



Education

Classification: the cornerstone for Compliance- and Cost-driven Information Management

Edgar StPierre, EMC

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced without modification
 - ◆ The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.

Classification: the cornerstone for Compliance- and Cost-driven Information Management

Without a clear understanding of all the information under management in your environment, it is impossible to get a handle on information growth, compliance-related risk mitigation and information management costs. The practice of information classification is fundamental to an effective information-centric ILM strategy. Information classification requires that I.T. administrators work with Line-of-Business and knowledge workers to gain an understanding of the data to be managed. Once a clear set of goals and policies are established you can efficiently organize your information into tiers of service that will meet the performance, protection, and compliance requirements of your business.

This session will explore the different types of classification methodologies and techniques used to drive data placement and service delivery today, and how that relates to the delivery of a multi-tiered service catalog.

About the SNIA DMF

This tutorial has been developed, reviewed and approved by members of the Data Management Forum (DMF)

- The DMF is an industry resource to those responsible for the accessibility and integrity of their organization's information
- The DMF focuses on the technologies and trends related to Data Protection, ILM and Long-term digital information retention

DMF Workgroups:		
Data Protection Initiative (DPI)	Information Lifecycle Management Initiative (ILMI)	Long-term Archive and Compliance Storage Initiative (LTACSI)
Defining best practices for data protection and recovery technologies such as Backup, CDP, Data deduplication and VTL	Developing, educating and promoting ILM practices, implementation methods, and benefits	Addressing the challenges of retaining, securing, and preserving digital information for the long-term

- 1. Classification**
- 2. Service Level Management**
- 3. Automating Policy Management**

What's Driving The Need for Classification *TODAY*

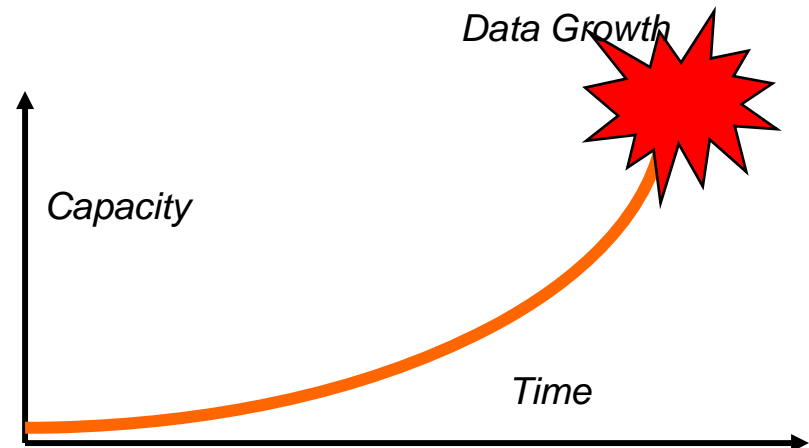
Corporations are *saving everything*, because

- ◆ They are unsure about the value of their information
- ◆ They are being litigated
- ◆ They are complying with government regulations

Resulting in

- ◆ Massive amounts of information growing at fantastic rates
- ◆ Information security breaches
- ◆ Lots of money being spent for governance and compliance

Corporations are balancing *IT Infrastructure and Management Costs* against *Information Risk Management*



What's Driving The Need for Classification *TODAY*

Corporations are *saving everything*, because

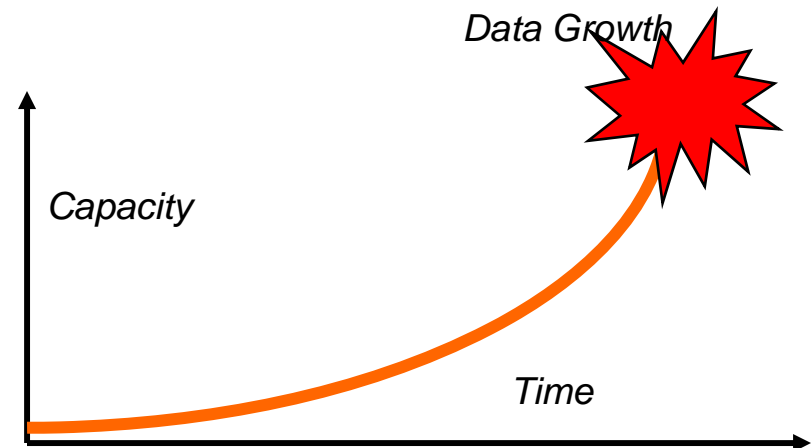
- ◆ They are unsure about the value of their information
- ◆ They are being litigated
- ◆ They are complying with government regulations

Resulting in

- ◆ Massive amounts of information growing at fantastic rates
- ◆ Information security breaches
- ◆ Lots of money being spent for governance and compliance

3 drivers for automated classification:

1. Risk Management
2. Reduce Storage TCO
3. Improve Productivity



Classification Driver #1

Risk management

- ◆ Compliance:
 - › Payment Card Industry Data Security Standard (**PCI**)
 - › Health Insurance Portability and Accountability Act (**HIPAA**)
 - › New Federal Rules of Civil Procedure (**FRCP**)
 - › EU Directive on Privacy and Electronic Communications (2002/58/EC)
- ◆ Information Security
 - › Protecting Personally Identifiable Information (**PII**)
 - › Data Leakage Prevention (**DLP**)

- ◆ *\$80B spent on compliance by 2009*
- ◆ *Compliant records growing 60%/yr at > 2PB in 2007*
- ◆ *Fastest growing application segment of storage industry*

Top 10 Customer Data-Loss Incidents Since 2000*

Number of affected customers	Date of initial disclosure	Company / Organization
94,000,000	2007-01-17	TJX Companies Inc.
40,000,000	2005-06-19	Visa, et al
30,000,000	2004-06-24	America Online
26,500,000	2006-05-22	U.S. Department of Veterans Affairs
25,000,000	2007-11-20	HM Customs and Revenue
8,637,405	2007-03-12	Dai Nippon Printing Company
8,500,000	2007-07-03	Fidelity National Information Services
6,300,000	2007-09-14	TD Ameritrade
6,000,000	2008-05-11	Chilean Ministry of Education
5,000,000	2003-03-06	Data Processors International

* Source: Fred Moore, Horison, Storage Spectrum 2006

*From <http://etiolated.org/> July 2008

➤ eDiscovery and records management coming together

- ◆ Driven by huge costs and risks
- ◆ Changes to the Federal Rules of Civil Procedure
 - › Electronically Stored Information (**ESI**) is subject to production (the way it is managed from cradle to grave will impact costs and risks of eDiscovery)
 - › There will be an early “**meet and confer**”
 - › Word “**preserving**” appears in the rules for the first time
 - › There is a need to understand the “**sources**” of ESI
- ◆ Average eDiscovery costs can run into the millions of dollars per event



Check out SNIA Tutorial:

Compliant Storage: The Risks of Retention and Deletion in the Face of FRCP



Classification Driver #2

Storage TCO

- ◆ External disk storage purchase projected to grow at 52% annually (capex)
- ◆ Capacity is #1 storage issue – driven by email & unstructured data
- ◆ Significant transition to disk-based archival storage
- ◆ Digital archive capacity will increase nearly tenfold between 2005 and 2010



Check out SNIA Tutorial:

The Secret Sauce of ILM: The ILM Assessment Core

October 2008

IDC report – by 2011:

- 1.773 zettabytes, ~60% CAGR
- 70% created by individuals
- 85% managed by organizations

Storage Capex vs. Opex

- ◆ Opex may be as high as 75% of TCO*
- ◆ Impact of data management is rising
- ◆ Greater savings possible through more effective processes

*Source – wikibon: http://www.wikibon.org/Storage_CAPEX_vs_OPEX

Classification Driver #3

Improved productivity

- The average knowledge worker spends *six hours per week* searching for information
 - ◆ 50% of all searches fail to locate desired information
 - ◆ 15% of the average knowledge worker's time is spent recreating existing information
- **Need**
 - ◆ Better organization of information
 - ◆ Accurate search
 - ◆ Consistent management of information
 - ◆ Shortened "time-to-information"





Education

Classification

Classification, Taxonomy & Ontology

➤ Classification:

- ◆ Organizing entities in groups for management purpose
- ◆ Generic term can be applied across many levels of implementation

➤ Taxonomy:

- ◆ Classification based on concepts or syntax
- ◆ Tree – single inheritance
- ◆ Strong taxonomies often used in classification systems

➤ Ontology:

- ◆ Taxonomy-based classification plus:
- ◆ Graphs & relationships among entities (various levels of sophistication)
- ◆ Enables automated reasoning & interoperation of intelligent agents
- ◆ Standards allow programs to access info structure & content
 - › RDF = Resource Description Format
 - › OWL = Web Ontology Language

➤ Folksonomy:

- ◆ The “wisdom of the masses”
- ◆ How many enterprises actually classify their information

The Benefits of Coming Together: Example #1

Legal, Records Manager, I.T collaborate to save money by employing automated classification of eRecords of Exited Employees

<i>Case Study Measurements and Assumptions</i>	<i>Results</i>
Average number of electronic files per exited employee	601
Average number of emails per exited employee	7418
Average time to manually classify each document	2.25 minutes
Number of documents manually classified per day	180
Annual Salary of a qualified records technician (loaded)	\$55,000
Cost to manually classify per exited employee	\$12,473
Cost to outsource the autotclassification of documents	\$800-1200 per exited employee

