Angels in our Midst: Associative Metadata in Cloud Storage

Tom Coughlin and Mike Alvarado

Introduction

Driving new business models with metadata

As more valuable personal and commercial content is stored in private and public clouds how do we ensure this content is protected? How do we find and connect with what is valuable in virtualized environments? We need some "Guardian Angels" to protect our interests. Guardian angels, as used here, are instantiations of our interests in shared environments. These guardian angels collect metadata not just from content characteristics but also from individually accumulated associations with this content. Personalized guardian angels can, if allowed, represent our anonymous interests in a shared environment through an "Invisible College."

Metadata is information that allows organizing, indexing, finding and using content. Metadata can exist at many levels ranging from simple descriptions of a thing or event, to observations of feelings about the thing or event and introduced by an observer, to statements or expressions of meaning or intention. Types of metadata and their independent or contingent relationships evolve with time. Our developing technologies enable us to create associated data that can help define richer details about the world and events around us. Increases in the associative capacity of information combined with ever more sophisticated tools to analyze and process it enables visualizing the relationship of that thing or event with other things and events.

Metadata created by individual associations can be the basis of important new industries and business models. Individuals can pool their collective knowledge to make better buying decisions and organizations can use anonymous associative metadata to create better allocation of valuable resources as well as to offer individuals and groups faster access to valuable content and thus generate better income streams through content monetization.

Organization of paper

This white paper will discuss business and other drivers that are changing and expanding the role and function of metadata, including its structure and meaning. It will also discuss a concept for finding and tracking the ***emergent*** meaning of content over time from a personal level using software that quietly observes an individual user and watches the developing relationships between various content (both local and remote).

These observations then become metadata that allows users to find and further benefit from these developing relationships. We refer to this software as a "Guardian Angel."

Even more powerful data relationships may be found from comparing associative content found by these guardian angels with each other through a shared interaction, which facilitates aggregating distributed data, as well as the comparison and development of associative knowledge while preserving and protecting content privacy and security. We refer to this relationship between personal associative databases as an "Invisible College." The Invisible College enables the creation of even higher, richer and more complex levels of association and from more points of view than an individual could create alone. It is also a potentially useful tool for managing the evolution of and creating a purpose-driven system of data systems.

This white paper will develop these concepts through a detailed discussion of the types of metadata, followed by an analysis of metadata creation for active content. An analysis of the evolving role of social relationships and the development of the Meaning Levels of metadata will show how social learning and associations can help create and capture contextual data. The paper also examines requirements for standards to be used for associative metadata and will close with an example based on smart girds and meters before presenting final conclusions.

Types of Metadata

We will discuss a model for various types of metadata as a basis for examining how various metadata levels can be applied to real problems in order to create new business opportunities and strategies. Metadata is data about some information that may be kept in a storage system or moved within a business management system or across a network between multiple electronics systems. Metadata is used to align data with an intended purpose and to achieve useful outcomes. Uses of metadata can range from a successful search and access of content, to adjustment of system resource allocations to synchronize to usage patterns, to content protection and privacy, to reporting and accounting of individualized or aggregated and autonomous information.

Data and content generation are exploding. With more bits available, people and organizations are finding more ways to productively use them. In addition, the speed of interfaces and networks is similarly expanding. As the capacity, efficiency, availability and connection speed of storage devices and data and application systems has grown, there has also been created positive feedback loops that in turn lead to generating even more content. As the amount of raw content expands without bounds there is a great need for descriptions of the content that can be used to find and use this content. Thus there is an evolving need for new and novel metadata that enables individual users and organizations to optimize the value of their digital infrastructure and assets.

New mobile electronic devices provide a host of options for personal use and highly mobile business use cases. Mobile users need to have access to critical information wherever they are. Multi-faceted business relationships also require that information be

sharable and transportable. Without a ubiquitous infrastructure for hosting and delivering data, business is severely curtailed. Properly used, metadata allows the broadest range of locations, device type and policies. This enables people to enjoy access to information they need, whenever they need it, wherever they need it and with the right characteristics to meet user goals.

