SNIA
Advancing storage &
information technology

A Data Protection Taxonomy

June 2010

Mike Dutch
EMC Senior Technologist
SNIA Data Protection and Capacity Optimization Committee

Table of Contents

Introduction 1
The taxonomy in context 1
Aspects of data protection 2
What is “data protection”? 3

Taxonomy Overview 4
An example of using the taxonomy 5

Who 6
Market Segments 6
Sales Channels 6

Where 7
Location Type 7
Data Relocation 7

What 8
Device Type 8
Data Type 9
Content Type 9
Operational Environment 10

Why 11
Operational Recovery 11
Disaster Recovery 12
Retention and Preservation 14
Governance, Risk Management, and Compliance (GRC) 15

How 15
Protection Technology 16
Data Copy Services 16
Data Security Services 17
Storage Technology 19
Resiliency 19
Capacity Optimization 20
Power Management 21
Access Technology 22
Management Technology 24

Summary 25

About the Author 25

About the DPCO Committee 25

About the SNIA 25

List of Figures

Figure 1. Data Protection Taxonomy 4
Figure 2. Data Protection Taxonomy Example 5

Introduction

Once upon a time, I/O devices were viewed as peripherals, existing only to satisfy the whims of the master processing unit. Then the magical wizard of business cast a spell upon the data processing glass house, transforming it into a *data center*, observing that information is the lifeblood of commerce. Protecting that data became essential to business survival and data protection was likened to an insurance policy, a necessary cost of doing business. Now, the magical wizard of business is once again casting spells, transforming the vassalage of backup and recovery into a reservoir of competitive advantage.

Can data protection really help companies thrive in the real world or is this story just a fairy tale? How can business people make sense of even the buzzwords they are bombarded with, much less of the alleged solutions which technologists are so fond of offering at a dizzying pace? The answer to these questions of course depends on what data protection means and on whether you use technology to implement a strategic model of data protection or simply use it as a tool to react to the latest crisis.

The purpose of any taxonomy is to enable the subject matter to be studied better. This paper presents a data protection taxonomy, that is, a classification of aspects relevant to data protection, that the industry is encouraged to use as a reference when asking and answering questions about data protection. This taxonomy is defined in terms of who, where, what, why, and how.

The paper begins by putting the taxonomy in context, suggesting a broad view of what data protection is, providing a high level view of the taxonomy, and showing an example of how to use it. This is followed by a drill down into each category of the taxonomy to detail subcategories and provide illustrative comments.

The taxonomy in context

Let's begin by noting the obvious… that a taxonomy is not an end unto itself. It is intended to help make considered decisions; it does not replace the need to ask and answer fundamental questions. The following example is indicative of the context in which a taxonomy fits. It also alludes to the fact that it is important to not make a decision more complicated than it needs to be.

As an example, suppose the merits of an acquisition are being presented to a company's decision makers. A taxonomy discussion could represent a brief but critical element of a technology review used to establish the boundaries of the subsequent analysis. In this example, the taxonomy is used within the context of identifying a business opportunity and how to best provide the capabilities to address it. After reviewing the market size and dynamics, the taxonomy can be used to identify missing capabilities required to deliver a solution for specific use cases. At this point, the build/buy landscape can be evaluated for relative strengths and weaknesses, and a best fit recommendation is made based on differential advantage analysis.

When this taxonomy is used to assess whether a company has particular capabilities in-house, it can also be used to identify opportunities for cross-functional cooperation. This ability to identify where elements of a solution may be obtained can help us operate effectively within dynamic organizations.

Aspects of data protection

Some key aspects considered by our data protection taxonomy include:

Usability

- Consistent recovery (backup and recovery, snapshots)
- Multiple PiT recovery (versioning, CDP)

Accessibility

- No single point of failure
- Multi-platform/interface/environment
- Remote access
- Replication (migration, disaster recovery)

Performance

- Multipathing
- Caching
- Hardware acceleration

Security

- Authentication, Authorization, Accounting
- Information Rights Management
- Data Leakage Prevention
- Encryption (on the wire, at rest)

Compliance

- Retention and preservation
- Legal discovery/hold

Resilience

- Media protection (Redundant Array of Independent Disks, erasure codes)
- CPU protection (Redundant Array of Independent Nodes, High Availability, Fault Tolerance)
- Data integrity (data is not altered or destroyed in an unauthorized manner)

Efficiency

- Policy-based automation
- Information Lifecycle Management, tiering, classification
- Capacity/Performance Optimization
- Power management

Manageabilty

- Central administration
- Real time monitoring and alerting
- Historical reporting and trend analysis
- Consumability (licensing, billing)

What is “data protection”?

Somewhat surprisingly, the scope of what “data protection” should encompass is not obvious. Making and managing a copy of data doesn’t sound particularly challenging but it remains one of the most common pain points for businesses and consumers despite decades of technological progress. A broad view of data protection is defined here in the interest of promoting a comprehensive yet easy to understand definition.

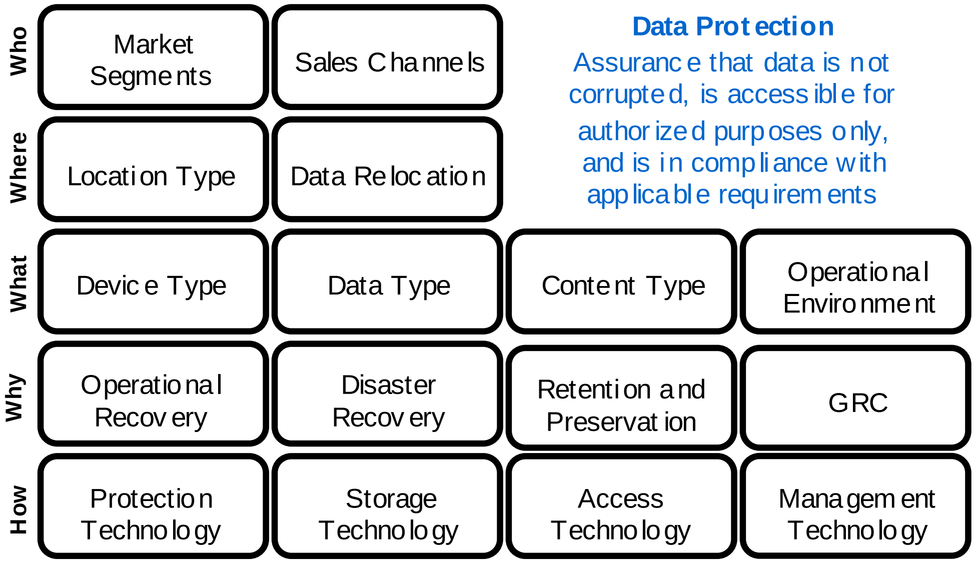
Data Protection means assurance that data is not corrupted, is accessible for authorized purposes only, and is in compliance with applicable requirements. Protected data should be usable for its intended purpose. Usability may require that steps be taken to provide data integrity, application consistency, versioning, and acceptable performance.

