
Commonwealth of Virginia

Enterprise Technical Architecture (ETA)

Information Domain Report

Version 1.0, July 10, 2006

Prepared by:

[Virginia Information Technologies Agency](#)
ETA Information Domain Team

(This Page Intentionally Left Blank)

Information Domain Team Members

David Froggatt VITA, Enterprise Applications & Architecture Solutions
 Art Ritter Department of Social Services
 Dick Jones Department of Transportation
 Dennis Moen Department of Health
 Tony Shoot..... Northrop Grumman
 Todd Kissam VITA, Policy, Practice and Architecture
 Mike Hammel (Team Facilitator) VITA, Policy, Practice and Architecture

The [Information Domain](#) team began its work by delineating the team's goals, objectives, and scope of work. Discussions included how the domain interfaces with other architecture domains, present and future directions, and how often the information provided in this document is to be updated. The team also reviewed input from publications and individuals with specialized knowledge. The results of the team's efforts, research and deliberations are provided throughout this document.

Information Domain Report: Version History		
Revision	Date	Description
1.0	07-10-2006	Initial

Review Process

Technology Strategy and Solutions Directorate Review

The domain report was reviewed and approved by Jerry Simonoff, Director and Paul Lubic, the Associate Director of Policy, Practices, and Architecture Division.

Online Review

Participation of all Executive Branch agencies was encouraged through a review and comment period via VITA's Online Review and Comment Application (ORCA). Technology businesses and the general public were also actively encouraged to use ORCA to provide comments. All comments were considered and many resulted in modifications to the final document. Additionally, the Domain team provided the reviewers with responses to their comments.

(This Page Intentionally Left Blank)

Table of Contents

<i>Executive Summary of Information Domain</i>	<i>1</i>
<i>Overview</i>	<i>3</i>
Definition of Key Terms	7
Agency Exception Requests.....	9
<i>Information Scope.....</i>	<i>11</i>
Overall Information Domain Scope	11
Scope of this Report	12
<i>Future Information Domain Initiatives</i>	<i>13</i>
Development of a Federated Metadata Repository.....	13
Provide Data Warehousing Capabilities	13
<i>Domain-wide Principles, Recommended Practices and Requirements</i>	<i>15</i>
Domain-wide Principles.....	15
Domain-wide Recommended Practices	17
Domain-wide Requirements:.....	18
<i>Information Domain Technical Topics.....</i>	<i>21</i>
Reporting	21
Ad Hoc End-User Reporting.....	22
Standardized/Canned Reporting	25
Online Analytical Processing (OLAP).....	27
Query	29
Precision/Recall Ranking.....	29
Classification	29
Pattern Matching.....	30
Data Management	31
Data Standards	31
Data Classification (security and access)	34
Metadata Repository/Management	36
Data Cleansing.....	37
Data Profiling.....	38
Enterprise Data	39
Federated Data	40
Business Intelligence	41
Data Warehouse / Data Marts	41
Operational Data Stores	46
Extraction, Transformation and Loading	46
Data Storage Structures.....	48
Data Mining	49
Demand Forecasting and Management.....	49
Balanced Scorecard.....	49
Decision Support and Planning.....	49

Business Analytics Suites	49
Dashboards.....	50
Knowledge Management	51
Information Retrieval.....	51
Information Mapping/Taxonomy.....	51
Information Sharing	51
Categorization	51
Knowledge Engineering.....	51
Knowledge Capture	51
Knowledge Discovery.....	52
Knowledge Distribution and Delivery	52
<i>Glossary</i>	53
<i>Appendices</i>.....	55
References and Links.....	55

Executive Summary of Information Domain

The Information Domain Report is written to assist business and technical leaders in state agencies and central services in making sound decisions related to data warehouse design and acquisition of data warehouse, business intelligence, and other reporting tools and products. The Information Domain also provides a framework for defining responsibility for data integrity and distribution. A well-defined ETA Information Domain will enable the Commonwealth of Virginia to leverage the most value from its data assets. It includes the technology topics of Reporting, Data Management, Business Intelligence, and Knowledge Management.

There is a distinction between the concepts of “[Data](#)” and “[Information](#)”. “Data” is stored in multiple applications systems on multiple platforms using multiple methods across the state and is used to perform day-to-day operations. If this distributed data is grouped together in a meaningful format, it can provide valuable “Information” to the state’s business organizations, decision-makers and the public.

“Data” is captured using online transaction processing (OLTP) applications to perform mission-critical daily operations. Typically, many users simultaneously add, modify, delete and view data using these applications. OLTP applications are characteristically designed to perform transactions one record at a time. The Database technologies used to support OLTP applications are included in the ETA Database Domain.

“Information” is often derived from online analytical processing (OLAP) systems used for analysis, planning, and management reporting. An OLAP system usually references information that is stored in a data warehouse.

“Data” and “Information” are extremely valuable assets to the Commonwealth of Virginia. The ETA Information Domain defines an infrastructure for providing high quality, consistent data to be used as the basis for decision support and executive information services as well as traditional transaction applications statewide. Currently, data is distributed and defined differently by the majority of application systems across the state. Data is often application specific using a variety of formats and semantics. No consistent agency or statewide standards were in place (and for that matter, still aren’t) when the applications systems and corresponding databases were developed. Application systems have historically been monolithic; developed and operated independently from each other. Application development is often driven by statutory, policy, or business needs. Many do not share any logic or data across system or organizational boundaries. The majority of the databases were designed for access by single application systems within a single agency, not for access by multiple application systems in multiple agencies simultaneously. By promoting the concept of federated data, the state will benefit in the areas of reuse, accuracy, security and currency thus making data more shareable than the historical monolithic model. The establishment of a metadata repository is an essential method to achieve and maintain federated data.

The ETA Information Domain documents the approach for the Commonwealth to manage its Information and Data. The goal of the ETA Information Domain is to support the creation of an architecture that:

- Separates transaction-processing systems from large ad hoc queries that are required by analytical, executive decision systems.
- Insulates transaction data systems from the performance and security risk of public Internet inquiries.
- Provides a cross-organizational view of data.
- Promotes cross-organizational sharing of data.
- Provides access to data not found in transaction systems such as summary data and historical data.
- Facilitates enhanced end user access and provide more timely answers to business questions by end users.
- Defines and disseminates information on the stewardship of data to ensure accuracy, security, privacy and ownership.
- Defines data consistently across the state using federated data guidelines.

Overview

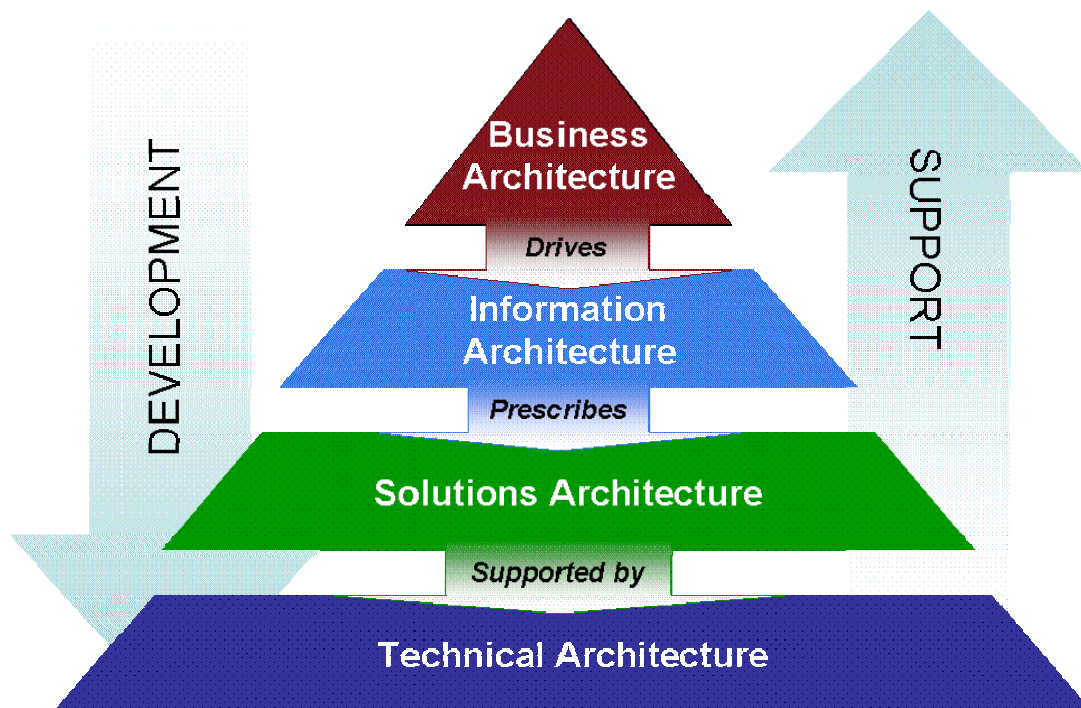
The Commonwealth's Enterprise Architecture is a strategic asset used to manage and align the Commonwealth's business processes and Information Technology (IT) infrastructure/solutions with the State's overall strategy.