The characteristics of required metadata include basic metadata as well as non-basic metadata, which are used to create meaning:

Meaning Metadata:

1. Relevance to other information “units”
2. Purpose
3. Significance/importance
4. Other commerce related attributes such as cost, discounts, etc.
5. History of personal associations showing developing connections between content
6. Content similarity in a broader sense
7. “Train of thought” traces

Basic Metadata:

1. Scope (unit or size)
2. Source
3. Type (legal, accounting, development, etc.)
4. Application or associated applications
5. Access path
6. Retention time
7. Locations (multiple)
8. Access control bits
9. Access time/performance specific to access path
10. History of all changes and accesses (time stamps)

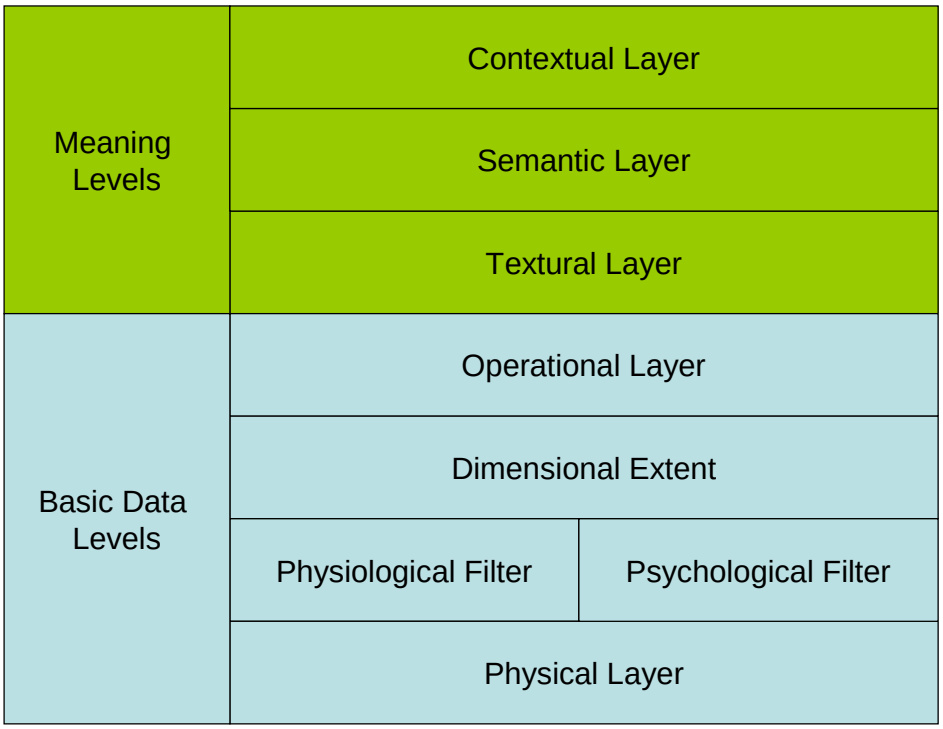
Figure 1 shows a graphical representation of a proposed metadata model[1].

Layer 1, Physical Layer (*Sensory and source information)*: This is where physical infrastructures are managed, maintained and protected. This encompasses basic information about content sources (where and when) as well as sensory information of various sorts. Sensory information could include sound, sight, touch or smell in some defined fashion.

Layer 2, Physiological and Psychological Filtering: This metadata defines what sort of personal or experiential filtering is applied to the signal. This filtering may relate to the characteristics of the channel used to transmit (physiological) or experience (psychological) the metadata, which may differ depending upon data type—e.g. speech or music may undergo different psychological filtering. Thus speech can be converted

to text, which can be a filtered metadata giving useful and searchable information about what was said but speech to text conversion would not pass on the psychological import of a piece of music.

Figure 1. Metadata Layer Model[1]



:

Layer 3, Dimensional Extent: This layer provides the ability to enable composing services from resources aided by service level aggregators. This layer reflects content complexity, which is described as a set of orthogonal dimensions describing content. For an image or video this layer may indicate whether it is flat or does it have depth as well. Likewise for audio content this could be used to describe the number of "voices" or the level of presence of the content, e.g., surround sound has more audio dimensions than a monaural sound. This concept of dimension could be applied to all of the senses with an interesting expansion of our ways of understanding dimensions in touch and smell,

Layer 4, Operational Layer: The fourth layer is focused on providing information agility for the purpose of optimizing responses to information requests as they arise. To ensure efficient operation, standard operating procedure responses would be formalized and infrastructure aligned to scenarios and instances. This level of metadata gives instructions about how to recreate content in its intended form using defined hardware and software. For instance this level could include information on what operations are performed on dimensional extents such as the number of frames per second, sampling rate, bit-depth, etc for video content. This level also contains data access and control

information, redundancy and location information, size, retention time and all other higher-level basic data metadata.