This definition of data protection goes beyond the notion of data availability, defined as the amount of time that data is accessible by applications during those time periods when it is expected to be available. Unacceptable performance can lower productivity levels such that access to applications and related data is effectively unavailable. Note that data security and compliance issues are also intimately involved as the ultimate goal of data protection is to reduce risks, costs, and downtime while increasing business value and agility.

This view supports traditional backup and archive processes but acknowledges that alternative approaches exist. By focusing on what the business seeks to achieve rather than on how IT has traditionally delivered its services, the range of potential services within the scope of data protection is expanded. This doesn’t imply we should explore all peripheral issues in depth; the objective is to avoid being blindsided by emerging technologies.

Taxonomy Overview

Figure 1 is a high level overview of the data protection taxonomy. It is represented as boxes which represent distinct lenses through which to view a data protection solution.



Who	Market Segments	Sales Channels		
Where	Location Type	Data Relocation		
What	Device Type	Data Type	Content Type	Operational Environment
Why	Operational Recovery	Disaster Recovery	Retention and Preservation	GRC
How	Protection Technology	Storage Technology	Access Technology	Management Technology

Data Protection
Assurance that data is not corrupted, is accessible for authorized purposes only, and is in compliance with applicable requirements

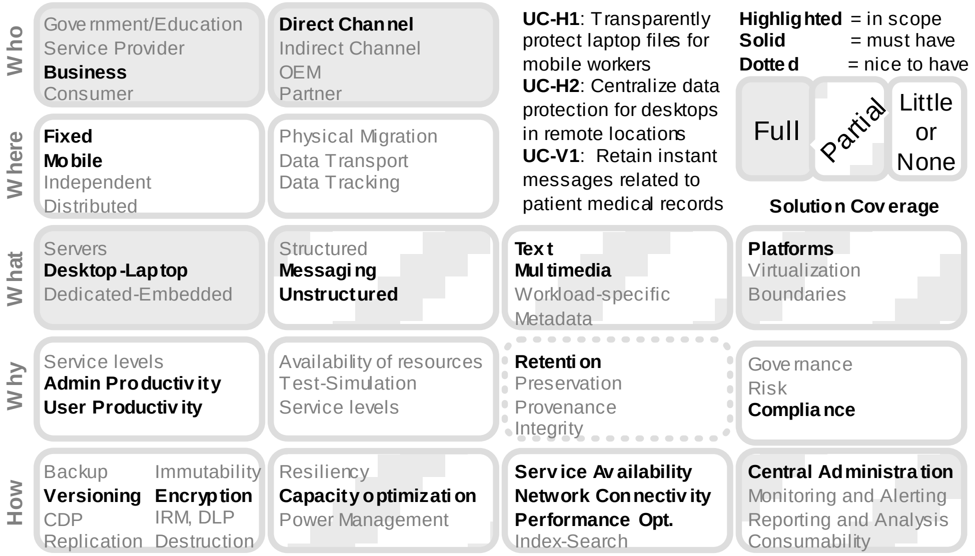
Figure 1. Data Protection Taxonomy

Each lens is independent of every other lens. For example, if you are looking through the protection technology lens, you don't have to worry about sub elements "bleeding" into the view of the device type lens. Of course, there are many relationships between these lenses and the taxonomy encourages examination of these relationships. In fact, by looking through multiple lenses, unexplored aspects of a solution may become apparent; ultimately a solution may be made more complete and relevant to the customer.

Each row of boxes addresses a particular question, namely who, where, what, why, and how. Each lens categorizes a straightforward notion and both the high level category and its subcategories avoid the use of unnecessary jargon.

An example of using the taxonomy

Figure 2 is an example of using the taxonomy when providing a data protection solution for a desktop or laptop computer.



Who	Government/Education Service Provider **Business** Consumer	**Direct Channel** Indirect Channel OEM Partner		
Where	**Fixed** **Mobile** Independent Distributed	Physical Migration Data Transport Data Tracking		
What	Servers **Desktop-Laptop** Dedicated-Embedded	Structured **Messaging** **Unstructured**	**Text** **Multimedia** Workload-specific Metadata	**Platforms** Virtualization Boundaries
Why	Service levels **Admin Productivity** **User Productivity**	Availability of resources Test-Simulation Service levels	**Retention** Preservation Provenance Integrity	Governance Risk **Compliance**
How	Backup **Versioning** CDP Replication Immutability **Encryption** IRM, DLP Destruction	Resiliency **Capacity optimization** Power Management	**Service Availability** **Network Connectivity** **Performance Opt.** Index-Search	**Central Administration** Monitoring and Alerting Reporting and Analysis Consumability

UC-H1: Transparently protect laptop files for mobile workers
UC-H2: Centralize data protection for desktops in remote locations
UC-V1: Retain instant messages related to patient medical records

Highlighted = in scope
Solid = must have
Dotted = nice to have

Figure 2. Data Protection Taxonomy Example

This solution is intended to address three use cases. The first two are considered horizontal or functional use cases because they apply across a range of industries. Both are considered must-have requirements which the solution should address before being shipped. The third use case covers a nice-to-have requirement specific to the health care industry.

The first use case is to transparently protect laptop files for mobile workers. Unlike server-centric data protection solutions, the focus is on maintaining the productivity of the laptop users. The second use case is to centralize data protection for desktops installed in remote locations. The focus here is on centralizing administration so a single storage administrator can manage backup and recovery operations across all the remote offices in the enterprise. The third vertical use case involves retention services for instant messages related to patient medical records.

The first difference between Figures l and 2 is that subcategories are shown in each box of Figure 2. This is because the high-level categories we just reviewed do not provide sufficient granularity to detail the market environment, functional requirements, or available resources to satisfy the solution. These subcategories will be presented subsequently for each taxonomy lens. However the current purpose is to show how this taxonomy can be used, not to explore this example solution, set of use cases, or to explain the subcategories within each lens.

The second difference is the shading of each box. A fully shaded box indicates that a company has the required capabilities in-house. A partially shaded box indicates that a company has some but not all of the required technology to satisfy the solution. An unshaded box indicates that a company has little or none of the required capabilities.

