. (after Van Beveren, 2002)

Metadata. Data about Data. (CCS02)

The term 'metadata' is often used in a rather restricted way, for example limiting the scope to keywords, terms populating (XML, HTML) tags and so on. It is intended here to be interpreted to embrace 'Information about Data' and any means of describing or setting in context other Data or Data Record.

Research Activity. The process through which Research Data and Context Data are accumulated and developed.

Research Data Development Process. One of a set of processes that are commonly carried out during the Research Activity which changes or adds to the Research Data associated with a research activity or project.

Data Case. The set of Data Records associated with some discrete Research Activity (project, task, experiment, etc.).

Record. Information in any medium, created, received and maintained as evidence of an activity.

Data Record (DR). The Data Object which contains the Data.

Research Data. Data that are descriptive of the research object, or are the object itself.

Research Data Record (RDR). A Record containing Research Data.

Context Data. Data that support the Research Activity but do not describe the research object nor are the research object itself.

Context Data Record (CDR). A Record containing Context Data.

Associative Data Record. A Context Data Record containing Context Data which records the association between other data records or data.

Research Object Data Record. A Data Object (being either a physical or electronic data record) containing data which is, itself, the object of the research enquiry.

Experimental Apparatus Data Record. A Data Object (being either a physical or electronic data record) containing symbolic representations which are functionally analogous to the physical experimental apparatus familiar in much laboratory-based research.

Child Record. A Data Record identified as having been created from the precursor (Parent) record through any of the Data Development Activities.

Parent Record. A Data Record which is the precursor of one or more successor (Child) records which have been created from or associated with the precursor through any of the Data Development Activities.

Data Object. Either a Physical Object or a Digital Object. (CCS02)

Digital Object. An object composed of a set of bit sequences. (CCS02)

Manifestation. The way in which the intangible underlying Data (contained in a file) is embodied (i.e. presented for interpretation).

Note: Underlying Data can be manifest in different ways; e.g. as a spreadsheet, as a graph, or as an HTML or PDF page.

Addition (Add). To supplement existing Data at the Data level, for example when inducting Data into an existing file.

Aggregate. To combine similar Data from different sources for the purpose of increasing sample size (cf. Augment).

Annotate. To add 'information or additional marks formulated on a document for enhancing it with brief and useful explanations' (Evrard & Virbel, 1996).

Note: annotation may be made on the subject document itself or on a 'stand-off' document with the annotation linked to the annotated content.

Associate. To make explicit the relationship between items of Data, Data Records or Data Cases.

Example: Two Data Records may be Related by such things as file names, metadata, explicit reference in one to another, by an embedded link.

Augment. To add Research Data Records or Context Data Records to a Data Case.

Collect. To acquire and bring together pre-existing Data.

Note: This is concordant with the DCC's use of the term 'receive' in relation to pre-existing data collected from external sources.

Duplicate. To make another, identical, copy of a file.

Data Cleaning. A special case of Refinement, where the Refined Data are a normalized version of the source Data; that is, with systematic errors corrected, calibration taken into account, invalid Data removed, etc.

Collate. Give order to Data assembled from different sources.

Note: This can occur at any organizational level, that is at the Data level, the Data Record level or the Data Case level.

Delete. To expunge or obliterate.

Pretty self-explanatory, but rather important in relation to Data management.

Derive. (cf. Refine) To create new Data by applying logical inference, extrapolation, or similar algorithmic process to pre-existing Data.

Note 1: Derivation constitutes creating a new description or representation of an analysis of the subject described by or being the existing Dat
.

Examples: a histogram representing the frequencies of occurrence; a narrative or commentary on a set of interviews.

Note 2: NASA's data-processing terminology includes the use of the term 'derivation' with a similar interpretation; however it seems to conflate the three concepts of 'Derivation' and 'Refinement' and 'Manifestation'. There 'derived data' are defined as 'derived results, as maps, reports, graphs, etc.'; this being data processed through 'NASA Process Levels 2 though 5'. There is some argument to say that 'Derivation' and 'Refinement' occupy different positions on the same continuum. However, they seem to be different in character and thus may have different management implications. If it is found that they do not the two concepts could usefully be conflated (as in the NASA definition).

Extract. To make a new Research Data Record from portions of the Data in one or more Research Data Records.

Format Migration (Migrate). The transfer of digital information from one format to another (with the intention of preservation of the full information content). (DCC glossary)

Generate. To act on or interact with a research subject thereby creating Research Data.

Populate. To add content in the form of attribute values to an existing framework. Such frameworks include the database, table and form.

Recording (data). Encoding the output from Data Generation in a carrier format (e.g. bitstream, written notes).

Refine. To re-express Data in a different form or according to a different Data model (cf. Derive).

Example of typical refinement functions: rounding, normalization, removing duplicates, stop-words, noise or outliers, simplification, etc.

Transform. To Derive or Refine Data in such a way that new or changed Data results.

Note: this interpretation excludes Format Migration because in Migration – in contrast to Derivation and Refinement – there is expressly no intention to change the Data.

Initiating Event. Any happening subject to research scrutiny and which is the precursor to the recording of data as manifest in a Data Record. Examples of events include interviews, meetings, design episodes and so on.