The Enterprise Architecture is also a comprehensive framework and repository which defines:

- the models that specify the current ("as-is") and target ("to-be") architecture environments,
- the information necessary to perform the Commonwealth's mission,
- the technologies necessary to perform that mission, and
- the processes necessary for implementing new technologies in response to the Commonwealth's changing business needs.

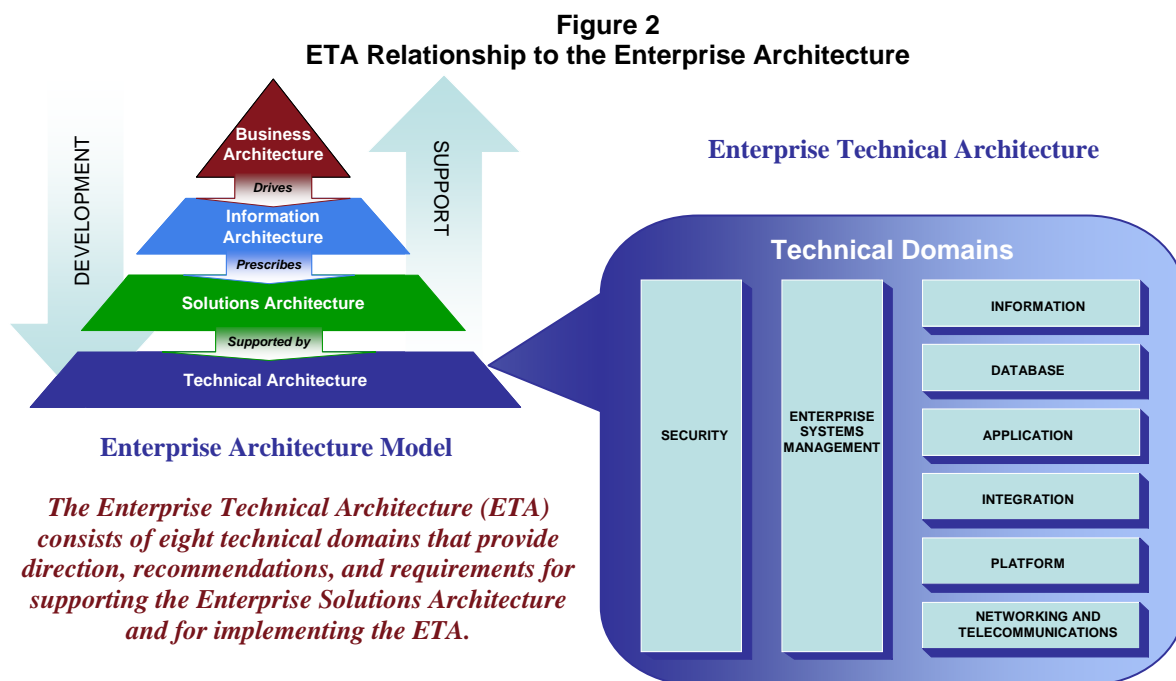
The Enterprise Architecture contains four components as shown in the model in Figure 1.

Figure 1
Commonwealth of Virginia Enterprise Architecture Model



The Business Architecture drives the Information Architecture which prescribes the Solutions Architecture that is supported by the Technical (technology) Architecture.

The Enterprise Technical Architecture (ETA) shown in Figure 2 consists of eight technical domains that provide direction, recommendations and requirements for supporting the Solutions Architecture and for implementing the ETA. The ETA guides the development and support of an organization's information systems and technology infrastructure.



Each of the domains is a critical piece of the overall ETA. The Networking and Telecommunications and Platform Domains address the infrastructure base and provide the foundation for the distributed computing. The Enterprise Systems Management, Database, Applications, and Information Domains address the business functionality and management of the technical architecture. The Integration Domain addresses the interfacing of disparate platforms, systems, databases and applications in a distributed environment. The Security Domain addresses approaches for establishing, maintaining, and enhancing information security across the ETA.

This report addresses the Enterprise Technical Architecture Information Domain and includes requirements and recommended practices for [Virginia's agencies](#)^{1,2}.

¹ This report provides hyperlinks to the domain report Glossary in the electronic version. In the electronic and printed versions, the hyperlinks will have the appearance established by the preferences set in the viewing/printing software (e.g., Word) and permitted by the printer. For example, the hyperlinks may be blue and underlined in the screen version and gray and underlined in the printed version.

²The Glossary entry for agency is critical to understanding ETA requirements and standards identified in this report and is repeated here. **State agency or agency** - Any agency, institution, board, bureau, commission, council, or instrumentality of state government in the executive branch listed in the appropriation act. ETA requirements/standards identified in this report are applicable to all agencies

This report was developed by the Information Domain team, which was commissioned to identify domain related requirements and recommendations. Identified requirements and technology product standards from this domain report will be combined with requirements and technology product standards from other technical domain reports into a single ETA Standard for review and acceptance by the Information Technology Investment Board ([ITIB](#)).

Concerning local governments, courts, legislative agencies, and other public bodies, while they are not required to comply with a requirement unless the requirement is a prerequisite for using a VITA service or for participating in other state-provided connectivity and service programs, their consideration of relevant requirements is highly recommended. This architecture was designed with participation of local government and other public body representatives with the intent of encouraging its use in state and local interconnectivity efforts.

including the administrative functions (does not include instructional or research functions) of institutions of higher education, unless exempted by language contained in a specific requirement/standard.

Commonwealth of Virginia: To-Be ETA

The to-be Enterprise Technical Architecture envisioned for the Commonwealth will be one where the Commonwealth's citizens and other customers who wish to access Virginia services will do so by utilizing an Enterprise Portal via standard web browsers.

Where appropriate, these online government services will be developed, delivered and supported using a Service-Oriented Architecture (SOA) based on open and industry standard solutions. Selected legacy applications will be exposed to the SOA using web services.

The SOA will be supported by an Enterprise Service Bus that provides Orchestration and Choreography Services to the agencies.

Central integration and coordination will be managed by an Integration Competency Center (ICC) that supports agency needs such as: asynchronous message queuing and persistence.

Large complex *mission critical* applications that need to be reliable, scalable, secure and highly available will be n-tiered and will utilize business rule and workflow engines.

Enterprise application software for the core government administrative business functions will be consolidated and the underlying business processes modernized. An Application Management Center of Excellence will service and manage the new enterprise applications that replace existing legacy and silo-based applications.

Data will be exchanged among systems, agencies, institutions of higher learning, localities, the federal government, and partners using XML based standards such as the Global Justice XML Data Model and the National Information Exchange Model.

The number and types of software tools and products used by the Commonwealth will be decreased to reduce complexity. This will create the opportunity for agencies to refocus their current in-house IT resources to achieve higher levels of expertise on the fewer required products resulting in, among other benefits, a lower dependence on outside contractors.

Agency software applications and customer services will be delivered and supported by an IT infrastructure that will:

- Be responsive, agile, modular, scalable, reliable, secure, and highly available (24x7)
- Utilize ITIL (IT Infrastructure Library) best practices
- Have extensive and proactive technology refreshment
- Utilize a shared services model for technology delivery
- Have a single secure state-wide [network](#) and Intranet
- Have a state-of-the-art data center and back-up facility
- Consolidate agency servers in their most cost-effective locations

- Unify statewide electronic mail services
- Employ innovative procurements, supplier partnerships, and financing arrangements to fund, expedite, and ensure the performance of future initiatives
- Provide a statewide customer care center
- Improve the cost performance of IT utilized by the Commonwealth

Transition:

The Commonwealth will transition from silo-based, application centric and agency centric information technology investments to an enterprise approach where applications are designed to be flexible. This allows agencies to take advantage of shared and reusable components, facilitates the sharing and reuse of data where appropriate, and makes the best use of the technology infrastructure that is available.

The implementation of the to-be architecture will take some time. It is not the intent of the Commonwealth to force agencies to replace their existing systems. The migration to the to-be architecture will occur as Agencies consider new information technology investments or make major enhancements/replacements to their existing systems. It is important to note that the Commonwealth ETA is not static; it needs to continue to evolve to support changing business strategies and technology trends.

Rationale:

Agencies can achieve the following benefits resulting from the Commonwealth's implementation of the ETA:

- Better responsiveness to changing business needs and rapidly evolving information technologies.
- Greater ease of software application integration and application interfacing.
- Easier secure access to data and information to enable interagency collaboration and sharing.
- Increased levels of application interoperability within the Commonwealth, with other states and municipalities, and with the Federal government.
- Increased sharing and re-use of current information technology assets.
- Faster deployment of new applications.
- Reduction in costs required to develop, support and maintain agency applications.

Definition of Key Terms

All of the Information Domain ETA standards and requirements considered to be critical components for implementing the Commonwealth's ETA are included in this report.

The report presents three forms of technical architecture guidance for agencies to consider when planning or when making changes or additions to their information technology:

- Requirements – mandatory enterprise technical architecture directions. All requirements are included within the ETA Standard.
- Technology Component Standard Tables - indicate what technologies or products that agencies may acquire at a particular point in time. These are mandatory when acquiring new or replacing existing technology or products. All technology component standard tables are included within the ETA Standard.
- Recommended Practices - provided as guidance to agencies in improving cost efficiencies, business value, operations quality, reliability, availability, decision inputs, risk avoidance or other similar value factors. Recommended Practices are optional.