Once the requirements of a solution are understood, two actions are performed. First, the lens subcategories within each box that are relevant to the solution are highlighted. In this example, it is clear by the bold face highlighting that this desktop-laptop data protection solution is targeted to business customers via the direct sales channel. Second, the "must have" capabilities are distinguished from the "nice to have" capabilities by outlining the box with solid or dotted lines respectively. In this example, the relatively lower priority of providing medical record IM retention services is indicated by the dotted outline.

Let's review the purpose of this paper again before diving down into details. We're going to categorize various aspects of data protection to provide a broad view of its scope and provide terminology so customers and vendors can discuss data protection based on a common understanding. Our goal is to help people ask and answer fundamental questions about data protection so they can make informed decisions that best fit their particular situation.

Who

Let's begin our "deep dive" into each lens of our data protection taxonomy. The first row of boxes identifies who the customer is ("Market Segments") and who will sell the solution ("Sales Channels").

Market Segments

The subcategories of the "Market Segments" category are government, education, service provider, business, and consumer. When highlighting a specific opportunity, feel free to use whatever category makes sense. This taxonomy is intended to stimulate your thinking, not to make it more rigid.

We'd also encourage right brain thinking[1]… random, intuitive, holistic/synthesizing, and subjective. This is not to say that left brain thinking… logical/sequential, rational, analytical, objective is not valuable. The point is that standardization of categories is useful but don't be limited by it.

Sales Channels

Sales channels are important because they drive delivery of a product or service. The Sales Channel subcategories listed here are not definitive. For example, while the term "Original Equipment Manufacturer" (OEM) theoretically refers to a company that manufactures a product or a component, it may also refer to a company that purchases a component for use in its products or a company that simply rebrands another company's product (which may also be termed value-added resellers (VAR) or resellers respectively). The term "GSI" refers to Global System Integrators. The term "Partner" is a

[1] Right Brain / Left Brain: What Is It All About? <http://painting.about.com/od/rightleftbrain/a/Right_Brain.htm>

catch-all that can include a competitor, such as when a company licenses its technology for use by a competitor. Whether "Service Providers" should be categorized separately, or even subcategorized, is a business unit prerogative.

Where

The second row of boxes identifies where the solution will be deployed ("Location Type") and where data may need to be located ("Data Relocation").

Location Type

There are four data location types: fixed, mobile, independent, and distributed. The category is called "location type" to emphasize the point that there may be multiple locations of the same type.

Most locations are "fixed" currently. A fixed location is a specific known location where a solution is run. This means it may be relevant where data processing occurs. The fixed location subcategories of unconstrained and constrained are actually introducing notions that are not strictly relevant to location type. However, in the interest of distinguishing an IT glass house from a center of business operations, the separation is meaningful. "Datacenter" and "Remote Location" are considered examples of a fixed location type rather than as types of locations themselves. This acknowledges that a datacenter is not necessarily larger than a remote location and avoids confusing the term "remote location" with the very different notion of remote access.

If the location is not fixed it may be "mobile." This typically refers to edge devices like smart phones where a location can be determined but may dynamically change.

Cloud computing introduces the notion that regardless of whether the computing or data location is known, the processing does not depend on knowing where the data resides, that is, it is independent of the location. Of course, location-awareness is an important consideration in many circumstances.

Finally, a specific dataset may reside and/or be processed on multiple systems, each of which may be distributed geographically and managed separately.

Data Relocation

Physical Migration refers to moving information from one physical system or location to another or from one physical media-format to another, such as from an older generation tape drive technology to a new higher density tape drive technology. It may also be used to refer to moving data between tiers of storage to maintain physical readability, accessibility, and integrity, or to achieve other storage and efficiency benefits. Note that while the movement of data between tiers may be classified as physical migration, the policies used to trigger such migration may be classified as Information Lifecycle Management or may be multi-tier storage system policies that may or may not be externalized to the customer. Additionally, when virtual machines are moved between physical hosts, access to their data needs to be maintained. When storage is not shared between these physical hosts, physical migration will be required.

Note that logical migration is a different notion than data relocation.

Logical migration refers to moving information from one logical-format to another, such as from an old application version to a new version, to preserve readability, interpretability, and integrity. Logical migration converts the content and representation information of an information object into a new information object (a transformation) and maintains an audit trail of the change, documenting the conversion event.

Data Transport refers to creating a copy of data and storing it at another location. It differs from physical migration in that an additional copy is maintained. Relocating data may involve the shipment of physical media (sometimes called sneaker-net or TAM (truck access method) or may be performed by sending sufficient information electronically to create a physical copy offsite. It is important to distinguish the process used to transport the data, not just the end result, because the impact on specific use cases varies considerably. For example, the risk of losing or damaging physical tape media during transport is eliminated by using electronic vaulting. Another common example of transporting data is for seeding full backups or replicas when network bandwidth is constrained. In this case, a copy may be created at the source location, physically transported to the destination, and from then on updated by electronically sending only changed data.

Data Tracking refers to controlling and tracking the creation of data copies at another location. Beyond creating backup copies on physical tape in native or optimized formats, capabilities that help customers efficiently validate the recoverability of data from offsite locations, including cloud storage, are essential to maintain confidence in dynamic storage infrastructures.

What

The third row of boxes in Figure 1 identifies what types of devices are addressed (“Device Type”), what types of data are addressed (“Data Type”), what types of content are addressed (“Content Type”), and what types of platforms, virtualization, and boundaries (“Operational Environment”) are addressed by the solution.

Device Type

The Device Type refers to the solution delivery vehicle and correlates to particular market segments. Traditionally, it refers to a server or a personal computer in desktop or laptop form factors. It can also refer to devices dedicated to and optimized for specific server, network, or storage roles; these are termed appliances and may be physical or virtual (which means running in a virtual machine). Edge and consumer devices such as mobile phones, tablets, home theater PCs, audio/video components like digital video recorders, and other devices such as internet picture frames may also present unique challenges and opportunities for data protection.

Data Type

There are several reasonable ways to categorize data type.

One way is to categorize data types is in quadrants, indicating whether the data is subject to change or not (dynamic or fixed) and whether its format and content are prescribed or not (structured or unstructured).

Databases are structured data. Dynamic databases support transactional applications like Online Transaction Processing, Enterprise Resource Planning, Customer Relationship Management, Content Management Systems, collaborative applications, and Social Media. Fixed content databases are used by Business Intelligence and Analytics applications.

Filesystems are unstructured data. Dynamic files include Virtual Machine disk images and files used by Office, CAD, and media production applications. Read-only files are used for archive repositories, medical imaging, broadcast and audio/video applications.