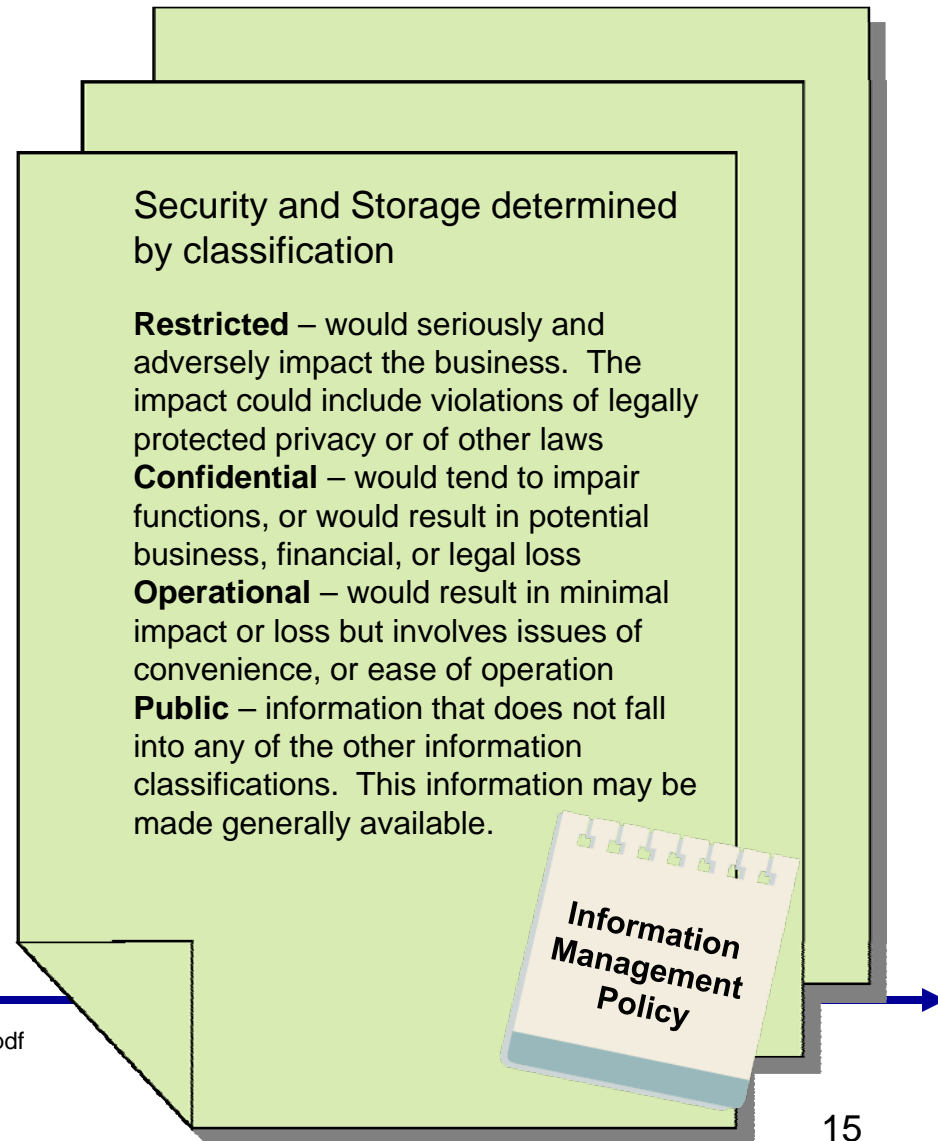
Source: ARMA 2005 International Conference
 "Classifying e-Records of Exited Employees: Case Study Using an Automated Classification Tool", Session Number T027

The Benefits of Coming Together: Example #2

Hospital establishes cross-functional group that will be responsible for information management

➤ Information is *Restricted, Confidential, Operational, Public*

- ◆ *How important is it to keep this information confidential?*
- ◆ *How important is it's integrity?*
- ◆ *How important is it's availability?*



Source: <http://www.psychiatry.ufl.edu/security/docs/InformationClassificationDept.pdf>

The Benefits of Coming Together: outcome

Information Type	Confidentiality	Integrity	Availability	Classification
Medical record	High	High	High	Restricted
Administrative Documents with Private Health Information (PHI)	High	Medium	Medium	Restricted
Email with PHI	High	High	High	Restricted
Patient claims or billing information with PHI	High	Medium	High	Restricted
Research information with PHI	High	High	Medium	Confidential
Budget Information	Medium	Medium	Medium	Confidential
Financial Reports	Low	High	Low	Confidential
Personnel and fiscal operations records without PII	Medium	Medium	Medium	Confidential
Research Proposals	Medium	High	Medium	Operational
Memos without PII	Low	Medium	Medium	Operational