The following terminology and definitions are applicable to the technology component standard tables presented in this report:

Strategic:

This technology is considered a strategic component of the Commonwealth's Enterprise Technical Architecture. It is acceptable for current deployments and shall be used for all future deployments.

Emerging:

This technology requires additional evaluation in government and university settings. This technology may be used for evaluative or pilot testing deployments or in a higher education research environment. Any use, deployment or procurement of this technology beyond higher education research environments requires an approved *Commonwealth Enterprise Technical Architecture Exception*. The results of an evaluation or pilot test deployment should be submitted to the **VITA Technology Strategy and Solutions: Policy, Practice and Architecture Division** for consideration in the next review of the Enterprise Technical Architecture for that technology.

Transitional/Contained:

This technology is not consistent with the Commonwealth's Enterprise Technical Architecture strategic direction. Agencies may use this technology only as a transitional strategy for moving to a strategic technology. Agencies currently using this technology should migrate to a strategic technology as soon as practical. A migration or replacement plan should be included as part of the Agency's IT Strategic Plan. New deployments or procurements of this technology require an approved *Commonwealth Enterprise Technical Architecture Exception*.

Obsolescent/Rejected:

This technology may be waning in use and support, and/or has been evaluated and found not to meet current Commonwealth Technical Architecture needs. Agencies shall not make any procurements or additional deployments of this technology. Agencies currently using this technology should plan for its replacement with strategic technology to avoid substantial risk. The migration or replacement plan should be included as part of the Agency's IT Strategic Plan.

Agency Exception Requests

Agencies that desire to deviate from the requirements or the technology component standards specified in this report shall request an exception for each desired deviation and receive an approved *Enterprise Technical Architecture Change/Exception Request Form* prior to developing, procuring, or deploying such technology or not complying with a requirement specified in this report. The instructions for completing and submitting an exception request are contained within the *Commonwealth Enterprise Architecture Policy*.

(This Page Intentionally Left Blank)

Information Scope

The ETA Information Domain describes the logical structure of databases and the methodology used to correlate data in multiple databases. It also provides a framework for defining responsibility for data integrity and distribution.

Overall Information Domain Scope

The following hierarchy describes the technologies to be addressed in this domain. Included are the priorities () for the development of each component:

Information Domain

Reporting

- Ad Hoc (1)
- Standardized/Canned (1)
- OLAP (1)
- Query (3)
- Precision/Recall Ranking (3)
- Classification (3)
- Pattern Matching (3)

Data Management

- Data Standards (2)
- Data Classification (security and access) (2)
- Metadata Repository/Management (2)
- Data Cleansing (3)
- Data Profiling (3)
- Data Recovery (and backup) (3)
- Enterprise Data (3)
- Federated Data (3)

Business Intelligence

- Data Warehouse / Data Mart (1)
- Operational Data Stores (1)
- Extraction, Transformation and Loading (1)
- Data Storage Structures (1)
- Data Mining (2)
- Demand Forecasting and Management (3)
- Balanced Scorecard (3)
- Decision Support and Planning (3)
- Business Analytics Suites (3)
- Dashboards (3)

Knowledge Management

- Information retrieval (3)
- Information Mapping/Taxonomy (3)

Information Sharing (3)
Categorization (3)
Knowledge Engineering (3)
Knowledge Capture (3)
Knowledge Discovery (3)
Knowledge Distribution and Delivery (3)

Scope of this Report

This report will address all of the components identified above having a priority of one or two.

Future Information Domain Initiatives

Components with a priority of three will be addressed at a future time.

To meet the mission of the Information Domain, other initiatives have been identified to give future direction toward accomplishment of the domain goals.

Development of a Federated Metadata Repository

- Provide agencies with the ability to share their metadata (database related information) with other state agencies.
- Enhanced statewide application collaboration and interoperability.
- Provide a centralized source of information about data for the enterprise.
- Allow the state to leverage its data resources.

Provide Data Warehousing Capabilities

- Allow for replicated and combined data from multiple agencies without changing the originating system.
- Create a user friendly, centralized location for information & decreased burden on programming staff.
- Reduce demand on mission critical operational systems by separating transaction processing from analytical processing.
- Develop data conversion and Data Warehouse retention standards
- Create Federated data by creating uniform data elements and values; cleanse and convert data from transaction-based databases into Federated Data Warehouse.³
- Include ability to allocate costs of Enterprise Data Warehouses maintained by central, e.g. VITA, (rather than agency) organizations.

³ An excellent discussion of “Federated Data Warehouse” can be found in an article written by Colin White titled “[The Federated Data Warehouse](#)” and published in the March 2000 issue of *DM Review* Magazine.

(This Page Intentionally Left Blank)

Domain-wide Principles, Recommended Practices and Requirements

The following [principles](#), [recommended practices](#) and [requirements](#) pertain to all components, in all situations and activities related to the ETA Information Domain. Component specific principles, recommended practices and requirements will be discussed in the next section of the report.

Domain-wide Principles

In addition to the principles identified in the “[Commonwealth of Virginia Enterprise Architecture – Conceptual Architecture](#)”, the Information team identified the following five Information Domain-specific principles:

INF-P-01: Data Ownership – Data and Information are Commonwealth assets, are shareable, and should be treated as if they do not belong to any particular office, program, individual or agency unless prohibited by law.

Rationale:

The Information Domain Team acknowledges that according to the *Code of Virginia*, data is owned by the agency which creates it. The intent of this principle is not to change the *Code of Virginia* (which is clearly beyond the scope of this project) but rather to change how we, as stewards of this valuable asset, think about how data can be used for the greater benefit of the Commonwealth of Virginia. Data is a strategic asset to be shared and easily accessed across the Commonwealth with customers and partners to support better customer service and management decision making. This requires a shift in the data management culture regarding data from an ownership to a stewardship mentality. This will include responsibility for the accuracy, timeliness and integrity of data without any proprietary restriction on its use. Data management owners will follow an established change management process and will notify all affected parties when changes are made to the data. The needs of the Commonwealth as a whole will be considered in every collection, creation, storage, processing, and dissemination activity.

The value of information is not realized if it is held in isolated pockets. Information must be shared to maximize effective decision-making across lines of business and with partners. Information is necessary for decision making to support accelerated business process cycles. Increased access leads to improved integrity and relevance of data.

Implications:

Data and thus information will be reliable. Customers will receive faster, better and higher quality service. Supporting policies regarding security, privacy, confidentiality, information sharing, information integrity, utility and data relevance must be developed and implemented. There will be a need to promote interoperable

information management, such as data warehouses and data access methods that facilitate information availability for decision-making. Data warehouses, metadata and data accesses tools may need to be developed to facilitate information availability for decision-making. Metadata (information about the data, such as source, units of measurement, and collection methods) will need to be developed and made available.

INF-P-02: Data Standardization – Enterprise Data (ED) will conform to a standardized set of data elements and definitions.

Rationale:

Effective information sharing and exchange depends on a shared definition of standard data elements throughout the Commonwealth. To support program decision making, the timeliness and integrity of each data element should meet the information needs of the most demanding user. ED data definitions should be consistent with definitions used by the Commonwealth's suppliers of information, the customers who use that information, and all applicable standards (e.g. State, Federal, national, international, etc).

Implications:

Definitions included in laws and regulations will be clearly stated and, if possible, will reflect the most useful common definition. Collaboration among program and appropriate staff offices will result in clearly defined information and data needs. Redundant data collection, storage, processing, and dissemination will be eliminated. Information will be more easily exchanged and shared with customers and constituents using data standards.

INF-P-03: Information Access – Easy and timely access to data and information without security and privacy being compromised is the rule rather than the exception.

Rationale:

Productivity, decision-making, and customer service all benefit from easy, direct, and timely availability of information. Information should be attainable in the appropriate place, time, format and context as required by applicable state and federal statutes, e.g. FOIA, ADA, Section 508 and paperwork reduction acts. Beyond the legal requirements, easy and timely access to data and information makes sound business sense.

Implications:

For unrestricted information, the right to know should be presumed unless policy or law specifies otherwise. The business necessity of sharing information must be established. Technology must be deployed to distribute and allow access to information. Classification of information must be clearly stated and the rules well defined. Secure information must not be accidentally released.

INF-P-04: Security and Privacy – IT systems will be implemented in adherence with all security, confidentiality and privacy policies and applicable statutes.

Rationale:

This will ensure that confidential and proprietary information is safeguarded and that data that should be public is indeed public.

INF-P-05: Reduce complexity – The enterprise-wide and agency architecture must reduce integration complexity to the greatest extent possible.

Rationale:

Reducing complexity increases the ability of the enterprise to adapt and change. It also helps to minimize product and support costs.

Domain-wide Recommended Practices

INF-RP-01: Total Cost of Ownership – Agencies should adopt a total cost of ownership (TCO) model for applications and technologies which balance the costs of development, support, training, disaster recovery and retirement against the costs of flexibility, scalability, ease of use, and reduction of integration complexity.

Rationale:

Using a TCO model will lead to higher quality solutions, enable improved planning and budget decision-making and reduce the IT skills required for support of obsolete systems or old standards.