Messaging data such as email, instant / real-time messages, text (SMS) / multimedia (MMS) messages, and voice data has characteristics of both unstructured and structured data. Semi-structured data may also be created by augmenting unstructured data with metadata.

Another way to categorize data types distinguishes how storage is deployed in traditional IT environments and in emerging content depot environments. Within traditional enterprises, data types are classified as structured and unstructured as before, but also identifies replicated copies of the active data used for backup, test, analytics, or retention purposes. Content depots provide access to large quantities of fixed content. Content depots include consumer-focused companies, telecommunications companies, eDiscovery companies, and cloud-based IT companies.

Content Type

Content type provides an application-centric lens into the data. Whereas data type focuses on how bytes are used or formatted, content type focuses on the context and meaning of the data… in other words, categorizing the information conveyed by the data rather than the data bytes themselves.

Many of the examples mentioned when discussing the data type lens illustrate how specific applications use databases and filesystems. When the topic concerns compliance with data formats, the data type lens is most appropriate. However, when the information conveyed by the data is the topic under consideration, the content type lens may be more useful when considering opportunities.

The subcategories of content type reflect three sub-views of the content:

- Text and Multimedia reflect how applications use dynamic and fixed content respectively.
- Workloads move the focus from the data type to the application type. It is intended to engender brainstorming opportunities in very different market segments. Business applications include file and print serving IT infrastructure as well as B2B and B2C requirements. Scientific

applications often leverage high performance computing clusters. Workgroup applications concentrate on collaboration requirements. Consumer markets will prioritize ease of use and scalability solution requirements.

- Metadata may be system-generated or user-generated[2]. System-generated metadata may or may not be visible to the end-user. User-generated metadata may be created by an application (e.g., Office document metadata and Exchangeable Image File Format (EXIF) photo metadata) or explicitly by the end-user (e.g., "folksonomy" tagging).

Operational Environment

The final "WHAT" lens of the taxonomy is termed Operational Environment. It is intended to spawn ideas related to product and service deployment.

The Platform subcategory is straightforward. It indicates the machine architecture and the software stack to be supported by the data protection solution. Different platforms reflect different needs, priorities, and skill sets. For example, whereas support for hierarchical filesystems is expected in open systems environments, support for cataloged data sets is the norm in mainframe data centers.

The virtualization subcategory reflects the trend toward abstracting all levels of the IT infrastructure, not just hardware components like memory or servers. It is intended to explicitly recognize differences in how physical and virtual assets are managed. It includes application containers (such as partitioning, virtual machines, and desktop virtualization), storage virtualization (including logical volumes and file virtualization), and connectivity solutions (such as virtual I/O and virtual network switches).

The Boundaries subcategory is a powerful notion intended to evoke deployment models that may not otherwise be explicitly considered.

Administrative boundaries may be the most familiar sort of boundary. It means the set of resources over which an IT administrator has control over and visibility into. Administrative boundaries are often set to facilitate delegation of responsibility, separation of duty, and scale operations. For example, one person may be responsible for data backup at each location while another person may be responsible for reporting service levels across all locations.

Ownership boundaries may not sound as familiar because they are often implicit. It simply means that the owner of an asset is free to set policies on how that asset can be used. When the IT infrastructure is owned by a single person or organization, operational procedures can be coordinated if there is motivation to do so. However, if the IT infrastructure used to accomplish a task has multiple owners, it may be impossible to resolve differences in policy or to even have visibility into what those policies are. For example, consider the SETI@home distributed computing project. Any individual can

[2] Storage Industry Resource Domain Model <http://www.snia.org/education/storage_networking_primer/sirdm/>

download software that allows units of work to run as a background process that uses idle computer power. After computation on the work unit is complete the results are reported back to the SETI@home servers for consolidation. Public cloud[3] service providers provide a similar example on a more commercial basis. Ownership boundaries are important because they may directly impact acceptable solution architectures.

Regulatory boundaries reflect operational differences due to legal or voluntary constraints on IT practices. For example, separate governments may impose not only different but contradictory requirements concerning data retention. If one government requires one type of record be securely deleted within 1 year of its creation but another government requires the same type of record be retained for 10 years, an IT infrastructure subject to both jurisdictions that is used to store such records will need to reconcile its data retention practices. Notice that regulatory boundaries need not span geographies as data subject to different regulations can be stored on the same device.

There are also interesting privacy issues related to boundary issues relevant to data protection. For example, suppose a company stores their backup data at a remote location for safekeeping. If the data at this remote location is legally confiscated (such as by police with a warrant), the company's backup data may also be searched.

Why

The "WHY" aspect of data protection, the forth row of boxes in Figure 1, includes operational recovery, disaster recovery, retention and preservation services, and governance, risk management, and compliance (GRC) services.

Operational Recovery

SNIA defines operational recovery and backup as follows:

- Operational Recovery is the recovery of one or more applications and associated data to correct operational problems such as a corrupt database, user error or hardware failure.
- A backup is a collection of data stored on non-volatile storage media for purposes of recovery in case the original copy of data is lost or becomes inaccessible.

The definition of operational recovery highlights the need to resolve operational problems onsite. The backup definition highlights the obvious but often overlooked reason for backing up data… and that is to use it to recover data so normal business operations can continue.

An IT organization or service provider will be tempted to focus on the first subcategory, namely, that the service levels in an advertized service catalog should be agreed to, measured, and reported.

[3] SNIA Cloud Storage Reference Model <http://www.snia.org/forums/csi/>, NIST Working Definition of Cloud Computing <http://csrc.nist.gov/groups/SNS/cloud-computing/index.html>

However, it is important not to forget that the ultimate objective of operational recovery is human productivity. Computers are tools, not ends unto themselves.

Server recovery-centric data protection offerings focus on making the administrator productive and minimizing server downtime. Unfortunately, maximizing server utilization may be at odds with business productivity when administrator-productivity and business user-productivity do not coincide.

User-centric data protection offerings focus on making the consumer productive. This requires an astute understanding of end-user requirements. Transparent data protection for mobile workers suggests that sales force teams engage their customers to intimately understand their data protection objectives.

Disaster Recovery

"Disaster Recovery" means data processing must be moved to an alternate site with its own infrastructure in order to resolve a service level impact or to mitigate a devastating event. Such disasters are relatively rare and many customers have never experienced a disaster in this sense. However, a prudent company will take proactive steps to ensure business continuity, treating data protection as an insurance policy on their mission-critical data. Note that while operational recovery is within the control of the IT organization, disaster recovery is the responsibility of the entire business and can be a people-intensive process.

