# Metadata and Mass Storage

**Richard R. Lee**
Data Storage Technologies, Inc.
Post Office Box 5073
Banner Elk, North Carolina 28604-5073
Tel: +1-704-963-7773
Fax: +1-704-963-7779
E-mail: rrl@dst.com.

## Abstract:

Mass Storage systems in the multi-TeraByte class today are stretching the limits of existing OS file systems and data management tools. Given the anticipated growth of the data centers' appetitie (60% CAGR), these multi-TB archives will grow into multi-PetaByte size by the year 2000. In order to meet this growth challenge the mass storage and OS communities will have to develop comprehensive Metadata standards. This effort is presently underway in the engineering and scientific community via large national programs with PB+ archives, but will also need the support and nurturing of the commerical marketplace as it is only a few short years behind in terms of this same need.

## Overview:

Metadata is by no means a new concept. It is widely used in today's relational database systems (data dictionaries), as well as in many large data gathering systems (NASA's EOSDIS will use it to capture each data granules unique history and identification information), but it is relatively new in respect to being applied to Data Management and storage systems. There are no specific constraints on the contents of the metadata. It can be any information required by the OS or data management system.

One of the main benefits of metadata is to allow end-users to "Browse" large data archives and repositories with complex queries without having to look at each file via only its descriptors.

In regards to storage systems and data management there are a number of types of metadata available with the two most prominent being; System-level and Application-level.

## Systems-Level Metadata:

Systems-level metadata provides additional functionality within a native file system and helps manage the bitfiles controlled by that file system. It includes physical information regarding the bitfiles size, locations, access-time requirements and pointers as to where to best store the bitfile within the hierarchy of storage devices available (in respect to providing performance within the constraints of that user's "class of service" authorization (class of service is determined by order of importance of the job, cost-per-unit stored, and other resource allocation parameters)).

This type of metadata can provide a HSM (Hierarchical Storage Manager) with details regarding usage patterns and access characteristics, performance requirements, and device selection parameters. This will then allow the HSM to best choose the type of storage device for storing, retrieving and migrating that bitfile. It will also provide information on bitfiles spanning multiple volumes and those associated with families of bitfiles and how to best co-locate them for the most efficient storage, retrieval and migration operations.

In the future it is envisioned in the that storage systems will evolve into a layered architecture (in line with the ISO OSI model), with systems-level metadata residing in the same layer as data management, effectively bridging-the-gap between the storage system devices and drivers and the native file system of the host platform(s) and its applications.

**Application-Level Metadata:**

Application-level metadata is used to enhance the user's access, management and manipulation of his or her bitfiles. It uses abstracts to indicate the type of data in the bitfile (what application created it, along with the type of file it is), what information is contained within it (a snapshot of the data for searching and browsing) and its relationship (if any) to other bifiles. Application-level metadata is best located on the highest-availability storage media for immediacy of access and browsing.

Although not directly associated with the storage and retrieval processes, this type of metadata provides increased efficiency in terms of managing end-user data as well as locating bitfiles without tying up entire archives and repositories with onerous file-by-file searches and complex queries.

**Standards:**

As the need for increased operational efficiencies and improved bitfile access has become more paramount to the success of such developments as the Digital Library and the WWW; Government , Academia and Industry have joined forces to develop a set of comprehensive Metadata Standards and Reference Models. These efforts have been designed to dovetail with other activities such as the IEEE Mass Storage Systems Reference Model, the ISO/NASA Digital-Archiving Information Systems Reference Model, and the OCLC Dublin Metadata Core Element Set in order t o provide for maximum interopability with systems in development already. **Figure 2** represents a proposed reference model for the relationship of metadata to the other elements of the system.

Although these models and the groups that support them have sometimes disparate requirements in terms of a final standard, all are participating in the development of these reference models and strawman standards to help flush out the numerous technical and implementation issues that must be addressed for any of them to be successful.

**"What does all this mean to me?"**

As enterprise computing becomes highly distributed and new data intensive client/server applications become the norm, the need to efficiently access and manage large geographically disperse data centers will soon become an everyday struggle. Current data management technologies and methodologies does not support either of these tasks adequately at all.

One of the technologies being pursued to alleviate this crisis is Metadata. Metadata offers incremental improvements in data management capabilities, along with increased efficiencies in the search, browse and retrieval operations that are becoming more critical each day within the data center.

Metadata can also control ever growing storage costs across the enterprise by better balancing of media type (cost-per-unit stored) vs. access time required; and processing costs (CPU cycles and network bandwidth) vs. user time (shorter waiting times for data equates to higher efficiency overall) when browsing and retrieving datasets from these data centers.

The data center of the future will range in size from 100 TeraBytes+ to Multiple PetaBytes based on current growth projections. It will contain millions of varying types of bitfiles (numerous databases, text, audio, video, etc.), and will be accessed by thousands of people each day via enterprise Intranets and other such networks. This can only be accomplished by dramatic increases in operational efficiency, such as that contributed by Metadata.

**Summary:**

Metadata has been in use by the database industry since the 1960's. It 's use in respect to mass storage is fairly new, but it is critical to the success of such national programs such as EOSDIS, Digital Libraries, etc. Numerous groups are involved in developing comprehensive standards for these programs, with the commercial computing community keeping close tabs on this progress. These universal standards will fill the gaps in today's OS's and Data Management tools, while providing "system harmony" to the entire enterprise desiring access its archives.