Layer 5, Textural Layer: The textural layer spans the entire data, user and infrastructure space. It provides an intermediation function to support composing the experience(s) that data user's wish to achieve. The first primary function is to provide an exchange mechanism between data sources and data requestors. The second primary function is to enable component integration to meet data user's needs. This level can be seen as a subclass of the next level (Semantic). It involves metadata that describes differences involving constructions built from the lower levels. Textual information could describe the differences between otherwise similar things with, for instance, different colors. A subset of this layer is data about use and interaction of content. This layer would include access frequency and change information and sequential relationships about data access.

Layer 6, Semantic Layer: This is a concrete definition of the object or experiences in a piece of content based upon generally agreed upon constructions – for example "a tree" - "my friend said..." (as input from an audio byte where it is recognized that it is your friend speaking and he/she said…) This layer would also include observations on participants and their interactions and evaluation of content or events (this would be structured as opt-in choices). Applets and widgets are examples of logical entities users can avail themselves to form representative proxies of users' worldviews.

Layer 7, Contextual Layer: This level refers to the description of experiences vis-à-vis content by a sensible sentient being. What users sense are not the devices, APIs, protocols, etc that compose information technology systems. Rather they encounter their personal libraries of experiences. Current computers cannot create true judgmental information for metadata as they cannot look at a scene and define it as "beautiful" or listen to music and define it as "melodical", analyze a smell and refer to it as "pungent". The contextual level is by its nature subjective or personal—specific to the participant. A collection of contextual level metadata from several sensible sentient beings could be represented as providing a sort of temporary consensus on the "meaning" of that content. This type of metadata determines the importance and personal evaluation of the context or event from the lower metadata layers.

Together the metadata layers in the stack allow creation of a body of metadata that includes the basic information as well as more advanced personal information such as feelings, valuation and associations.

Cloud Storage and Metadata

Cloud storage refers to digital storage assets accessed over TCP/IP networks. Clouds can be private (within a facility or organization) or public (accessible by individuals over the internet). With the introduction of associative personal metadata clouds have a unique capacity to enable new and profitable business models.

Cloud storage creates a need to protecting content within the cloud. This drives the need for a Virtual Information Perimeter. Clouds challenge assumptions about boundaries and perimeters, which in turn challenge conventional infrastructure views about processing and through virtualization of physical devices, the physical and logical components of the access and processing stack.

Because of the needs of multiple people interacting with personal or commercial content there are new ways that they will interact with this content. These new uses and interactions drive the need for new types of metadata to ensure data integrity, data security and privacy and new relationships with the content by individuals and social groups.

In short, metadata must continually evolve to meet innumerable new uses for content in a more public and sharing-centric environment. For instance, metadata in a cloud environment must be aware of simultaneously monitoring critical or private files for the purpose of determining their locations and access path in order to ensure data integrity or for evaluating the risk of discovery and access. Access to information through public networks provides new opportunities but it also poses new dangers. These opportunities and dangers are magnified by ubiquitous corporate and public networks and the vast number of handheld and removable storage devices accessing shared data.

Metadata will evolve to deal with future cloud environments and new ways for social integration and interaction with content accessed through the cloud. Metadata will evolve to meet the needs of the many new paradigms for hosting, delivering and using data. Metadata will be indispensible to dealing successfully with the risks and opportunities related to routinely using personal devices in business settings with the consequent potential for unauthorized removal or compromise of intellectual property.

Metadata must help us account for and simultaneously manage private and regulated information housed and protected in the same corporate data stores that may simultaneously contain obsolete, dangerous or illegal files. Commingling personal and corporate datasets may be essential to enable organizations to reach their objectives but they could also cause irreparable damage from shareholder lawsuits, government intervention, loss of market share, and reduced revenue--all stemming from a mishandling of comingled of private and corporate data.

Whatever data system solutions arise, they will have identifiable characteristics such as automated or semi-automated information classification and inventory algorithms with significance and retention bits; automated information access path management; information tracking and simulated testing access (repeated during the effective life time of information for quality assurance); and automated information metadata reporting.

These characteristics reflects the view that 20% of all data is generated by enterprises but enterprise data is 80% of the data for which someone is held ‘liable’ for by

regulations such as Sarbanes-Oxley (SOX). Liability for contractual data held by individuals will likely grow as the use of public cloud storage increases. Metadata about this content will be a crucial element in the management and protection of both corporate and personal content.

In the next section, a conceptual model will be presented to guide the evolution of metadata for active objects.