INF-RP-02: Software Acquisition Policy – Any agency that purchases software should develop and implement a Software Acquisition Policy that covers the following:

- Determination of what documentation is necessary as evidence of licensing for each type of software the organization owns.
- Designation of centralized, safe location(s) for license documentation.
- Delegation of the responsibility and accountability for purchasing new software, maintaining records and updating the agency software inventory.
- Storage of the original documentation.
- Storage of the original media.

- INF-RP-03: Software Use Policy** – Agencies should develop and implement a Software Use Policy that covers the following:
- The agency's rules for downloading, installing, and using software titles.
 - The review of the terms and conditions for each license to ensure proper usage.
 - A software installation authorization process.
- INF-RP-04: Support** – Agencies should maintain traditional paid-for support or have a risk-mitigation strategy for all software tools and hardware used to develop or support Commonwealth and/or agency "Mission Critical Applications".
- INF-RP-05: Architectural patterns** – Agencies should develop a limited number of "hardened" or standardized architectural patterns. Architectural patterns serve as reusable technology packages that contain all of hardware and software tools (A – Z) needed to develop and support an application/solution.

Objectives:

- Improve how agencies design systems in order to have modern, common, and effective platform reuse.
- Improve how agencies converse and collaborate in the design and sharing of experiences, practices and services.
- Improve how agencies articulate the benefits of systems built on platforms that are composed of standards, patterns and services.

Domain-wide Requirements:

- INF-R-01: Security and Privacy** – All Information Domain IT systems, as listed above, shall be implemented in adherence with all security, confidentiality and privacy policies and applicable statutes.

Rationale:

- Agencies must protect sensitive data (such as: SSN, Credit card numbers, etc.)
- Safeguards sensitive and proprietary information.
- Enhances public trust.
- Enhances the proper stewardship over public information.
- Ensures the integrity of the information.

Agencies should ensure that all of their mission critical applications are or can be adequately supported by the vendors of all hardware and software products used to

support those applications. Software needs to be at current and supportable release levels and support must be available from a reputable supplier that can be purchased as needed. Non-complying software tools and hardware require an approved Commonwealth Enterprise Technical Architecture Exception.

INF-R-02: Software Tools Version/Release Support – All software used to support Mission Critical Information/Business Intelligence Applications shall be on version/release levels that are fully supported by the vendor or third party and have traditional paid-for support available.

Rationale:

Unsupported software is no longer being updated to fix newly discovered security vulnerabilities or other problems that occur due to environmental changes.

Currently no central repository exists in the Commonwealth that captures software tool use for the types of components found within the Information Domain. Maintaining an ‘as-is’ inventory is critical in moving the enterprise to a desirable ‘to-be’ state.

INF-R-03: Maintain Software Tools Inventory – The Commonwealth shall collect data on agency use of software tools, maintain an up-to-date inventory, and perform research in order to create a more effective and efficient environment in support of the Information Domain.

Rationale:

A better understanding of existing tool use would enable the Commonwealth to negotiate lower prices based on larger quantities of products purchased from fewer vendors. Also, reducing the number of products can result in the availability to the agencies of higher levels of support that can be provided with the same or fewer resources.

Many Information Domain activities involve the storage of artifacts by electronic means. These files, source code listings, reports, models, etc. need to be accessible and available today and in the future. In order to ensure that these artifacts remain accessible, they must either be in an open format or, at a minimum, the software tool which manages the artifact needs to have the ability to easily export the item to a format that is accessible and can be transported to other tools as needed.

INF-R-04: Artifact Accessibility – All electronic repositories of Information/Business Intelligence source code, metadata, development artifacts, models, documentation, etc. shall have their contents accessible either by an export facility or by a direct access method. This ability is required to allow the

repository contents to be transferred from one methodology or tool to another as needed.

Information Domain Technical Topics

The Information Domain defines all the components, interfaces and processes for implementing and managing an integrated, cohesive information policy. The following description of each component defines the scope of the component and discusses its place within the Information Domain. In addition, various Recommended Practices may be suggested for the component. Also, for each component, one or more Requirements and/or Product Standards may be specified. Requirements are conditions which must be met, i.e. are required and Product Standards are specifications for the use of specific hardware and software relative to the particular component:

Reporting

The Reporting topic includes the components: Ad Hoc End-User Reporting, Standardized/Canned Reporting, Online Analytical Processing (OLAP), Query, Precision/Recall, Ranking, Classification, and Pattern Matching.

An Enterprise Business Intelligence (EBI) Suite consists of software that will include the first three topics below (Ad Hoc Reporting, Standardized / Canned Reporting, and On-Line Analytical Processing). Examples of the leaders of EBI suites according to Gartner are Business Objects (which has acquired Crystal Reports), Cognos, SAS Institute and Information Builders. Near leaders in this area, using Gartner's analysis, include Hyperion, SAP and MicroStrategy. While one can use separate components from different vendors if licensing allows, there are advantages to going with one vendor to get an integrated suite of tools for business intelligence reporting purposes.

The Data Warehouse Institute summarizes selecting a set of Business Intelligence (BI) tools by looking at the following three major categories of requirements:

- Breadth – “Does it support the diversity of users, from casual report consumers to sophisticated report authors?”
- Depth – “Does it support the features and functionality required to meet user requirements in each category?”
- Scalability – “Does it scale up to support thousands of users with adequate response times for both ad hoc and scheduled queries?”

Gartner, in their January 9, 2006 report on Magic Quadrant for Business Intelligence Platforms, First Quarter 2006, makes the following recommendations for the development and integration of business intelligence tool suites. The following is quoted directly from page 4 of this report:

- “BI Infrastructure – All the tools in the platform should use the same security, metadata, administration, portal integration, object model and query engine, and share the same look and feel.

- **Metadata Management** – In particular, this is a major component of the platform. Not only should all tools leverage the same metadata, but the offering should provide a robust way to capture, store, reuse and publish metadata objects.
- **Web Services Integration** – This enables the BI platform for schedule, deliver, manage and administer information through industry standard Web services integration (such as SOAP and WSDL).
- **Programmatic development environment** – this provides a set of programmatic development tools coupled with a software developer's kit for developers to create BI applications and integrate them into a business process and/or embed them in another application.
- **Visual development environment** – this enables developers to build BI applications without coding by using a wizard like component assembly process.
- **Prepackaged business context** – BI application developers can use queries, reports, performance metrics and business roles to jump-start a project. Also, this capability can provide a data schema for a specific horizontal and/or vertical subject area.
- **Data movement** – this facilitates the movement of data from specific applications and data stores and the loading of the data into a repository optimized for analysis and reporting.”

Recommended Practices:

Data security is an important consideration, and much is restricted based on who the user is and who is trying to view it. For example, in a human resources data mart, it is likely that top management and HR management might need to see all employee records, but a unit manager can only see the records of employees in his or her work unit. Once the security rules are worked out, an organization will only want to apply them once and have all data access tools follow the exact rules.

INF-RP-06: Security Defined Once – Row and column level security should only need to be defined once and used by all BI tools.

Rationale:

The best solution is to implement row level security at the data base level rather than in a reporting tool, but if the latter needs to be done, the reporting tools should all implement the same security model so that different levels of security are not assigned by mistake in the different tools.

Ad Hoc End-User Reporting

Ad hoc query provides business analyst the ability to pose specific questions to produce a result without needing the programming of a report by IT. The ad hoc nature of these queries implies a short shelf life where some situation is being researched or a new opportunity is being explored. The tools falling into this category offer the ability, often through a point and click interface, to search the database and produce a result that can

then be displayed, further refined and analyzed. Knowledge of SQL or other database or programming languages should not be necessary when using these tools, but the ad hoc report writer needs to be very familiar with the data that they are reporting against. The results are often exported to another desktop application such as Microsoft Excel, Word, or PowerPoint. The ability to export the results to MS Excel is especially important.

IT professionals do need to create and maintain the metadata behind the scenes that allows users to navigate through and select the available data sources and data elements. Ideally, these “catalogs” are presented to users in a business friendly manner. Because reports can be created and run on the fly, an important consideration is that ad hoc query tools provide the ability to govern queries for performance and deliver auditing capabilities.

Note: although the stated goal of ad hoc reporting is to free up IT people from having to write reports and turn them over to the users, a review of case studies and documents reveals that this goal is difficult to achieve, and that many users are frustrated with ad hoc query software. As a result, it may end up sitting on the shelf, underutilized. It is relatively easy to free up IT resources at the expense of having end-users struggling to develop useful reports using the ad hoc query tool.

Recommended Practices:

Query tools are very powerful and can easily be used to produce extremely time-consuming queries with no ability to tune them. This can certainly include the use of MS Access queries against corporate data marts using an ODBC connection.

INF-RP-07: Query Restrictions – Ad hoc query tool usage should be restricted to a relatively small subset of very knowledgeable users.

Rationale:

Users who will be given access to ad hoc query tools need to receive adequate training before they are given access to production databases with the tools. Vendors can also “oversell” the ease of use and need to use their tools in order to sell additional licenses. A typical business executive or line worker does not need the advanced features of ad hoc query. Instead, every effort should be made to anticipate typical information needs for most users, and this information should be provided in standardized, pre-programmed reports and/or dashboards.

