Definitions

Collection: A user defined group of datasets.

Comment: High level information associated with a file or dataset left by users.

Content Based Retrieval: A means of indexing collections of data where instead of indexing by text or keywords items are indexed by signatures and users query the collection with example files in order to retrieve files with similar contents.

Content Management System: A system used to store, manage, and curate collections of files and datasets.

Data Extraction: A transformation that creates new data from the given data. An example would be the execution of analysis code on an image file's contents to determine if a face is in the image. Clowder utilizes extractions to automatically generate metadata, signatures, and previews from a file's contents and provide users with means of finding, relating, and utilizing data that may be difficult otherwise.

Dataset: A group of files that through some defined relationship or corresponding metadata are strongly tied together and not representable otherwise by the individual files.

Extractor: A tool which takes a file, section of a file, data set, or collection as input and through some analysis of the contents produces some higher level information, e.g. metadata, or other derived product, e.g. preview, to aid users in searching/organizing data (both automatically and/or manually).

File: The lowest level unit of information that can be tracked. This is a file from a file system.

Metadata: Simply data about data. Available on datasets and individual files.

Preview: Special representation of a dataset or a file used by a previewer to visualize information about the dataset or file on the web. Often used to provide a smaller version of a dataset or file when bandwidth is a consideration.

Section: A subset of a files contents (e.g. a sub-image, a line from a document, a frame from a video, etc...). A sections is tied to a file.

Signature: A typically numerical representation for some semantic aspect of a files contents. This can be thought of as a hash of the files contents. Various means of generating these signatures are typically available and focus on different aspects of a files data (e.g. color distributions in an image vs edge distributions). Signatures are used in content based retrieval to index and find similar data to a given example.

Space: A group of collections, data sets, and files with defined user access rights.

Tag: A short string, e.g. one or two words, associated with a file or data set used to categorize or index its contents.

Technical Metadata: Automatically generated metadata produced by the system via extractors.

User Metadata: Metadata associated with file or dataset, entered by a human user.

Versus: A framework for decomposing content based comparisons into reusable parts that can be mixed and matched to meet a variety of user needs when content based indexing and retrieval is a viable means of allowing users to search a collection of data.

Versus Metadata: Signatures, typically numerical in nature, generated by versus to represent some semantic aspect of a files contents. Used for content based retreival.