Guardian Angels and the Invisible College

Active data in a shared environment (private or public cloud) is data accessed by a number of people. The interaction of individuals with data can itself be a generator of metadata from simple interest (hits) information to more complex evaluation of the value and "meaning" of the data.

We referred to the idea of a "guardian angel" that has access to network traffic as well as content in personal and shared storage. This angel is capable of finding and tracking the emergent meaning of content over time both from a personal level using software that quietly observes an individual user and watches the developing relationships between various content (both local and remote), to create metadata that allows the user to find and use these developing relationships. All individual users can have their own "guardian angels" to create their own person metadata universe with both basic and higher level metadata. To protect privacy, individual users would control how much or what sort of metadata he/she wishes to share.

There can be a great potential and mutual value that can come from sharing some of the metadata that is available from the "guardian angels" that individuals would have. By comparing the valuation and meanings derived by each individual and comparing them to an aggregate, an emergent sense of "meaning" can be determined. To protect individuals, aggregated information would anonymous. The aggregate metadata generated by the polling of the individual "guardian angels" would be kept within an invisible college. The invisible college could be accessed by any of the members who contribute their angel metadata to create a higher level sense of meaning while protecting their actual identity through anonymity.

The invisible college could also be a great storage and content management tool. It would include information that may make it possible to determine what content and type of content will be popular and have this arranged in edge delivery or edge publishing and propagation systems for rapid access, protection and dissemination. It could also allow better management of hierarchical storage systems determining which sort of content is popular and having that available on higher performance storage.

Angels in the Standards

Several standards efforts have been started that incorporate some aspect of the Angel and invisible college concepts, though not with the same goal, intent or understand as is contained in this paper[1,2]. The evolution of value for shared cloud assets will be driven by even greater reliance on metadata. The Intel work on the Fodor application for combining localized personal metadata with GPS maps[3] represents a single company's attempt to bring much greater capacity for amalgamating and interpolating individual metadata sets into an "uber" metadata set (like an Invisible College).

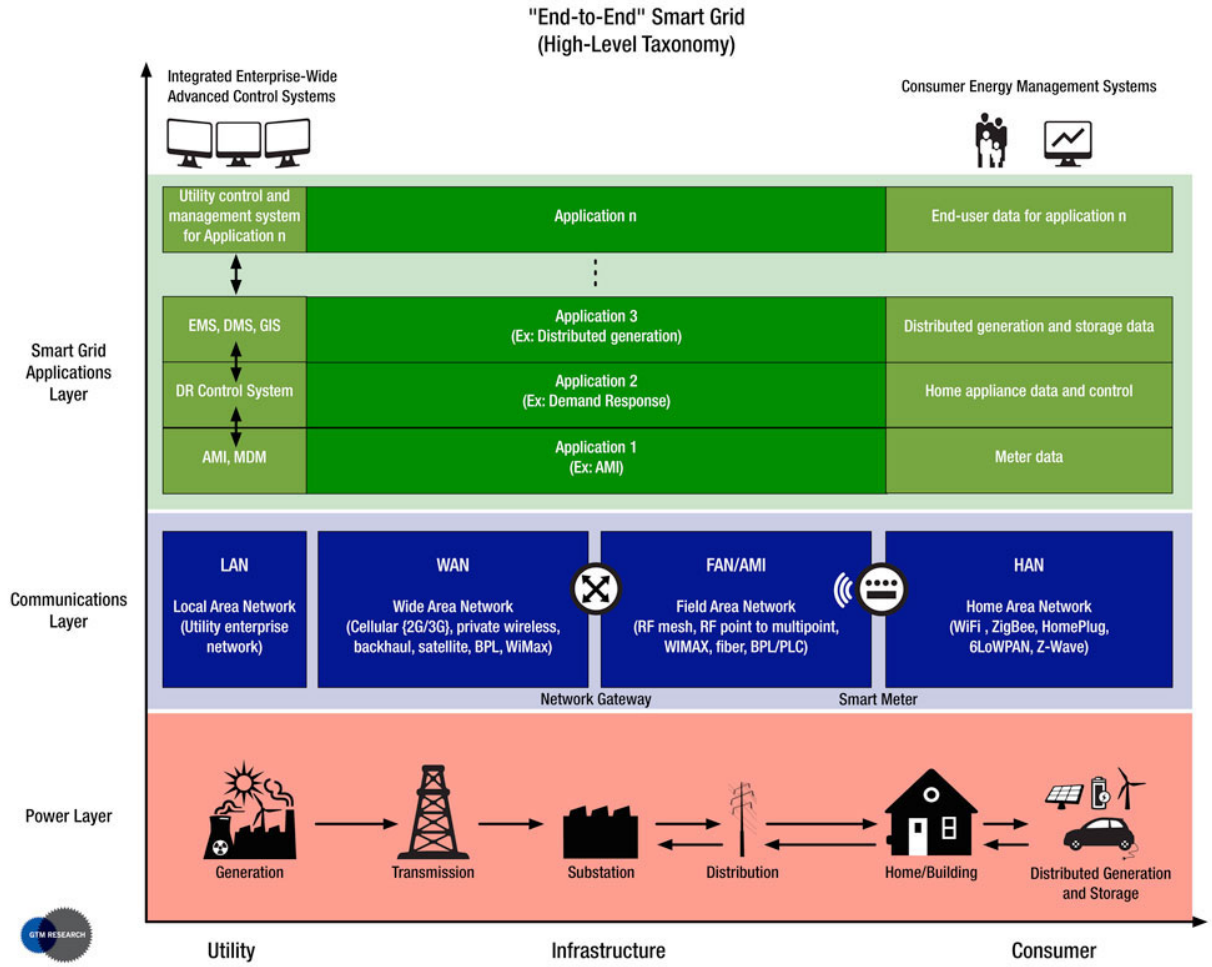
We need an open set of standards for associative individualized metadata (held by a guardian angel) combined with an anonymous aggregation of the individualized guardian angel metadata (the invisible college). An open ended metadata standards that allows organic growth and use of associative metadata will be a key element in widespread use of this concept.