INF-RP-08: Savvy users – Ad hoc query end-users do not have to be programmers, but they do need to be technically savvy in addition to knowing the data really well.

Rationale:

If an ad hoc query tool requires programming knowledge, then it may be a great tool for the programming staff but not as an end-user tool. Examples of the former

include SQL*Plus and TOAD. Even the easiest tools will require some degree of computer proficiency. Users who will be given access to ad hoc query tools need to receive adequate support from the professional IT staff.

INF-RP-09: Ease of use – Professional IT staff should work diligently to maintain a business-friendly view of the available data.

Rationale:

- Views of the business names for tables and columns, rather than the actual table and column names, should be used. The whole intent of an ad hoc report tool is to allow knowledgeable business users to access information. Using business friendly names should help meet this objective.
- Breaking the information available in manageable sized pieces makes it less likely that a user will try to run an inappropriate query. Therefore, many special purpose business “catalogs” are preferable to one or two huge “catalogs” giving access to everything.
- Also, it will likely be less confusing if there is a different model for each major subject area. The necessary table joins can be made in the metadata model so that queries work properly.

INF-RP-10: Work with subsets – Users should easily be able to work with a small subset of the data of interest while building the report, rather than the full production data set.

Rationale:

This will improve performance while the report is being constructed.

INF-RP-11: IT oversight – Reports developed by end-users that will become production reports for use by others should normally not be converted to production without the review of IT staff.

Rationale:

This practice minimizes the risk of an inefficient report developed by an end user will become widely used by others.

Requirements

Following are requirements that all newly acquired Information/Business Intelligence ad-hoc reporting software tools must support.

INF-R-05: Ability to share queries – An ad hoc end-user tool shall be able to share an ad hoc query with others.

This enables the reuse and efficient utilization of agency resources

Rationale:

Knowledgeable users will likely create very useful reports because they are so conversant with the data. Allowing such reports to be shared with other users as “standardized” reports reuses work without burdening IT resources excessively.

INF-R-06: Intuitive interface – The tool used to build a report shall have an intuitive interface, with “point and click” features for adding elements, filtering data, and sorting the results, with no programming knowledge required.

Rationale:

To keep the tool from becoming “shelf-ware”, it must be easy to use and intuitive. The goal is to give power users access to their data, not require them to become programmers. Many case studies exist of ad hoc query tools that were too difficult to learn or remember, and were thus purchased and never used.

Standardized/Canned Reporting

Static reporting is a repeatable, pre-calculated request for information. Static reporting is a means of documenting results of such requests in a standard format on a routine basis to a targeted audience. Reports can be pre-formatted or interactive, often including the ability to drill down into more details from summary level reports. Where reporting of this nature is often viewed as hardcopy, it may take on newer form as the Intranet can become a vehicle for fast dissemination of information. Standardized reporting includes the ability to distribute and schedule the reports using a variety of capabilities, such as by e-mail and “bursting” (where each user automatically gets their own view of the data depending on their user class). Gartner views reporting as having reached the mature mainstream status on the “hype” cycle.

The ability to export the results of a report to MS Excel is an important capability of a reporting tool.

Recommended Practices:

Relatively few users will have access to ad hoc reporting, while nearly anyone can access a set of standardized reports.

INF-RP-12: Standardized reports – Most information should be delivered by way of standardized, not ad hoc, reports.

Rationale:

Standardized reports can be tuned for optimal performance, and their use monitored to determine which ones are most useful. Between knowledgeable business users and

a good IT staff, a comprehensive set of useful information available in standardized reports should be feasible.

INF-RP-13: Use of Web – Network attached users should view reports on the web, using web-browser software, without the need for client side software.

Rationale:

Users are likely to be geographically dispersed. It is not feasible to install and update a client plug-in for anyone needing access to reports. Web accessible software is rapidly becoming a standard.

Implication:

Users who must work offline, may have to use client side tools to perform some reports with local data stores.

INF-RP-14: Built-in job scheduling – Facilities to schedule the running and delivery of reports should exist within the software. This must include the ability to deliver reports by way of electronic mail, and the electronic bursting of reports.

Rationale:

Scheduling reports can take the peak load off of the report server. These capabilities are now normal features of the robust report tools that are in the market place.

As an example of “drill down” reporting, the first report might summarize data for a time period by regions of Virginia. A second report might summarize the same data for the counties and cities within a selected region. A third report could show the details that make up the summary for a selected county on the second report.

Not all reports will need this capability, but when suitable, it is very useful to start viewing information at the higher levels and drill into more detail as needed.

INF-RP-15: Use drill-down – Reports should be able to be linked to other reports so that one can drill into more and more detail, starting with a summary view of the data.

Rationale:

This is standard functionality that users will generally expect.

INF-RP-16: Use of graphics – The report tool should allow displaying reports graphically.

Rationale:

Information is often more clearly presented in graphical formats.

INF-RP-17: Multiple reports – The report tool should allow multiple reports and/or graphical displays in a single window or web page.

Rationale:

Having multiple displays on the same page makes it easier to have a consolidated view of one's information and is more user-friendly if designed properly.

“Standardized reports” does not mean “static reports”. Users interacting with standardized reports need to be able to ask for just the data that they want within the confines of the report.

INF-RP-18: Report filters – Standardized report prompts must allow users to specify the values used to “filter” the data for a given run of the report.

Rationale:

Reporting tools exist that allow both “stand alone” prompt values and “cascading” prompt values. In the latter case, the value chosen for one prompt limits the allowable values for a second prompt. An example of a cascading prompt would be to select a region from one list and be presented with a list of only the counties and cities contained within that region. Cascading values can break a large number of values into much more manageable subsets and improve usability.

Online Analytical Processing (OLAP)

OLAP tools view information in the form of cubes, or multiple dimensions and allow the user to drill down to lower levels of detail and slice across different dimensions such as time or commodity. These tools are generally used by the business analyst in conducting their research to answer business questions as part of the decision making process. OLAP is decision support software that allows the user to quickly analyze information that has been summarized into multidimensional views. Traditional OLAP products, also known as multidimensional OLAP, or MOLAP, summarize transactions into multidimensional views ahead of time. User queries on these types of databases are extremely fast because the consolidation has already been done and the database does not have to be accessed each time. OLAP places the data into a cube structure that can be rotated by the user, which is particularly suited for financial summaries. Putting this functionality in the hands of the agency's power users allows them to ask their own questions and gives them quick and easy access to the information they need. One tool however does not fit all. The Business Intelligence tools arena still requires that we match the right tools to the right end user. It should be noted that Online Transaction Processing (OLTP) systems are updated as soon as transactions are entered at terminals or received over communications lines. It also implies that confirmations are returned to the sender. They are considered “real-time” systems and are covered in the ETA Database Domain. Gartner views OLAP as having reached the mature mainstream status on the “hype” cycle.

Recommended Practices:

INF-RP-18: OLAP tool restrictions – OLAP tool usage should be restricted to knowledgeable users who have the need to analyze data trends and summarize information.

Rationale:

OLAP, by its very name, is for analytical users. There is no need to make this available to casual users who do not have the need to analyze trends and “slice and dice” data in various combinations. People who just need basic data or information should get it from preprogrammed reports, not OLAP.

INF-RP-19: OLAP training – OLAP users should receive adequate training in the capability of the tool and should understand the data and its limitations as presented in the cube.

Rationale:

A typical OLAP tool has many features, and at least some training is needed to make users aware of how to use the tool properly.

INF-RP-20: Ability to save subsets – The OLAP tool should allow users to create and save their own customized subsets of data in the cube.

Rationale:

This type of functionality is useful when a user wants to compare information across dimensions, for example comparing the numbers for several different dates side by side rather than for just a single date.

INF-RP-21: OLAP Graphics – The OLAP tool should allow display of the data in a graphical fashion.

Rationale:

Information is often more clearly presented in graphical formats. Ideally, the tool will allow the selection of the best type of graph or chart by the user, depending on the nature of the data and how they are attempting to interpret or analyze it.

Requirements

Following are requirements that all newly acquired Information/Business Intelligence OLAP software tools must support.

INF-R-07: Drill-down capability – OLAP tools shall have the ability to drill into the details of a cell in an OLAP cube by going to the source database.

Rationale:

Once a user slices into just the dimensions of interest, they are likely to want the details of the summary number.

INF-R-08: OLAP Export – OLAP tools shall have the ability to export the results to a standard spreadsheet format such as .csv or .xls.

Rationale:

Advanced users may want to use the “what if” capabilities of Excel to do further analysis.

To keep the tool from becoming “shelf-ware”, it must be easy to use and intuitive. If software is intuitive, the designers have anticipated what the users will want and will have provided it, where they expect to have it. The goal is to give power users access to their data, not require them to become programmers. Some power user training and re-training should be expected with OLAP under the best of circumstances.

INF-R-09: Easy cube manipulation – The interface to manipulate data in the cube shall have “point and click” and “drag and drop” features for analyzing the available data.

Rationale:

More complicated features should be provided with easy to understand wizard interfaces.

Query

Defines the set of capabilities that support retrieval of records that satisfy specific query selection criteria.