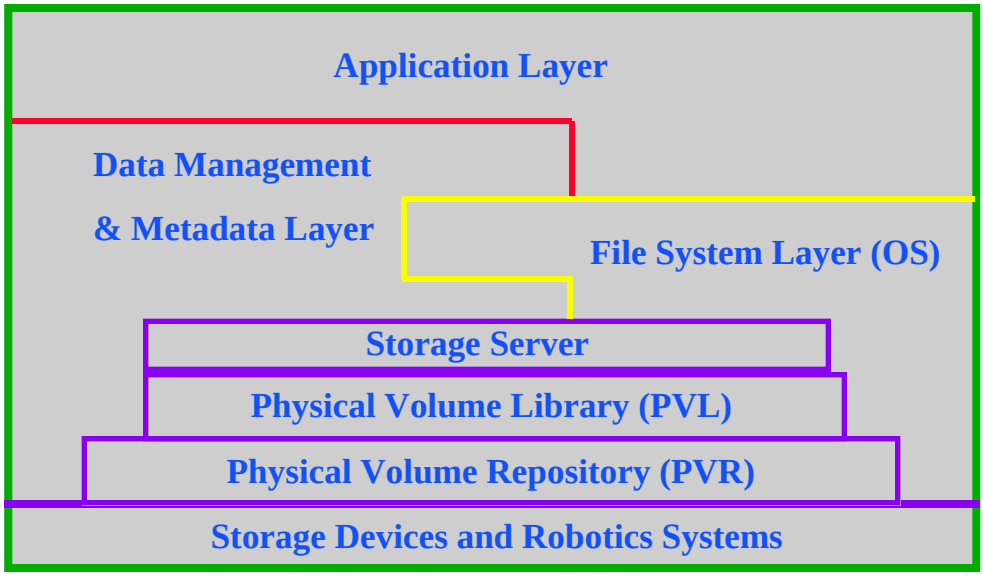
**References:**

1.- "A Reference Mode for Metadata" - Strawman 3/24/94, Francis Bretherton, Univ. of Wisconsin
2.- "Draft Whitepaper on Data Management", Draft 2/10/94, Robyne M. Sumpter, LLNL
3.- "OCLC/NCSA Metadata Workshop Report", 3/95, Stuart Weibel et al, OCLC On-line Computer Library Center, Inc.
4.- "NOAA-GFDL Metadata Issue Briefing", 7/94, Richard R. Lee, DST, Inc.
5.- "Resource Discovery and Use in a Distributed Digital Library: Metadata Issues", 11/95, Avra Michelson, Mitre Corp., Library of Congress Network Advisory Committee
6.- "The Use of Metadata in the Storage Environment", June 1996, Richard Lee, Storage Management Solutions

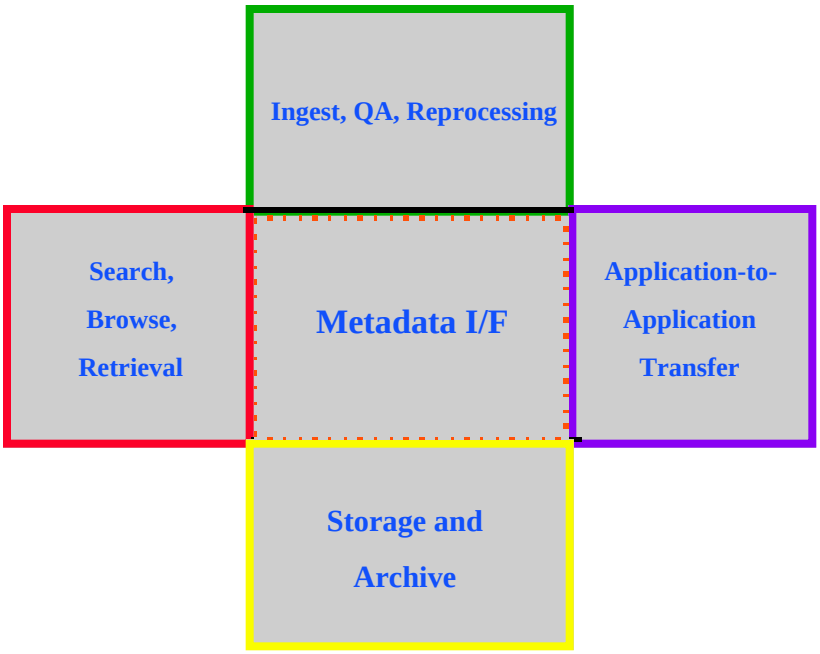
**Further Reading:**

1.- "IEEE Metadata Conference Proceedings" April 16-18, 1996, NOAA, Silver Spring, MD
2.- "The Metadata and Data Management Information Page" (http://www.llnl.gov/liv comp/metadata/metadata.html)

3.- “Bibliography for Metadata” web site (http://www.llnl.gov/liv comp/metadata/biblio.html)
4.- “CNRI “D-Lib Magazine”, July 1995” (http://www.dlib.org/)
5.- “Metadata Requirements for Evidence”, 10/95, David Bearman et al (dbear@lis.pitt.edu)
6.- “ISO Archiving Standards” web site (http://www.gsfc.nasa.gov/nost/isoas/overview.html)




**- Figure 1 -**
**“MSSRM & Metadata ISO Model”**




**- Figure 2 -**
**“Metadata Interaction Scheme”**