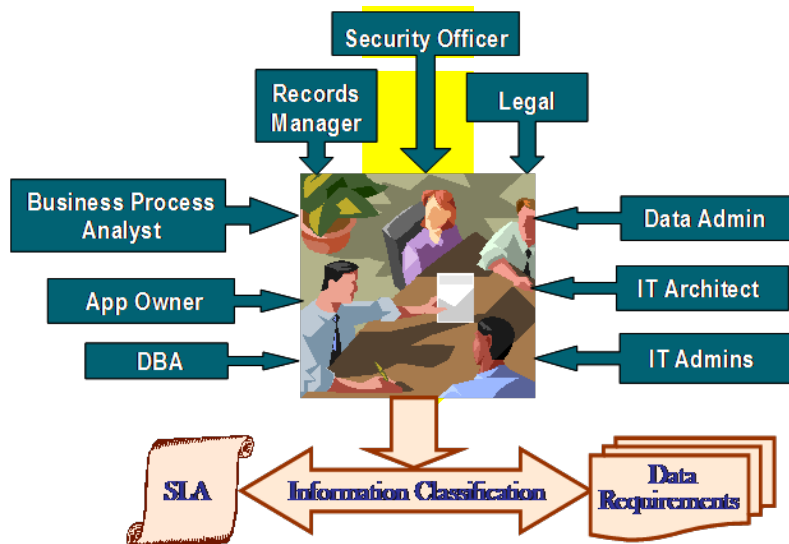


Source: <http://www.psychiatry.ufl.edu/security/docs/InformationClassificationDept.pdf>

October 2008

Gathering requirements from stakeholders

- Information is simply data to the data center
 - ◆ I.T. manages data: software, files, volumes, bits and bytes
 - ◆ Information is data with context: business decisions are based on *information*
 - ◆ Use a collaborative process to identify information service requirements



Stakeholder	Risk Management			TCO
	Security & Risk Mitigation	Litigation Support	Records Mgmt	Cost & Performance Mgmt
Chief Security Officer	✓			
Chief Legal Officer	✓	✓	✓	
Corporation Counsel		✓	✓	
Records Information Manager		✓	✓	
Chief Compliance Officer	✓	✓	✓	
Chief Risk Officer	✓			
Chief Financial Officer	✓			✓
Line of business	✓	✓	✓	✓
Chief Information Officer	✓	✓	✓	✓

- Collaboration enables I.T. to:
 - ◆ Identify and mitigate competing stakeholder requirements
 - ◆ Create data management policies

- Classification must address multiple perspectives
- Support both overlapping and non-overlapping requirements



Education

Service Level Management

How is Data Classified?

Automated Data Classification methods:

Classify by business process or application policies

- > All data assigned same classification
- > Simple; good start; a first approximation
- > Net effect: ranking of applications to service tiers
- > Somewhat effective for TCO Management
- > Possibly effective for Risk Management (very coarse grained solution)

Classify by metadata-based policies

- > Time last accessed, owner, file name, path, etc
- > Useful for aligning data to tiers of service
 - E.g., the CEO's email receives different service than yours
- > Or for placement of data to appropriate stage within a service tier
 - E.g., Hierarchical Storage Management (HSM) for a file server
- > Effective for TCO Management
- > Possibly effective for Risk Management (coarse grained solution)