SNIA defines disaster recovery and key recovery service level metrics as follows:

- Disaster Recovery is the recovery of data, access to data and associated processing through a comprehensive process of setting up a redundant site (equipment and work space) with recovery of operational data to continue business operations after a loss of use of all or part of a data center.
- Recovery Time Objective (RTO) is the maximum acceptable time period required to bring one or more applications and associated data back from an outage to a correct operational state.
- Recovery Point Objective (RPO) is the maximum acceptable time period prior to a failure or disaster during which changes to data may be lost as a consequence of recovery.

You'll notice that while both RTO and RPO are defined in terms of a time period, RPO is speaking to the amount of acceptable data loss. A Consistency Service Level Objective identifies the requirements of the recovery service. Many applications require data to be recovered, and therefore backed up, in a transaction-consistent state, in order to make use of the data. However, some applications, like databases, can accept data that was backed up in an I/O-consistent state, using undo and redo capabilities to transform the data into a transaction-consistent state.

Recovery at an alternate site implies availability of the resources necessary for resumption of key business operations at that site. This includes:

- The physical facilities to support the IT infrastructure and the work space for the personnel performing the recovery.

- Hardware, software, and services sufficient to support the key business operations to be resumed
- Current step by step instructions such that persons, skilled but not familiar with your specific environment, can perform recovery, and…
- Access to data that can be used to resume access of critical business applications.

Note that successful recovery requires continuous review of the disaster recovery plan to ensure that key business applications remain identified and protected and that documentation of manual and automated processes remain current.

Regular validation of the disaster recovery plan is another key component of maximizing the value of your data protection "insurance policy." It is impractical to regularly read all the data required for recovery at an alternate site to verify that it will indeed be available if and when a disaster occurs. However, it is important that steps are taken to manage the risk of data being inaccessible from the recovery site when needed. Simulating and testing the disaster recovery plan on a regular basis is necessary to validate the plan is practical, complete, and current. Also not to be overlooked is the requirement to validate that all business operations, not just those key operations running at the recovery site, can be returned to full production once it is determined that normal business operations can be resumed.

Retention and Preservation

Archive is often used as a verb to refer to a variety of services supporting business objectives. For clarity, we'll think of archive as a noun, consistent with the SNIA definition:

- An archive is a collection of data objects, perhaps with associated metadata, in a storage system whose primary purpose is the long-term preservation and retention of that data.

The key takeaway is that backup copies and archive repositories must be managed separately as they support very different business processes and requirements. However, this is not to say that both types of data can't be stored efficiently using a capacity-optimizing storage system.

Retention is the process of keeping and controlling digital objects for specific periods of time[4]. Retention is implemented as a policy defining a retention period. Retention policies ideally define the period of time[5] the digital objects are to be retained along with their administrative, legal, fiscal, historical, business, security, or other disposition requirements.

Preservation refers to the processes and operations involved in ensuring the ability to read, interpret, authenticate, secure and protect against the loss of data or information throughout its lifecycle. Preservation services:

- Read and interpret information in its original context over time and HW/SW obsolescence
- Protect information from loss or change[6]
- Verify and protect the authenticity, availability, and security for the entire information object, including its data and metadata

The change from the old way of thinking about 'archive' practices is that preservation services are required to deal with legal, security, and compliance risk as well as long-term retention requirements from creation to expiration.

Provenance is an archival term used to express the documentation of the history (meaning things like origin, source, and changes) and chain of custody of a digital record that is being ingested into a preservation repository, such as those found in libraries, museums, or historical archives. Authentic metadata and a chain of custody are the closest analog to provenance for electronically stored information (ESI) in the datacenter. Chain of Custody refers to a process that tracks the movement of evidence through its collection, safeguarding, and analysis lifecycle by documenting each person who handled the evidence, the date/time it was collected or transferred, and the purpose for the transfer.[7]

[4] Building a Terminology Bridge: Guidelines to Digital Information Retention and Preservation Practices in the Datacenter <http://www.snia.org/forums/dmf/knowledge/white_papers_and_reports/SNIA-DMF_Building-a-Terminology-Bridge_20090515.pdf>

[5] Digital Media Life Expectancy and Care <http://www.caps-project.org/cache/DigitalMediaLifeExpectancyAndCare.html>

[6] Media Preservation <http://en.wikipedia.org/wiki/Media_preservation>

[7] NIST SP 800-72: Guidelines on PDA Forensics, Appendix B. Glossary <http://csrc.nist.gov/publications/nistpubs/800-72/sp800-72.pdf>

Services that maintain integrity are focused on assuring the consistency, accuracy, and correctness of stored or transmitted data. The objectives of integrity processes are to be able to detect and prevent change or corruption. Integrity is often provided by using 'hashing' methods like digital fingerprinting or by locking the media from change using technologies such as write-once-read-many (WORM) immutable storage.

Governance, Risk Management, and Compliance (GRC)

The final "WHY" lens through which data protection should be explored is Governance, Risk Management, and Compliance. Data protection plays a key role in managing the risk of outages to IT-centric business operations and in supporting compliance activities. The GRC lens can help determine the right level of investment in data protection technology.

GRC is a concept that can be used to increase efficiency and reduce the complexity of your business. Michael Rassmussen[8] describes GRC as a business philosophy that ultimately is about the integrity of the organization. He describes governance as the culture, policies, processes, laws, and institutions that define the structure by which companies are directed and managed. SNIA defines risk, risk management, and compliance as follows[11]:

- Risk is the potential that a given threat will exploit vulnerabilities of an asset or group of assets to cause loss or damage to the assets.[9]
- Risk Management is the process of assessing and quantifying risk and establishing an acceptable level of risk for the organization.[9]
- Compliance is the state of being in accordance with a standard, specification, or clearly defined requirements. The "compliance market" is centered around storage and systems that support the retention and discovery of data as required by law or regulation.

Essentially, GRC aligns data protection with the business. Its broad reach is apparent in the scope of the compliance infrastructure taxonomy as defined by IDC[10].

How

The final row of boxes in Figure 1, the four "HOW" lenses of our taxonomy, describe the technologies used to implement data protection solutions. "Protection Technology" identifies how data protection and data security can be provided. "Storage Technology" identifies how to make data more resilient and how to store data more efficiently. "Access Technology" identifies how to deal with service and network outages and how to move data more efficiently. "Management Technology identifies how to administer, monitor, report, and make the offering more consumable.