Initiating Object. Any object subject to research scrutiny and which is the precursor to the recording of data as manifest in a Data Record. Such objects include physical objects from which data are generated or from which data is obtained by observation or, singularly, a Research Object Data Record.
First-Generation Data. The Data resulting either from Data Collection or from Data Generation.

This term is intended to identify Data which has not been the subject of Derivation or Refinement.

Data Rawness. The inverse measure of the number of processing steps leading to the creation of a set of Data.

Information Loss. Removal of Information from an instance or set of Data.

Examples are such things as rounding down or up of real numbers to integers, deletion of the record of units in a data set, or disassociation of Context Data and the Data they explain.

Information Gain. Addition of Information to an instance or set of Data, for example when Aggregating or Annotating.

Function Loss. Removal of or reduction in the capacity to compute or manipulate.

An example is migrating the contents of a live spreadsheet to a PDF format where the content stays the same, but the facility changes.

Function Gain. Increase in the support for computation or manipulation of Data.

An example is transferring the Data in a hand-written sheet into a spreadsheet providing such facilities as ordering, summing, etc. Likewise, Function Gain is a characteristic of Format Migration through optical character recognition.

State Loss. Deletion or discarding of earlier Data state(s) or version(s).

State loss occurs either as a result of discarding unregarded Data or overwriting an existing version of the Data. It characteristically occurs when the item of interest is the final outcome of an iterative process, for example in the application of the Delphi Method or when continuous updating of Data occurs, for example in the automatic updating of computer code in response to closed-loop feedback.

Related. Two or more items of Data or Data Records which have an implicit or explicit connection. Explicit connections are made through Association.

Process Repeatability. A measure of the practical possibility of a Generation process being repeated such that Data Reproducibility is possible in principle.

Data Reproducibility. Data are reproducible if they can be regenerated through repeat of a Generation process such that their functional content remains the same.

Knowing the in-principle ability or non-ability to reproduce (or re-Generate) Data is important for Data management; it impinges not only on considerations of Data acquisition and maintenance but also experimental repeatability, inference validity and so on. However, the interpretation of repeatable is legitimately variable dependent on process and requirement. In some disciplines strict repeatability is a requirement for experimental and inferential validity, in others the concept is a non sequitur. In some processes, identical input will produce identical output. Some processes, given identical input, will produce the same average output, from which the same conclusions can be drawn. For some processes the concept of identical input and output are inappropriate concepts; what will be of interest is whether for the same general input conditions the same interpretation or inference can be drawn.

Reversible. True of a process P if and only if there can exist, at least in principle, a process P' which, when given P(I) as input, produces I as output.

Non-Reversible. True of a process P if and only if there cannot exist, even in principle, a process P' which, when given P(I) as input, produces I as output.

Data Use. Using Research Data for the current Research Activity or purpose to infer new Knowledge about the research subject.

Data Re-use. Using Research Data for a Research Activity or purpose other than that for which it was intended.

Supporting Data Re-use. Managing existing Research Data such that it will be available for a future *unknown* Research Activity.

This concept is one for which no verb has been coined. It is one of a set which also includes data 'Purposing' and 'Re-purposing'. It combines many of the activities implied in the verbs 'archive', 'preserve' and 'curate'.

Data Purposing. Making Research Data available and fit for the current Research Activity.

This is the activity all researchers are familiar with when making Research Data available for their own research.

Data Re-purposing. Making existing Research Data available and fit for a future *known* Research Activity.

Note: This definition emphasizes the activity as being one of explicit intention, and thus differs (as does the spelling) from the definition used for 'repurposing' in the Data Documentation Initiative's Combined Life Cycle Model for research data, viz.:

> *Repurposing: The data may also be used within a different conceptual framework; examples include sampling or restructuring the data, combining the data with other similar sets, or producing pedagogic materials.*

RAID Diagram. A 2-D graphical representation – based on the RAID Modelling method – of the Data Records within a Data Case and their associations and principal characteristics. The RAID diagram visualizes a part of the data contained within the RAID Record.

RAID Modelling. A method of modelling the development of Data during the Research Activity which identifies Data Records, their relations, associations and Metadata .

RAID Record. The underlying record of a Data Case containing information about Data Records constituting the case, their temporal ordering, relations and association and principal characteristics. Also a data file in which the record is contained identified as such by the RAIDmap application data type extension *.rmap*.

References

Burton, A. & Treloar, A. (2009). Designing for Discovery and Re-Use: the 'ANDS Data Sharing Verbs' Approach to Service Decomposition. *International Journal of Digital Curation,* 4(3).

CCS02. Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. Blue Book CCSDS 650.0-B-1. Also published as ISO 14721:2003. URL: http://public.ccsds.org/publications/archive/650x0b1.pdf.

Cohen, F.S.(1950) Field theory and judicial logic, *Yale Law Journal,* 59, pp.238-272.

Evrard, F. and Virbel, J. (1996) Realisation d'un prototype de station de lecture active et utilisation en milieu professionnel. Rapport du contrat, 9300571, ENSEEIHTINPT, Toulouse.

Van Beveren, J. (2002). A Model of Knowledge Acquisition that Refocuses Knowledge Management, *Journal of Knowledge Management,* 6 (1), pp.18-22.

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a Standard for the Social Sciences. *International Journal of Digital Curation,* 3(1).