Precision/Recall Ranking

Defines the set of capabilities that support selection and retrieval of records ranked to optimize precision against recall. Precision is related to the search objectives of returning only the information that is relevant to a query and to rank the results by relevance to the users specified search criteria. Recall conflicts with precision with the goal of not leaving out any relevant matches. Precision and recall can be thought of as inversely related. With searching, a determination must be made to give more priority to either precision or recall. The weight, or ranking, of precision vs. recall can be assessed with mathematical formulas.

Classification

This component defines the set of capabilities that support selection and retrieval of records organized by shared characteristics in content or context. The premise behind classifying information is that it can be stored in such a way to allow efficient and effective access to the right information that an organization needs. Classifications can

be informal and rely on descriptive text, or they can be formal, relying on a classification scheme with pre-determined values.

Pattern Matching

Defines the set of capabilities that support retrieval of records generated from a data source by imputing characteristics based on patterns in the content or context.

Data Management

The Data Management topic is concerned with the components that affect the quality, management, meta-management, accessibility, and recovery of electronic data resources.

It includes the components of Data Standards, Data Classification (security and access), Metadata Repository/Management, Data Cleansing, Data Profiling, Data Recovery (and backup), Enterprise Data Management (EDM), and Federated Data.

There are many reasons why data warehouses and business intelligence (BI) systems fail to meet expectations, but one of the most significant points of failure has been a lack of attention to the data itself. Where data is taken into account it is usually only in terms of data integration (i.e. bringing together data from different sources) rather than in terms of data quality. Moving data from its various sources into an easily accessible repository is only part of the challenge in delivering a data warehouse and BI. Without paying attention to the accuracy, consistency and timeliness of the data, BI quickly leads to poor decision-making, increased cost and lost opportunities.

According to industry analyst firm Gartner more than 50 percent of business intelligence and customer relationship management deployments will suffer limited acceptance, if not outright failure, due to lack of attention to data quality issues. The impact of poor data quality is far reaching and its affects are both tangible and intangible. If data quality problems are allowed to persist, executives grow to mistrust the information in the data warehouse and will be reluctant to use it for decision-making. Before long the data warehouse becomes another costly white elephant that has failed to deliver benefit to the business. Data quality is the primary focus of the following Data Management components.

Recommended Practices:

- INF-RP-22: Data management team** – Establish a team to ensure data quality through data analysis, metadata collection, and business process knowledge.
- INF-RP-23: Periodic analysis** – Data columns should be periodically analyzed.
- INF-RP-24: Document business rules** – Research and document in the metadata repository the definition and business rules for the data.

Data Standards

It is important to address the issues of data and data quality through the use of data standards.

Data standards are important in the quest for data integration and consist of a framework used to classify or define data. These standards may include Data Element Naming, Database Object Naming, Metadata Requirements, Data Modeling, and Geo-Spatial Requirements. Information can be captured and exposed via a variety of data types. For example, information can be captured as text, numbers, images, maps, graphics, video and audio. The software used to create data files stores these files in different data formats. These formats can be proprietary and therefore controlled and supported by just one software developer. Formats can also be non-proprietary or open.

Open formats are defined as specifications for data file formats that are based on an underlying open standard and are fully documented and publicly available. The Data Standards component addresses the acceptable formats in which data can be presented and captured.

Information that traditionally has been presented in text form is increasingly being enriched through the use of multimedia data types such as graphics, audio and video. The variety of data formats used however raises concerns regarding interoperability and accessibility. The target state is the ubiquitous use of open formats to capture and store data within applications and in individual data files.

Text and Numeric Fields - Current operational databases are almost completely text and numeric data fields. Since they are discrete values, these can be individually retrieved, queried and manipulated to support some activity, e.g. reporting needs or analysis. These data types will continue to play a significant role in all our databases.

Images - Scanned pictures of documents, photos and other multi-dimensional forms can be stored in databases. The scanned image is a single data field and is retrieved and updated as a single fact. Software outside of the DBMS is used to manipulate the image.

Geographic Data - Geographic data is information about features on the surface and subsurface of the earth, including their location, shape, description and condition. Geographic information includes spatial and descriptive tabular information in tabular and raster (image) formats. A geographic information system (GIS) is a hardware and software environment that captures, stores, analyzes, queries, and displays geographic information. Typically geographic information is the basis for location-based decision making, land-use planning, emergency response, and mapping purposes.

Multimedia: Voice, Animation and Video - Multi-media applications are increasing as we employ new modalities of communicating with users. Voice can be stored in a database to capture instructional, informative messages that can then be played back rather than displayed as text. This facilitates those situations where keyboards and visual displays are difficult to utilize. Graphics, animation and video, likewise, offer an alternative way to inform users where simple text does not communicate easily the complexity or the relationships between informational components. An example might be

graphic displays of vessels and equipment allowing drill down to more detailed information related to the part or component. Video may be useful in demonstrating some complex operation as part of a training program.

Objects - Objects are composites of other data types and other objects. Objects form a hierarchy of information unlike the relational model. Objects contain facts about themselves and exhibit certain behaviors implemented as procedural code. They also "inherit" the facts and behaviors of their parent objects up through the hierarchy. Relational databases store everything in rows and columns. Although they may support large binary object (LOB) fields that can hold anything, an object database can support any type of data combined with the processing to display it.

Recommended Practices:

INF-RP-25: Common data elements – Establish common data element and data object naming standards for all Enterprise Data.

Rationale:

The set of guidelines for naming data elements establishes a naming convention, or classification scheme, that will make it easier to determine if a data requirement is already being met or if it is a new requirement that needs to be fully defined and the data collected and distributed as necessary.

1. The names of data elements should:
 - a) Be based on the names of data entity attributes identified in the data model.
 - b) Be clear, accurate, and self-explanatory.
 - c) Be named according to logical (business process) and not physical considerations. Physical characteristics include any connotations regarding technology (hardware or software), physical location (databases, files, or tables), organization (data steward), or function (systems, applications, or programs).
 - d) Consist of the minimum number of words that categorize the data element. Fewer words may be too general while more words may be too narrow or restrictive. Modifiers may be used with class words, generic elements, and prime words to fully describe generic elements and data elements. Modifiers are often derived from the data entity attribute names and the entity names identified in the data model or an approved data model linked (mapped) to the data model.
 - e) Include only alpha Numeric characters (A-Z "CAPS"; 0-9), and underscore (_).
 - f) Have each component of the name separated by an underscore.
 - g) Have multiple word prime words connected with underscore. Examples of multiple prime words might be "Purchase_Order," "Medical_Facility," or "Civil_Government."
2. The following are not permitted in data element names:

- a) Words which redefine the data element or contain information that more correctly belongs in the definition.
- b) Class words used as modifiers or prime words.
- c) Abbreviations or acronyms. (Except those traditionally accepted in the Commonwealth; VITA has a list of approved Abbreviations and acronyms.)
- d) Plurals of words.
- e) The possessive forms of a word, i.e., a word which denotes ownership.
- f) Articles (e.g., a, an, the).
- g) Conjunctions (e.g., and, or, but).
- h) Verbs.
- i) Prepositions (e.g., at, by, for, from, in, of, to

Requirements:

Following are requirements that all newly acquired Information/Business Intelligence software tools must support.

INF-R-10: Standard file formats – Agencies shall ensure that all software tools or packages that create files or data stores do so in a format that is based on an underlying open or de facto standard or provides the capability to export to such a format.

Rationale:

Examples of the open or de facto standards which software should be consistent with would include PDF files from Adobe, generic .CSV (comma separated value) files, the Data Interchange Standards Association (DISA), Object Management Group (OMG), World Wide Web Consortium (W3C), Federal Geographic Data Committee (FGDC), and the National Information Exchange Model (NIEM).

Data Classification (security and access)

Within the context of this component, the term “data” generally refers to electronically maintained information. It must be classified according to its degree of sensitivity in a universally understandable manner. The degree of sensitivity can be determined by applying the appropriate State, Local or Federal laws or regulations to the data. As an example, the data sensitivity classifications might be personal data, confidential, secret, or top secret. Sensitivity levels are determined by the type of information that is in an automated system. The information that has the least amount of sensitivity might include things such as summary revenue and expense data for the Commonwealth. This type of information is widely available from many sources. Data that is made generally available without specific custodian approval and that has not been explicitly and authoritatively classified as confidential is not considered sensitive. Highly sensitive information would include information that must be protected to meet state and federal Privacy Act requirements including data such as social security numbers, credit card numbers, criminal and medical histories, etc. It is also data whose loss, corruption, or unauthorized

disclosure would be a violation of state and federal statutes, mandates and regulations. Confidentiality is understood to be a continuum wherein some data/information is more sensitive than other data/information and shall be protected in a more secure manner. The term "in a universally understandable manner" implies there should be standard definitions for the different sensitivity classifications. In addition, the data needs to maintain its security classification as it traverses any physical or logical boundary such as an agency, computer-related device, network, or software application system.

Requirements:

INF-R-11: Sensitivity classification – Data that is sensitive shall be classified by the agency according to its degree of sensitivity in a universally understandable manner.