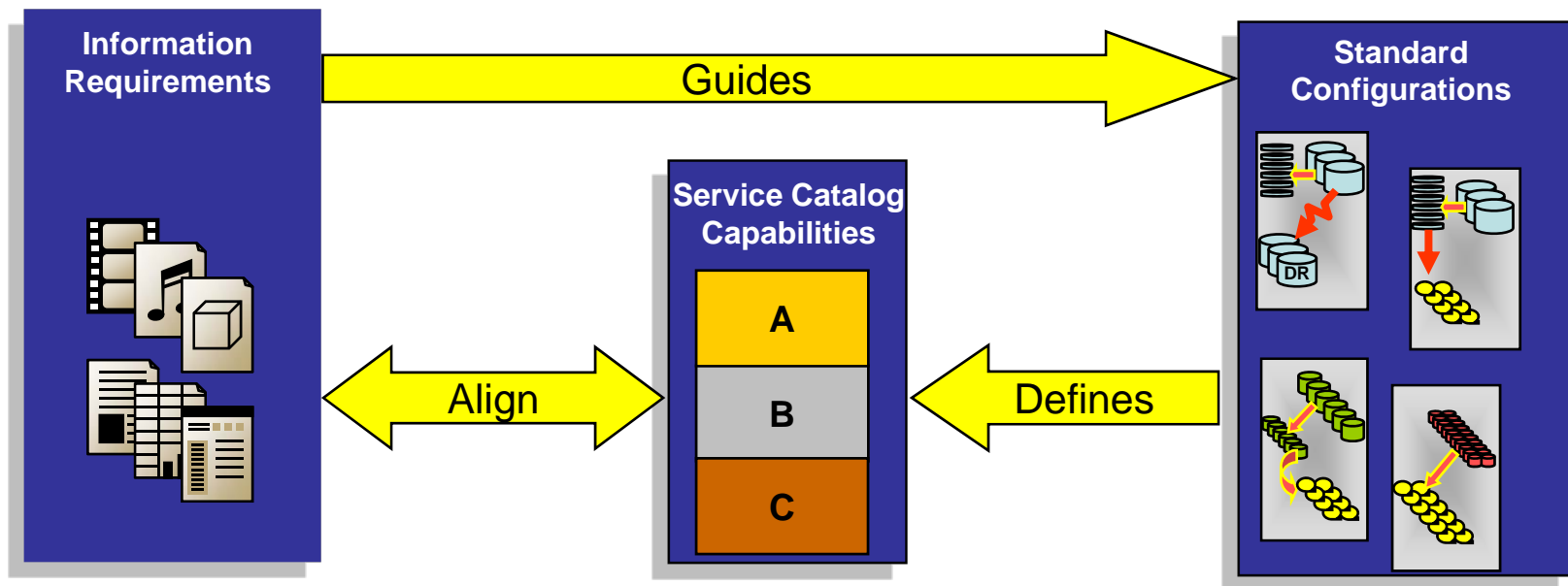
Classify by content-based policies

- > Content-driven alignment of data to service level requirements
- > Added value for Business Intelligence, Compliance & eDiscovery
- > Mostly not applicable for TCO Management
- > Most effective for Risk Management

Service Level Alignment

To address TCO, Service Delivery, and/or Risk Management

- Information Classification gathers requirements to guide configurations
- Standard configurations described using Service Level Metrics
- Data classification can align data to resources by Service Levels



Service catalog capabilities may also include information policies

Service Delivery – Customer IT Sample*

	Scheme	Specification	Class 1	Class 2	Class 3	Class 4
Storage	Guaranteed Performance	Performance throughput per port (IOPS)	5,000+	Up to 5,000	Up to 3,500	Up to 1,500
		Response time (ms)	< 8ms	7-14ms	12-30ms	12-30ms
	Availability	Maximum unplanned downtime per year (mins)	< 26.5	< 26.5	< 52.5	< 263
Archive	Performance	Response time	< 1 second	< 1 second	< 24 hours	
		Throughput	<= 300 Mbps	<= 700 Mbps	<= 280 Mbps	
	Availability	Maximum downtime (yr)	<5.25 mins	<52.56 mins	< 175.2 hours	
	Retention & Disposition	Retention period	< 30 years	< 10 years	< 3 years	
		Data shredding compliance	Yes	No	No	
	Data Integrity	Guarantee of authenticity	Yes	No	No	
	Accessibility	Annual access frequency	< Hourly	< Hourly	Daily	
Offsite	Recovery point objective	< 1 minute	< 28 hours	< 38 hours		
Operational Recovery (OR)	Recovery Classification	Recovery classification	Complete app. restore	Complete app. restore	File or file sys. restore	File or file sys. restore
	Recovery Point Objective (RPO)	Amount of data loss	1 hour	24 hours	24 hours	30 days
	Recovery Time Objective (RTO)	Time required for recovery	< 30 minutes	< 30 minutes	7 GB/minute	.5 GB/minute
	Recoverability	Ability to recover backed up data	100%	100%	98%	95%
	Retention period	Time data is retained	2 hours	24 hours	3 Weeks	15 months
Disaster Recovery (DR)	Recovery Point Objective (RPO)	Amount of data loss	0 minutes	< 4 hours	24-48 hours	24-48 hours
	Recovery Time Objective (RTO)	Time to restore data	< 2 hours	<12 hours	< 48 hours	<72 hours

*Courtesy of EMC² Consulting

Adding Service Levels for Info-centric Policy Mgmt

	Scheme	Attributes	Class 1	Class 2	Class 3	Class 4
Primary Storage	Performance	Throughput (IOPS)	Up to 5000	Up to 3,500	Up to 1,500	Up to 500
		Response time (ms)	< 8ms	7-14ms	12-30ms	12-30ms

Indexed	Data search capability	Index method	Full Text	Metadata	None	
Archive	Performance	Response time	< 1 second	< 1 second	< 24 hours	
		Throughput	<= 300 Mbps	<= 700 Mbps	<= 280 Mbps	
		