One standard that holds promise for being influenced by the guardian angels and invisible college concepts is the Storage Networking Industry Association's Cloud Data Management Interface (CDMI)[4]. It is a functional interface that applications could use to create, retrieve, update and delete data elements from the cloud. By incorporating guardian angels and invisible colleges, a far richer set of contextual information could be leveraged along with cloud-hosted data objectives. These elements could also serve as storage operational managers too; by delivering and coordinating management actions with rich context, it could greatly enhance the value of cloud-based services and cloud-based information assets. To investigate this potential, a mapping is needed between the metadata layers and elements described in this paper and CDMI described objects.

A second standards effort could involve adding context to search. Search is a standardized method for locating data yet traditional search does not have any ability to retain or take advantage of context to further optimize search actions, objectives or constraints. Context rich search could greatly enhance the value of found objectives and dramatically reduce the incidents of extraneous data.

An Illustrative Example: Smart Grids and Smart Meters

Figure 2. Example of an End-to-End Smart Grid[5]



Energy consumption is a key macroeconomic issue. Smart meters and grids are tools for improving the value and economics associated with that use. They provide a vehicle for providing consumers feedback on usage and for utilities to manage and predict power generation and distribution requirements. With today's level of smart meter data, consumers have shown they can achieve a 10% energy consumption reduction without lifestyle change. A 10% reduction in use is equivalent to the total energy provided by US wind and solar, 113.9 billion kW•h/year or a new geothermal or nuclear power plant.

Consumers can play an even more active role in saving power by using the information assets generated by a combination of guardian angels and an invisible college. Motivating our thinking are well known issues about privacy and security in a smart grid. Smart girds and meters represent the ultimate cloud-oriented application. Consumers plug into an invisible infrastructure and get the resources they are seeking. A consumer's guardian angel collects associative information about personal uses of

power that can be anonymously shared with an invisible college associated with a smart grid. Much more detailed data allowing the determination of energy needs at any time during the day can be created using the collective and anonymous data in the invisible college.

Utilities face a daunting task of building the clouds supporting vital power infrastructure that effectively meet the need for efficient provisioning while effectively managing all the information that could flow onto these networks. Besides allowing for protection of privacy guardian angels through an invisible college it could enable monetization initiatives of this infrastructure for other purposes . For instance utilities (through unregulated subsidiaries) might be able **to supplant** major content and information utility suppliers such as Google and Facebook as the future's most reliable mechanism for all other applications whether personal or private.

Concerns about privacy breaches on Facebook have caused commentators to opine that Facebook and other **public** social media sites could lose tens or hundreds of millions of customers due to issues related to information ownership, privacy and security[6].. The guardian angels as described in this paper could be the essential guarantor of value in social media and the clouds that house these systems as well as a means for protecting personal privacy.