INF-R-12: Security classification – Data that requires a security classification shall maintain its security classification as it traverses any physical or logical boundary such as an agency, computer-related device, network, or software application system.

Rationale:

The classification of data is governed by the Code of Virginia §36-105.3 and §44-146.22 and 49 CFR Part 1520, FOIA, COV §2.2-3705.2 and other authorities as follows:

"Government Data Collection and Dissemination Practices Act." § 2.2-3800 - § 2.2-3809

Identity fraud; penalty; victim assistance § [18.2-186.3](#).

Fair and Accurate Credit Transactions Act of 2003 (*sections pertaining to identity theft effective 12/31/2004*)

Fair Credit Reporting Act (FCRA)

Identity Theft and Assumption Deterrence Act of 1998 Title 18 United States Code - Section 1028 *Fraud and Related Activity in Connection with Identification Documents and Information*

THE PRIVACY ACT OF 1974 5 U.S.C. § 552a - *Section 7 of the Privacy Act (found at 5 U.S.C. § 552a note (Disclosure of Social Security Number))*

Family Educational Rights and Privacy Act, 20 U.S.C. §1232

Metadata Repository/Management

Metadata is information about data. A metadata repository is an asset and inventory database that tracks and manages the information about the applications and databases that have been developed for conducting the state's business. A metadata repository can substantially help organizations coordinate and support the applications and databases, which in turn support the business. A metadata repository catalogs business and technical metadata, business rules, data ownership, and other information about the state's data. It provides the necessary foundation or infrastructure for an effective data management program.

The Metadata repository is itself a database containing a complete glossary for all components, databases, fields, objects, owners, access, platforms and users within the enterprise. The repository offers a way to understand what information is available, where it comes from, where it is stored, the transformation performed on the data, its currency and other important facts about the data. It describes the data structures and the business rules at a level above a data dictionary. Metadata has however taken on a more visible role among day-to-day knowledge workers. Today it serves as the main catalog, or map to a data warehouse. The central metadata repository is an essential part of a data warehouse. Metadata can be generated and maintained by an ETL tool as part of the specification of the extraction, transformation and load process. The repository can also capture the operational statistics on the operation of the ETL process. Ideally, access to data definitions and business rules in the metadata repository should be end user accessible. The end result is Enterprise metadata that is defined consistently across the state, is re-useable, shareable, accurate, up-to date, secure, and managed from a statewide perspective. This cannot be effectively accomplished without the use of a Metadata Repository. Specifically Enterprise metadata is required when there is a need for managing and accessing metadata across physical boundaries in a secure manner. Those physical boundaries might be the result of community-of-interest, line of business, system, department, or enterprise separation.

Kimball⁴ lists the following types of metadata in a data warehouse:

- source system metadata
 - source specifications — such as repositories, source schemas
 - source descriptive information — such as ownership descriptions, update frequencies, legal limitations, access methods
 - process information — such as job schedules, extraction code
- data staging metadata
 - data acquisition information — such as data transmission scheduling and results, file usage
 - dimension table management — such as definitions of dimensions, surrogate key assignments
 - transformation and aggregation — such as data enhancement and mapping, DBMS load scripts, aggregate definitions

⁴ [Ralph Kimball](#), The Data Warehouse Lifecycle Toolkit, Wiley, 1998

- audit, job logs and documentation — such as data lineage records, data transform logs
- DBMS metadata — such as
 - DBMS system table contents
 - processing hints
- front room metadata — such as
 - descriptions for columns
 - network security data
 - favorite web sites

Comments;

While this section has a heavy flavor of Data Warehouse it is important to note things that are not quantified, and measured are not easily managed. The Metadata Repository provides a place to inventory and manage all data within the enterprise.

Recommended Practices:

It is important to address the metadata repository to maximize the long-term value of data warehousing and to facilitate Enterprise Data sharing.

INF-RP-26: Enterprise metadata repository – Create and maintain an Enterprise level metadata repository.

Rationale:

- The repository contains metadata, or information about the data.
- The repository represents the shared understanding of the organization's data.
- The repository can be built incrementally, in stages, based on data warehouse design, application system design and implementation.
- The repository should support all types of data elements.
- Changes in the repository must occur before, and correspond to, the changes to data warehouse environments, and application systems.
- Metadata references many common data elements used by multiple application systems.
- If the repository databases provide for input from multiple ETL tool sets the metadata will easily transfer between repositories to facilitate the building of the Enterprise Information Model .
- There are two primary metadata repository products “Computer Associates Metadata Repository and ASG Software Solutions Inc. Rochade Repository”.

Data Cleansing

Data Cleansing tools provide for text cleaning, removal of duplicate data, column analysis, email checking, conversion of font case and are designed to save time. These cleansing tools support the removal of incorrect or unnecessary characters and data from a data source.

The quality of the data is acceptable if the data meet the following criteria:

- Are accurate
- Are stored according to data types
- Have integrity
- Are consistent
- Are in well designed databases
- Are not redundant
- Follow business rules
- Correspond to established domains
- Are timely
- Are integrated
- Satisfy the needs of the business
- Satisfy the user
- Are complete
- Do not contain duplicate records
- Do not contain data anomalies

Data Cleansing Considerations.

- Whether to cleanse the data in place or during conversion
- Disposition of the data which cannot be converted (e.g. “parentless” details)
- Written procedures for cleansing elements
- Roles and responsibilities when manual intervention is required

If the project involves a COTS package that is replacing an existing system, the team should investigate its impact on this deliverable by asking, at a minimum, the following questions:

- Does the COTS package have its own data cleansing and conversion rules? If so, do these rules adhere to Commonwealth standards?
- Will the project team need to create interim steps to match vendor conversion specifications?
- Will additional development tools be required for data cleansing and conversion?

Data Profiling

Data Profiling is a process whereby one examines the data available in an existing database and collects statistics and information about that data. Typical types of metadata sought are:

- Domain: whether or not the data in the column conforms to the defined values or range of values it is expected to take
 - for example: ages of children in kindergarten are expected to be between 4 and 5. An age of 7 would be considered out of domain
 - a code for flammable materials is expected to be A, B or C. A code of 3 would be considered out of domain.
- Type: Alphabetic or numeric
- pattern: a North American phone number should be (999)999-9999

- frequency counts: most of our customers should be in California; so the largest number of occurrences of state code should be CA
- Statistics:
 - minimum value
 - maximum value
 - mean value (average)
 - median value
 - modal value
 - standard deviation
- Interdependency:
 - within a table: the zip code field always depends on the country code
 - between tables: the customer number on an order should always appear in the customer table

Broadly speaking, most vendors who provide tools in the data profiling space divide the functionality into three categories. The names for these categories often differ depending on the vendor, but the overall process is in three steps, which must be executed in order:

- Column Profiling (Including the statistics and domain examples provided above)
- Dependency Profiling, which identifies intra-table dependencies. Dependency profiling is related to the normalization of a data source, and addresses whether or not there are non-key attributes that determine or are dependent on other non-key attributes. The existence of transitive dependencies here may be evidence of second-normal form.
- Redundancy Profiling, which identifies overlapping values between tables. This is typically used to identify candidate foreign keys within tables, to validate attributes that should be foreign keys (but that may not have constraints to enforce integrity), and to identify other areas of data redundancy. Example: redundancy analysis could provide the analyst with the fact that the ZIP field in table A contained the same values as the ZIP_CODE field in table B, 80% of the time.

Column profiling provides critical metadata which is required in order to perform dependency profiling, and as such, must be executed before dependency profiling. Similarly, dependency profiling must be performed before redundancy profiling. While the output of previous steps may not be interesting to an analyst depending on his or her purpose, the analyst will most likely be obliged to move through these steps anyway.⁵

Enterprise Data

The data and information to be shared among the various organizational entities of the Commonwealth to effectively manage its business as a whole, including:

- Local and Enterprise Structured data generated by both internal and external sources that is incorporated into State's operational data stores;

⁵ Data profiling. (2005, July 9). *Wikipedia, The Free Encyclopedia*. Retrieved 16:29, December 9, 2005 from <http://en.wikipedia.org/>.

- Unstructured hard-copy data, as well as data in electronic formats, such as audio, video, graphic, image, text.

Data maintained in support of an agency's operation are considered enterprise data if they meet any of the following criteria:

- If another agency uses the data and considers it essential;
- If integration of related information requires the data;
- If the Commonwealth must ensure the integrity of the data to comply with legal and administrative requirements; or if the Commonwealth needs the data for planning purposes.

Federated Data

A **federated database system** is a type of meta-database management system (DBMS) which transparently integrates multiple autonomous database systems into a single federated database. The constituent databases are interconnected via computer network, and may be geographically decentralized. Since the constituent database systems remain autonomous, a federated database system is a contrastable alternative to the (sometimes daunting) task of merging together several disparate databases. Through data abstraction, federated database systems can provide a uniform front-end user interface, enabling users to store and retrieve data in multiple databases with a single query--even if the constituent databases are heterogeneous. To this end, a federated database system must be able to deconstruct the query into sub queries for submission to the relevant constituent DBMS's, after which the system must composite the result sets of the sub queries. Because various database management systems employ different query languages, federated database systems can apply wrappers to the sub queries to translate them into the appropriate query languages.⁶

⁶ Federated database system. (2005, June 28). *Wikipedia, The Free Encyclopedia*. Retrieved 20:42, December 12, 2005 from <http://en.wikipedia.org/>.