[8] Corporate Integrity <http://www.corp-integrity.com/analysts/bio_michael_rasmussen.html>
[9] ISO/IEC TR 13335-1:1996 Information technology -- Guidelines for the management of IT Security -- Part I < http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=21733>
[10] IDC's Worldwide Compliance Infrastructure Taxonomy, 2009 (Doc #217572)

Note that this taxonomy has not focused on classifying only these technologies (the "how" lenses) because such an approach obscures business linkages and doesn't acknowledge that multiple sets of technologies can be used to satisfy service level objectives.

Protection Technology

The first technology lens, called Protection Technology, covers both the services used to create data copies and the technical controls that protect storage resources and data from unauthorized users and uses.

Data Copy Services

Data Copy Services include techniques used to create consistent point-in-time copies of data.

- Backup and Recovery
  - Files
  - Blocks/Images
  - Bare metal recovery
- Versioning
  - Create version when file changes
  - Snapshot/Checkpoint management
- Continuous Data Protection
  - Capture data changes
- Replication
  - Full copies and Changed Block Copies
  - Synchronous and Asynchronous modes
  - File synchronization
  - Log shipping

The Backup subcategory refers to copies that are created in response to scheduled events or ad hoc requests. Copies can be made at the logical file level or the physical block level. Copies of virtual machine images blur this distinction because each can be managed as a single file but visibility within the image can also be provided. When the disk on which the operating system itself fails, techniques referred to as Bare Metal Restore can be used to recover the data and operating environment.

The Versioning subcategory refers to creating copies in response to an event, such as when a change is made to a file. It also refers to managing multiple point-in-time copies of data sets throughout their lifecycle. Versioning is used to minimize recovery time by increasing the number of intermediate checkpoints from which an application can be restarted. File versioning products can be thought of as providing an "undo" function at a file level.

The Continuous Data Protection (CDP) subcategory refers to a class of mechanisms that continuously capture or track data modifications, enabling recovery to previous points in time. Technologies referred to as “near CDP”, which take frequent snapshots are included in the Versioning subcategory above. This category includes “true CDP” solutions which enable recovery of a data set as it existed at any previous point in time.

The Replication subcategory refers to a process used to continuously or periodically maintain a secondary copy of data. Data protection is just one of its many uses.

A point-in-time copy can also be classified according to whether it is a Full Copy (also known as clones or frozen images) or a Changed Block Copy (also known as delta copies or pointer copies) as well as by the method used to create the copy. Full copies are created by using Split Mirror (copy performed by synchronization; Point in Time (PiT) is split time) or Copy On First Access (PiT is replica initiation; copy performed in the background) techniques. Changed block copies are created using Copy on Write (original data is used to satisfy read requests for both the source copy and unmodified portions of the replica; updates to the replica are made to a save area) or Pointer Remapping (pointers to the source and replica are maintained; updated data is written to a new location and pointers for that data is remapped to it) techniques. Note that while the term Snapshot is used by many companies as a synonym for a changed block copy, snapshot may also be used to mean any type of point-in-time copy, including full copies; in fact, this is how SNIA defines[11] the term. SNIA uses the term “delta snapshot” to refer to changed block copies.

Replicas may be classified by the distance over which replication is performed (local or remote) as well as where the replication is performed (in the host, network, or storage array).

Synchronous replication is used at local distances where zero RPO is a requirement. Synchronous replication is a technique in which data is committed to storage at both the primary and secondary locations before the write is acknowledged to the host. Asynchronous replication is a technique in which data is committed to storage at only the primary location before the write is acknowledged to the host; data is then forwarded to the secondary location as network capabilities permit. Asynchronous replication may be used to replicate data to “remote” locations, where network latency would adversely impact the performance of synchronous replication.

Host-based replication includes logical volume mirroring, filesystem snapshots, file synchronization capabilities that support one-to-one, one-to-many, and many-to-one configurations, and log shipping. Log shipping is the process of automating the backup of a database and the transaction log files on a production server, and then restoring them onto a standby server. Network-based replication includes storage virtualization solutions that support cloning. Storage array-based replication solutions are the most popular.

Data Security Services

[11] SNIA Dictionary <http://www.snia.org/education/dictionary/>

Data Security Services are controls that protect storage resources and data from unauthorized users and uses. Various technologies may be used to implement the following services:

- Immutability
- Malware Protection
- Encryption and Key Management
- Information Rights Management
- Data Loss Prevention
- Media Destruction
- Identity Management

Data Security Services may also be classified as Integrity, Confidentiality, and Accountability services.

- Integrity means prevention of unauthorized modification of data. It includes detection and notification of unauthorized modification of data, recording of all changes to data, immutability (prevention of change after creation), and malware protection capabilities.

- Confidentiality[12] is the property that data cannot be accessed by unauthorized parties[11]. It includes encryption and key management, anonymity (prevent unauthorized disclosure of Personally Identifiable Information), Information Rights Management (protect sensitive information from unauthorized access), Data Loss Prevention (detect and prevent the unauthorized use and transmission of confidential information), and media destruction[13].

- Accountability services include authentication (proof of identity), authorization (proactively restrict access and usage), non-repudiation (indisputable proof that a particular entity performed a specific action), Identity Management, and audit logging.

Many data security technologies are components within other products while others form product categories unto themselves.

Those interested in understanding this subject in greater detail are encouraged to explore the resources offered by the SNIA Storage Security Industry Forum[14] including:

- An Introduction to Storage Security
- SSIF Solutions Guide for Data-At-Rest
- Storage Security Best Current Practices
- Storage Security Professional's Guide to Skills and Knowledge

[12] ISO-17799 <http://17799.standardsdirect.org/> defines confidentiality as "ensuring that information is accessible only to those authorized to have access"

[13] NIS 800-88 <http://csrc.nist.gov/publications/nistpubs/800-88/NISTSP800-88_rev1.pdf> defines media destruction to include shredding, disintegration, pulverization, melting, and incineration.

[14] SNIA Storage Security Industry Forum <http://www.snia.org/forums/ssif/>

Storage Technology

The Storage Technology lens focuses on "HOW" to make data storage more resilient and efficient.

- Resiliency
  - CRC, ECC/FEC, block identity
  - RAID (mirroring, parity)
  - Erasure coding
  - Disk scrubbing/cleansing
  - HA technologies (auto-restart, clustering)

- Capacity Optimization
  - Delta snapshots
  - Thin provisioning[15]
  - Compression (encoding)
  - Deduplication (remove redundant data)
  - Packing (many small objects in larger allocation unit)
  - Recompression (specialized encoding)
  - Regeneration

- Power Management

Resiliency

Resiliency technologies are used to provide self-healing capabilities and ultimately instill confidence that the storage system is reliable enough for operating a business.