	Retention	Retention period	< 30 years	< 10 years	< 3 years	
	Data Integrity & Authenticity	Guarantee of immutability	Yes	No	No	
User ID	Authentication	ID challenge method	Physical	Logical	Logical	None
	User Login	Scope of login	Unique	Unique	Federated	Federated
Audit	3 rd party log gather	Audit action level	Intervene	Alert	Monitor	None
	Audit Log Storage	Guarantee of immutability	Yes	Yes	No	No
Data Security	Encryption	Data at rest	Tamper-Proof	Centralized	None	
		Secure data movement	Enterprise	Application	None	
		Data shredding at deletion	Destroy	DoD Shred	Simple Delete	
	Information Rights Mgmt	Customer Data Access	PII-Private	PCI-CC	Confidential	Basic
Operational Recovery (OR)	Recovery Point Objective (RPO)	Amount of data loss	1 hour	24 hours	24 hours	30 days
	Recovery Time Objective (RTO)	Time to restore data	< 30 minutes	< 30 minutes	7 GB/minute	.5 GB/minute
	Retention	Retention period	2 hours	24 hours	3 Weeks	15 months
Disaster Recovery (DR)	Recovery Point Objective (RPO)	Amount of data loss	0 minutes	< 4 hours	24-48 hours	24-48 hours
	Recovery Time Objective (RTO)	Time to restore data	< 2 hours	<12 hours	< 48 hours	<72 hours

Corresponding Solution Profiles

	Scheme	Attributes	Class 1	Class 2	Class 3	Class 4
Primary Storage			Hi-end FC RAID-1	Hi-end FC RAID-5	Mid-tier FC RAID-5	Mid-tier ATA RAID-5
Indexed			Text Indexing policy			
Archive			CAS	ATA	Tape	
User ID			HW Token, Local Login	SW Token, Local Login	SW Token, LDAP Login	LDAP Login
Audit			SIEM Intervention, on CAS	SIEM Alerts, on CAS	SIEM Monitoring, on disk	No SIEM Monitoring
Data Security			Secure/HW key vs. central key manager			
			IPSEC, SSL, None			
			HW destroy, DoD "shred", standard delete			
			IRM Key Manager policies			
Operational Recovery (OR)			Many Snaps	Fewer Snaps	B2D / VTL	B2T
Disaster Recovery (DR)			Sync Mirror	Async Mirror	Remote DiskCopy	B2T

Service Level Management Summary

➤ Broad context

- ◆ Business / IT interaction (SLAs)
- ◆ IT /IT interaction (OLAs)
- ◆ Service Packaging (Service Catalogs)
- ◆ Plus impact to all data center process areas...

➤ Resource Management

➤ Information-centric Policy Management

Service Level Management offers a critical, abstract, expression language for resource and policy management