Smart homes that utilize guardian angel metadata to control resources and home automation could also be a valuable use of this technology. Guardian angels could have access to local environmental metadata generated by a smart home which could become part of the information pool for generating associative metadata. Home automation and monitoring will benefit enormously through the ability to interact and share information with the guardian angels for the building residents.

Home security vendors offer home automation services bundled with their monitored security solutions. Guardian angel integration will increase the value and accuracy of home security systems. Home health monitoring where the gathered data is incorporated into a person's guardian angel is a natural extension of the concept. Home health monitoring is popular in Japan and Korea where there is 90% penetration of broadband. Health maintenance organizations worldwide may be the real drivers for the adoption of guardian angels and invisible colleges which could be integrated with fully automated and digitized record management systems.

Conclusion and Ideas for Action

Enhancing and extending metadata to optimize the value proposition of clouds is a critical and pacing issue for industry. Guardian angels create a path for trust in a shared content environment and could be a valuable enabler to future cloud business models. The concept of a guardian angel composing an Invisible College challenges conventional notions of who is driving design and implementation of clouds. Industries need to develop an ethic of co-creating **with** their customers in a collaborative and

privacy-protecting manner to enable the efficient operation of vital infrastructure systems such as smart grids and meters, enhanced personal content systems and new solutions for healthcare and education. Regardless of which applications galvanize adoption of guardian angels and invisible colleges, the old self-contained stack of the past will eventually give way to a stack that spans users and infrastructure domains to create a unified data fabric well informed by context driven rich and continuously evolving metadata.

References

1. ***A Novel Taxonomy for Consumer Metadata***, T. M. Coughlin and S. L. Linfoot, Presented at the 2010 ICCE Conference in January 2010
2. Ontology for media resources, http://www.w3.org/TR/2010/WD-mediaont-10-20100309/
3. Intel Fodor Travel project: http://www.engadget.com/photos/intels-context-aware-presentation-and-fodor-travel-app-at-idf-2010/#3366248
4. SNIA Cloud Data Management Initiative, http://www.snia.org/forums/csi/programs/CDMIportal
5. Source: Green Tech Media
6. http://www.greentechmedia.com/images/wysiwyg/News/endtoendtax.jpg"Why Facebook Users Are Considering Leaving" http://mashable.com/2010/05/25/facebook-quit-survey-results/

About the Authors:

Tom Coughlin, President, Coughlin Associates is a widely respected storage analyst and consultant. He has over 30 years in the data storage industry with multiple engineering and management positions at high profile companies.

Dr. Coughlin has many publications and six patents to his credit. Tom is also the author of Digital Storage in Consumer Electronics: The Essential Guide, which was published by Newnes Press. Coughlin Associates provides market and technology analysis (including reports on several digital storage technologies and applications and a newsletter) as well as Data Storage Technical Consulting services. Tom is publishes a *Digital Storage in Consumer Electronics Report, a Media and Entertainment Storage Report*, *and a Capital Equipment and Technology Report for the Hard Disk Drive Industry*.

Tom is active with SMPTE, SNIA, IDEMA, the IEEE Magnetics Society, IEEE CE Society, and other professional organizations. Tom is the founder and organizer of the Annual Storage Visions Conference (www.storagevisions.com), a partner to the

International Consumer Electronics Show, as well as the Creative Storage Conference (www.creativestorage.org). He is also a Senior member of the IEEE, Leader in the Gerson Lehrman Group Councils of Advisors and a member of the Consultants Network of Silicon Valley (CNSV). For more information on Tom Coughlin and his publications. go to www.tomcoughlin.com.

Mike Alvarado has over 25 years IT and data storage industry experience including over 5 years of executive management responsibility. Mike's expertise is in gaining new product adoption, managing program and market risk and creating market influence strategies. Past technology experience includes storage systems, software, deduplication, storage protocols, content management, virtualization, replication and data protection. Mike's industry experience include past membership on the Board of Directors of the Storage Networking Industry Association (SNIA). He also served as a Technical Advisor to the Best Practices Committee of the Association of Storage Networking Professionals, which was a forum for sharing knowledge and experience to create greater awareness by end users of best practices in storage networking applications such as disaster recovery, data protection, data migration, and storage consolidation. Mike's most recently published paper addresses how to increase the pace and quality of technical innovation. He has an MBA degree from Santa Clara University and a B.Sc. from San Jose State University. He is presently completing a M.A. in Applied Anthropology from San Jose State University. To contact Mike Alvarado, please send an e-mail to mike@goingevergreen.org.