A cyclic redundancy check (CRC) is designed to detect changes to data. A disk drive calculates and stores a short, fixed-length binary sequence with each block of data. When the block is read, the disk drive repeats the calculation. The block contains a data error if the new check code does not match the stored check code.

An error-correcting code (ECC), aka a forward error correction (FEC) code is redundant data that is added to the data on the sender side. If the number of errors is within the capability of the code being used, the receiver can use the extra information to discover the locations of the errors and correct them.

Block checksums[16] can detect bit corruption and partial writes but does not detect corruption caused by lost or misdirected writes. Storing the block number within the checksum allows this type of

[15] A technology that allocates the physical capacity of a volume or file system as applications write data, rather than preallocating all the physical capacity at the time of provisioning.

[16] An Analysis of Data Corruption in the Storage Stack, Garth Goodson, SNIA Storage Developer Conference 2008

corruption to be detected by the filesystem at read time (and is less costly than using a write-readback-verify technique).

Redundant Array of Inexpensive Disks (RAID) is a class of data storage schemes that spread data across multiple disks, known as a RAID set, and mirror or add parity information to increase data reliability and/or increase I/O performance. Failure of one disk (or possibly more, depending on the RAID scheme) in the array will not result in loss of data. A failed disk may be replaced with a good one to allow the lost data to be reconstructed from the remaining data and the parity data.

The disks in the RAID set are virtualized as a single logical disk. If a disk in the RAID set fails the data on it can be reconstructed from the remaining data and parity data. RAID 6 (striped disks with dual parity) combines four or more disks in a way that protects data against the loss of any two disks. While data remains accessible during rebuild, performance can be degraded. Also note that the time to rebuild a RAID set increases as disk capacity increases, thereby increasing the risk of data loss. For this reason, other technologies, such as erasure coding, may also be used to provide redundancy.

Erasure coding can be used to provide increased protection from disk failures as well as speed rebuild time. Whereas RAID protection applies to all data in the RAID set, erasure coding can be selectively applied at a file system level. It also allows different levels of redundancy to be "dialed in," so the customer can tradeoff space for the number of failures they would like to survive. Erasure coding supports faster rebuild time through finer granularity; the rebuild process can also leverage multiple processors such as by using a grid architecture.

Data scrubbing/cleansing is the act of detecting and correcting or removing corrupt or inaccurate data. In contrast, data validation is performed at entry time, rather than on batches of data.

High Availability technologies can also be used to protect storage servers in a similar fashion as for application servers. This includes automatic restart and automatic failover in a fault tolerant configuration.

Capacity Optimization

Capacity Optimization and Power Management technologies are used to store data in a space and power efficient manner.

Capacity optimization refers to any method which reduces the consumption of space required to store a data set, including RAID, delta snapshots, thin provisioning, compression, and data deduplication. We've already discussed RAID for resiliency and delta snapshots in the context of Protection Technology.

- RAID 5/6 may be considered capacity optimization techniques because they can be used to regenerate data without storing full copies. RAID 5[11] is a form of parity RAID in which the disks operate independently, the data stripe size is no smaller than the exported block size, and parity check data is distributed across the RAID array's disks. RAID 6[11] is any form of RAID that can continue to execute read and write requests to all of a RAID array's virtual disks in the presence of any two concurrent disk failures. Several methods, including dual check data computations (parity and Reed Solomon), orthogonal dual parity check data and diagonal parity have been used to implement RAID Level 6.

- Delta snapshots[11] are a type of point in time copy that preserves the state of data at an instant in time, by storing only those blocks that are different from an already existing full copy of the data.

- Thin provisioning is a technology that allocates physical capacity as applications write data, rather than pre-allocating all the physical capacity at the time of provisioning.

- Compression is the process of encoding data to reduce its size.

- Data deduplication[17] replaces multiple copies of data—at variable levels of granularity—with references to a shared copy in order to save storage space and/or bandwidth.

- Other techniques such as de-constructing compound file formats, re-generating scaled images, and packing many small files into a space allocation unit may also be used to save space.

Note that capacity optimization technologies can complement each other in two senses. First, they may be used to optimize different infrastructure elements based on their applicability. Second, some of the techniques can be used together to achieve additive benefits.

Power Management

Energy efficiency ratings, like those used in the ENERGY STAR program, have typically been focused on "reading the meter" calculations. Yet, capacity optimization techniques can eliminate the need for additional hardware, yielding the largest energy savings. Significant energy savings can also be achieved by slowing down or turning off components that are not in use. SATA[18] arrays fronted by large caches (SSDs, flash, FC/SAS drives) can save power (about 6x GB/W) at a small cost in performance. Storage systems, e.g., Massive Array of Idle Disks (MAID), can also power down disk drives individually or in groups when not required. Such storage systems reduce the power consumed by the storage array at the cost of increased Mean Time To Data. Note that the ability to spin-down drives can be affected by data placement strategies. For example, if data is spread over all the drives in a storage system, none of the drives can be powered down when that file is accessed.

[17] Understanding Data Deduplication Ratios <http://www.snia.org/forums/dmf/knowledge/white_papers_and_reports/Understanding_Data_Deduplication_Ratios-20080718.pdf>

[18] Serial ATA <http://www.serialata.org/>

Access Technology

The Access Technology lens categorizes "HOW" data is located, accessed and optimized. As with all of the taxonomy lenses, the subcategories are not meant to be exhaustive; merely suggestive of the types of elements that should be reviewed.

- Service Availability
  - Constrained by product limitations
  - Constrained by business policy
  - Outage containment and resolution

- Network Connectivity
  - Impact of intermittent connectivity
  - Impact of distance on latency, throughput (IOPS), and data transfer rates (MB/sec)

- Performance Optimization
  - Network optimization
  - Hardware acceleration
  - Intelligent automation (e.g., caching, tiering, multipathing)

- Index-Search

Service Availability considers how products and company policies can have an impact on service availability. For example, a server may have a limited number of client-server connections or may prevent client connections when certain server activities are running. In such situations, service availability may be improved by product changes. In other cases, a business policy rather than a technical limitation may constrain service availability. For example, if business policy dictates that non-standard ports through a firewall may not be opened, a service that requires opening those ports to communicate with its clients will be blocked. Solutions to such problems include changing the ports which the product requires for operation or changing business policy if within the risk tolerance of the organization. Improving processes used to contain and resolve service outages is another key action businesses can take to improve service availability.