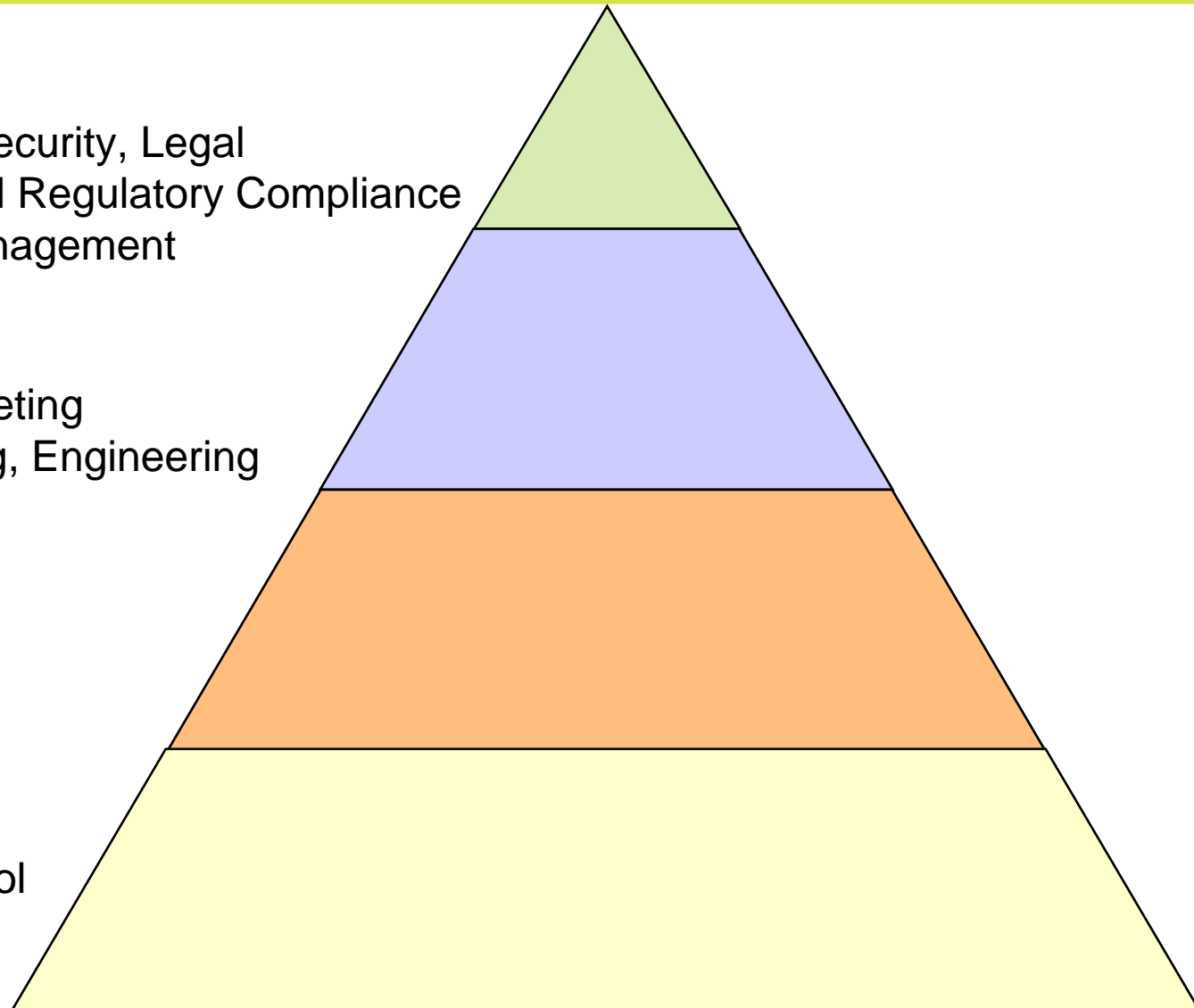


Education

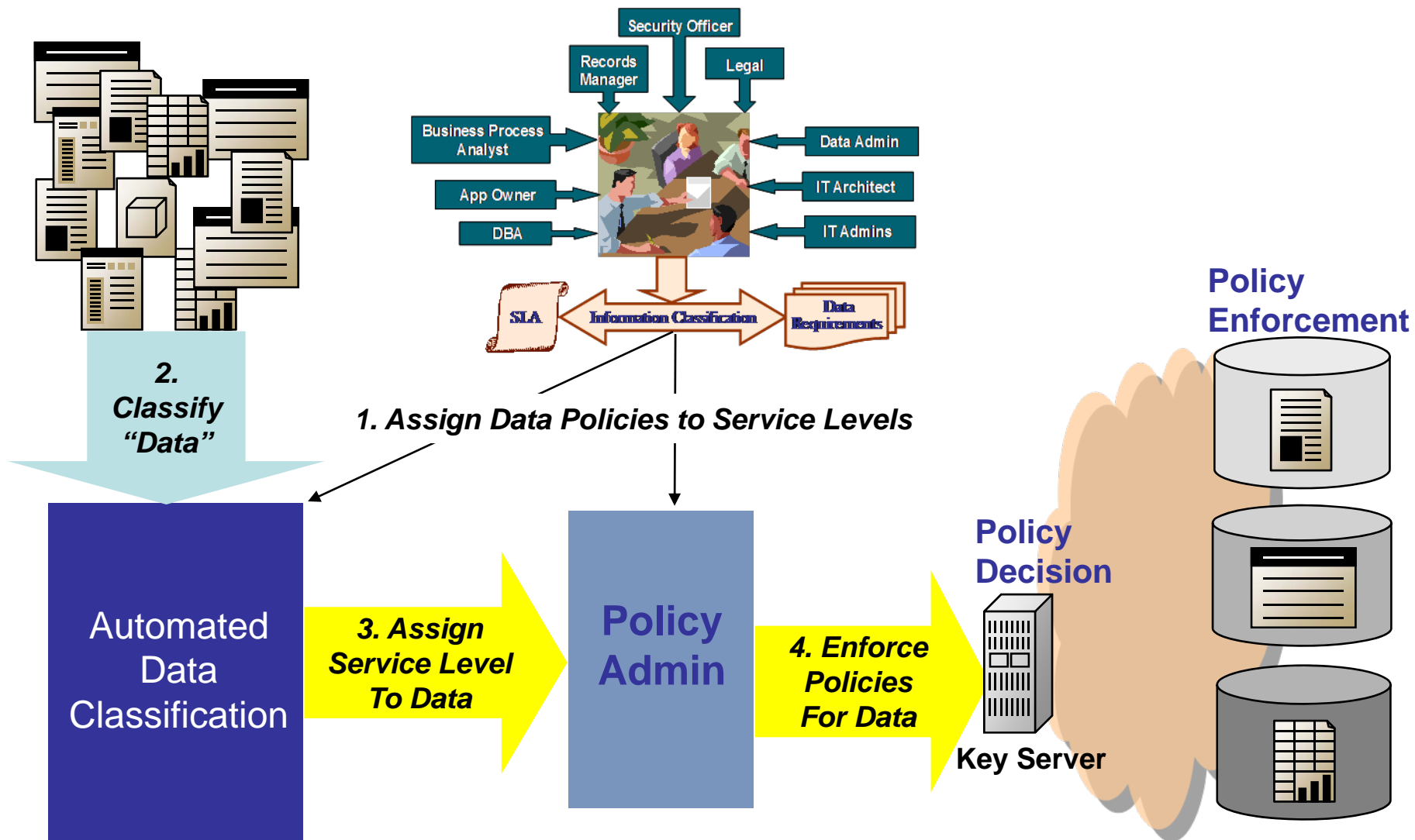
Looking Forward: Automating Policy Management