Network Connectivity impacts data access in at least two ways. When connectivity is required to perform a service, a product can be designed to continue service in the face of intermittent or no connectivity or suffer outages and discard any useful work performed before the interruption. For example, a network backup application that loses connectivity between the client and server could simply report an interruption and restart the data transfer when connectivity is restored. Alternatively, it could cache backups locally and send data in smaller chunks so progress is maintained rather than forever backing up but never being backed up.

The second method in which network connectivity impacts data access is related to distance. The farther two points are from each other, the longer it takes to send data between them. Latency-sensitive applications, such as synchronous replication, cannot tolerate long latency. I/O operations per second throughput and MB per second data transfer rates may also be degraded. However, sending data in a capacity-optimized format or using performance-optimization techniques can significantly accelerate effective performance.

Performance Optimization categorizes a variety of techniques used to improve both the speed at which data can be accessed and data availability. Speed issues are typically the result of bandwidth limitations and protocol inefficiencies. Bandwidth limitation issues are addressed by using QoS (prioritization), compression, and de-duplication technologies. Transport protocol chattiness issues are addressed by reducing the number of round trips by caching/pre-positioning data, by sending "virtual data", and by using larger TCP buffers. Application protocol chattiness issues are addressed by eliminating chattiness over the WAN either by pre-fetching data or by using application-specific modules to complete transactions locally. Chattiness mitigation modules are available for file serving (CIFS, NFS), email (MAPI), and branch office application (SQL) protocols. Remote access software can also be used to eliminate the need for servers at remote offices by running basic applications over relatively low-bandwidth connections or by streaming software. Note that WAN optimization controllers help improve remote access response time and may mitigate application outages caused by loss of network connectivity.

Hardware acceleration is another approach to improving the speed of data access. Three obvious examples include the use of flash drives to improve disk access, offloading processing from constrained processors to separate cards, such as TCP/IP Offload Engines, and using robotics to automate media handling. Intelligent use of high performance hardware is key to maximizing the benefits of such technology.

Automatically moving data between storage tiers based on policies that reflect costs, customer service level objectives, and real time measurements can have a large impact on how fast an application can access its data.

Multipathing is an example of a technology that enhances both SAN performance balancing I/O load and application availability through path failover.

Index-Search refers to improving the ability to locate data so that it may be accessed. There are many types of search, including full text search, metadata search, concept search, fuzzy search, proximity search, and the use of regular expressions. The point here is not to discuss the tradeoffs of each search type but rather to provide a subcategory where improving data access by locating it faster (or at all) is covered.

Management Technology

Management Technology is the final "HOW" lens in our data protection taxonomy. It includes central administration, instrumentation (real time monitoring and alerting, historical reporting, trend analysis, and capacity planning), and consumability.

The central administration subcategory includes automation technologies which allow administrators to manage rapidly growing quantities and types of data. Policy-based and model-based automation allow Service Management best practices, as described by the IT Infrastructure Library (ITIL[19]), to be implemented at scale. The use of management standards, like SMI-S[20], also promotes cost-efficiency, robustness, and more rapid time to market, as vendors can leverage the work of the entire industry, rather than reinventing the wheel. A key requirement going forward is also how seamlessly the entire IT infrastructure can be managed, regardless of the mix of platforms, virtualization, and cloud services in use.

The real time monitoring and alerting subcategory includes capabilities that enable feedback on system operations. Such feedback can be leveraged for system tuning or to manually correct out of compliance situations and speed problem resolution.

The historical reporting and trend analysis subcategory includes capabilities suitable for creating executive dashboards and preparing capacity forecasts.

The consumability subcategory is intended to capture items that have an impact on how a service is consumed by the customer. A deep appreciation of customer requirements is necessary to develop effective products with a minimal learning curve and operational costs. Any capabilities that offer improvements in these areas can be covered by this subcategory.

Licensing is intended to ensure the producer of an offering is compensated by the customer for services provided as specified within an agreement. However, many current schemes are perceived to be unfair by the customer, place unproductive and ongoing burdens on both the producer and the consumer, are incompatible across products, and can increase the costs and risks of operating a business. Capabilities which ease these concerns promote greater service consumption.

Billing is also a key component of providing on-demand services; without it, a pay as you go model for cloud services is problematic. On the other hand, strong billing capabilities raises the barriers to entry for other service providers.

In summary, simplicity, licensing, and billing capabilities are covered in the consumability subcategory.

[19] Information Technology Infrastructure Library <http://www.itil-officialsite.com>

[20] SNIA Storage Management Initiative <http://www.snia.org/forums/smi/>

Summary

As stated at the beginning, the SNIA Data Protection and Capacity Optimization Committee[21] would like to encourage the use of this taxonomy as a reference for use throughout the industry. Rather than inventing another taxonomy du jour each time we want to examine the data protection space, let's agree to treat this as a living taxonomy and evolve this reference to meet our common goals. This paper is designed to carry forward a consistent message as this taxonomy is socialized further. Feel free to share it to help set a baseline for mutual understanding on what data protection is and its terminology. The SNIA Tutorials[22] provide additional information about the topics mentioned in this white paper.

We have no doubt that innovative customers and companies will find ways to use this taxonomy to achieve real world business benefits. You can indeed transform the vassalage of backup and recovery into a reservoir of competitive advantage. We encourage you to share any comments or refinements with the general community by joining the DPCO Committee or sending an email to dpco@snia.org.

About the Author

Mike Dutch has been active within the storage industry for over 30 years in management, intellectual property, product management, software engineering, and field support roles. Mike has worked internationally for leading vendors with responsibilities spanning mainframe and open systems environments, software and hardware technologies, and business partners. Mike is also an honoree of SNIA's Profiles in Achievement "Hall of Fame"[23].

About the DPCO Committee

The mission of the Data Protection and Capacity Optimization (DPCO) Committee is to foster the growth and success of the market for Data Protection and Capacity Optimization technologies. It does this by educating the vendor and user communities, performing market outreach that highlights the benefits of relevant technologies and documenting implementation considerations and best practices, and advocating and supporting associated technical work. Join the SNIA DPCO Committee to gain insight into and contribute to market-leading technologies.

About the SNIA

The Storage Networking Industry Association (SNIA) is a not-for-profit global organization, made up of some 400 member companies and 7,000 individuals spanning virtually the entire storage industry. SNIA's mission is to lead the storage industry worldwide in developing and promoting standards, technologies, and educational services to empower organizations in the management of information. To this end, the SNIA is uniquely committed to delivering standards, education, and services that will propel open storage networking solutions into the broader market. For additional information, visit the SNIA web site at www.snia.org.

[21] <http://www.snia.org/dpco>

[22] <http://www.snia.org/education/tutorials/>

[23] <http://www.snia.org/about/profiles/>