Many Policy Domains

- **Corporate**
 - ◆ Information Security, Legal
 - ◆ Governmental Regulatory Compliance
 - ◆ Email, IM Management
- **Business Unit**
 - ◆ Finance
 - ◆ Sales & Marketing
 - ◆ Manufacturing, Engineering
- **User**
 - ◆ Security
 - ◆ Distribution
 - ◆ Retention
- **Resources**
 - ◆ Utilization
 - ◆ Consolidation
 - ◆ Change control
 - ◆ CMDB



DMF Vision: Policies driven by Automated Data Classification



- Assign policies to data based on classification
 - ◆ Application, metadata, and/or content
- Policies derived from Information Requirements
 - ◆ Data placement
 - ◆ Security
 - ◆ Information Rights Management
 - ◆ Data Leakage Protection
 - ◆ Data retention
 - ◆ Search & Index
 - ◆ Authentication & Authorization
 - ◆ And others...
- Automated data classification provides scalability

For more information on SNIA's Data Management Forum (DMF) visit the DMF website at

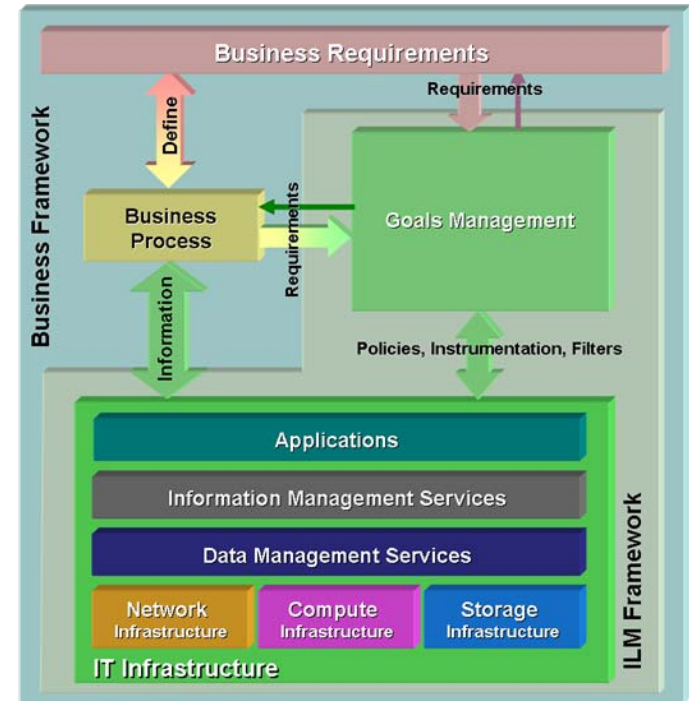
<http://www.snia-dmf.org>



Check out SNIA Tutorial:

Compliant Storage: The Risks of Retention and Deletion in the Face of FRCP

The Secret Sauce of ILM: The Professional ILM Assessment



Please send any comments on this tutorial to SNIA at: trackdatamgmt@snia.org

The DMF would like to thank the following individuals for their contributions to the development of this tutorial:

Edgar StPierre Sheila Childs
Bob Rogers John Field
Terry Yoshii



It's easy
to get
involved
with the
DMF !

- Find a passion
- Join a committee
- Gain knowledge & influence
- Make a difference

www.snia.org/dmf

Abbreviations used in this tutorial

- **Async:** Asynchronous
- **ATA:** Advanced Technology Attachment
- **B2D:** Backup To Disk
- **B2T:** Backup To Tape
- **CAS:** Content-Addressable Storage
- **DoD:** Department of Defense
- **FC:** Fibre Channel
- **IOPS:** Input/Output Operations Per Second
- **IPSEC:** Internet Protocol Security
- **IRM:** Information Rights Management
- **LDAP:** Lightweight Directory Access Protocol
- **HSM:** Hierarchical Storage Management
- **HW:** Hardware
- **ms:** milliseconds
- **PCI:** Payment Card Industry
- **PII:** Personally Identifiable Information
- **SIEM:** Security Incident and Event Management
- **SSL:** Secure Sockets Layer
- **SW:** Software
- **Sync:** Synchronous
- **TCO:** Total Cost of Ownership
- **VTL:** Virtual Tape Library