Business Intelligence

Business intelligence (BI) is a broad category of application programs and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions.

The Business Intelligence topic within the Information Domain includes the components of Data Warehouse / Data Mart, Operational Data Stores, Extraction, Transformation and Loading (ETL), Data Storage Structures, Data Mining, Demand Forecasting and Management, Balanced Scorecard, Decision Support and Planning, Business Analytics Suites, and Dashboards.

Organizations typically gather such information in order to assess the business environment, and cover fields such as marketing research, industry or market research, and competitor analysis. Competitive organizations accumulate business intelligence in order to gain sustainable competitive advantage, and may regard such intelligence as a valuable core competence in some instances.

Persons involved in business intelligence processes may use application software and other technologies to gather, store, analyze, and provide access to data (also known as *business intelligence*). Some observers regard BI as the process of enhancing data into information and then into knowledge. The software aims to help people make "better" business decisions by making accurate, current, and relevant information available to them when they need it.

Generally, BI-collectors glean their primary information from internal business sources. Such sources help decision-makers understand how well they have performed. Secondary sources of information include customer needs, customer decision-making processes, the competition and competitive pressures, conditions in relevant industries, and general economic, technological, and cultural trends.⁷

Data Warehouse / Data Marts

A data warehouse is a database designed to support decision-making in an organization or enterprise. It is refreshed, or batch updated, and can contain massive amounts of data. When the database is organized for one department or function, it is often called a "data mart" rather than a data warehouse. The data in a data warehouse is typically historical and static in nature. A data warehouse is something you do, not something you buy. A successful data warehouse does not have an end. Regardless of the methodology, warehousing environments must be built incrementally through projects that are managed under the umbrella of a data warehousing program. Most of the benefits of a data warehouse or data mart will not be realized in the first delivery. The first project will be the foundation for the next, which will in turn form the foundation for the next. Data

⁷ Business intelligence. *Wikipedia, The Free Encyclopedia*. Retrieved 15:43, December 27, 2005 from <http://en.wikipedia.org>.

warehousing at the enterprise level is a long-term strategy, not a short-term fix. Its cost and value should be evaluated across a time span sufficient to provide a realistic picture of its cost-to-value ratio. Many vendors and consultants respond to agency data warehousing needs by offering “data warehousing in a box”, or a data mart as their complete end-to-end solution. These pre-packaged data mart solutions sound like an easy way to bypass all the tough issues surrounding the design and integration of a data warehouse, but they are not. Integration of data is not something that can be pre-packaged.

From the enterprise perspective, the primary value of the data warehouse comes directly from the integration of data. A standalone data mart in each department or agency may independently meet some organizational needs, but it will not resolve problems that result from creating proprietary islands of information. A data mart is not a solution unto itself; it is a component of the overall data warehousing architecture. In order to enable data marts to contribute to the enterprise perspective they must utilize “conforming dimensions.” Conforming dimensions are attributes of data which occur within multiple data marts and are standardized across all occurrences. By implementing conforming dimensions, one can report across multiple data marts using a common context (e.g. geography, time, etc.).

A data warehouse or data mart is not special technology in itself. Each consists of a database component which is a relational data structure optimized for reporting and analysis. It collects and stores integrated sets of historical, non-volatile data from multiple operational systems and feeds them into one or more databases. Data warehouses and data marts may also contain numerous levels of summarized or aggregated data. The database component is structured to support a variety of elaborate analytical queries on large amounts of data that can require extensive searching.

Recommended Practices:

Accelerated decision making requires high quality data. If operational data has changed or additional data is needed, changes must be made in the information model and in the data warehouse itself.

The data stored in a data warehouse should conform to the information model.

INF-RP-27: Validity Audits – Perform periodic validity audits against the data warehouse information model to ensure a high level of confidence in the quality and integrity of the data.

Rationale:

- The source data populating a data warehouse should be verified for consistency and accuracy.
- The data warehouse should correspond to business needs.
- Ensuring the integrity and quality of data is the responsibility of both the business users and IS.

INF-RP-28: Address business needs – Data Warehouse (DW) implementations should start with a plan that addresses critical business needs.

Rationale:

- Start with strategic business needs.
- Optimize critical data access before less-critical data access.
- Optimize high-volume data access before low-volume data access.
- Optimize applications and data with high-service-level requirements before those with lower requirements.

INF-RP-29: Impact on Network – Consider the impact of network capabilities when designing Business Intelligence systems.

Rationale:

- The network is a partner to data marts and can impact performance and how the data mart is designed.
- Estimate the impact on the network when designing data marts.
- Network connectivity to disparate source information systems and available bandwidth are critical components of most Extraction, Transformation and Load (ETL) operations.

INF-RP-30: Plan ETL – During data warehouse design phase, determine the logic needed to load and convert data – plan and generate the extraction and transformation routines.

Rationale:

- Planning for data extraction and transformation should start at the same time the data warehouse design starts.
- Data extraction and transformation is an important process for populating the data in a data warehouse and for ensuring that the data in a data warehouse is accurate.
- The data warehouse will contain data from disparate operational data sources.
- Data extraction and transformation logic includes data conversion requirements and the flow of data from the source operational database to the data warehouse or data mart.

INF-RP-31: Choosing PM – Choose a data warehousing project manager who is both business process oriented as well as technology oriented.

Rationale:

- A business process oriented manager ensures that the data warehouse will meet the business needs of the end users. This also mitigates the risk of

placing too much emphasis on technology and insufficient focus on business requirements.

- The data warehousing project manager must manage the expectations and sponsorship of the data warehouse.
- The data warehousing manager can make sure the data is easy to use and understand.

INF-RP-32: Ensure quality of source – Assess the source data that will populate a data warehouse to ensure accuracy, quality, and veracity.

Rationale:

- Data needs to be accurate to ensure good business decisions.
- Data needs to be relevant to the business need and consistent across multiple sources.
- Data must be complete. It must contain the information necessary to answer the data warehouse business need.
- The data assessment also involves evaluating the business rules associated with that data. The appropriate business rules must be applied to the data to maintain accuracy.

INF-RP-33: Scalability – When implementing data mart solutions it is critical to future scalability and integration that all data mart solutions implement conforming dimensions (standardized versions of dimensions which are common to more than one Data Mart).

Rationale:

As a data warehouse grows in its number of data sources, conforming dimensions are pivotal in reporting which spans data sources.

INF-RP-34: Refresh Schedule – Identify specific requirements for data availability, freshness and recoverability.

Rationale:

- Some data in the data warehouse needs to be refreshed more frequently than others. If the original data is fairly stable and not as volatile, the data warehouse may only need daily, weekly, or even monthly loads. When the original data is frequently changing or is more volatile, it may be necessary to consider an Operational Data Store (ODS) as an alternative
- Evaluate the impact of data extraction to any OLTP systems accessed.
- Business process requirements should determine frequency of extracts.

INF-RP-35: Directing analytical queries – Direct all analytical queries against a dimensionally modeled data warehouse or data mart, not an OLTP databases.

Rationale:

- Data marts contain data that has been checked for consistency and integrity, and represent a cross functional view of the data.
- OLTP transactions should not interact with a data warehouse or data marts
- Separating end-user requests and OLTP maximizes the efficiency of both environments.
- Growth in OLTP is incremental, and requirements are predictable.
- Growth in data warehouses and end-user computing has been nonlinear, and requirements are very difficult to predict.
- Fosters and supports the concept of data stewardship.

Requirements:

INF-R-13: Read-only Data Warehouse – Access shall be restricted to read-only for end users of the data warehouse.

Rationale:

Updates should only occur through the OLTP source where the data originates.

INF-R-14: Database Standard – Data warehouses and data marts that use relational databases shall conform to all of the Requirements and Technology Product Standards for databases as defined in the Enterprise Technical Architecture Database Domain report.

Rationale:

Relational databases used for data warehousing or data marts should conform to the same standards for all databases used by the Commonwealth.

To ensure that data warehouse and data mart implementations are built to meet the current and future business needs of an agency, executive sponsorship and representation by the business community on the project is required. Without this leadership, business intelligence (BI) projects run the risk of not providing the anticipated rewards or even failing altogether.

INF-R-15: Business community representation – A representative of the business community shall be involved in the entire development life cycle of all BI projects.

Rationale:

This is paramount to ensure project objectives will meet user expectations.

INF-R-16: Executive sponsorship – Project sponsorship shall be obtained from one or more executives within the upper management of the related organization prior to initiating any Data Mart or Data Warehouse project.

Rationale:

Without this sponsorship any such project is far more likely to fail.

Operational Data Stores

The Operation Data Store (ODS) is a database that consolidates data from multiple source systems and provides a near real-time, integrated view of volatile, current data. An ODS differs from a warehouse in that the ODS's contents are updated in the course of business, whereas a data warehouse contains static data. Operational Data Stores are usually driven by a business need to perform faster or more flexible reporting on their operational (transaction) data. Operational Data Stores are also considered a good source of input into a data warehouse because someone has already gone through the hard work of identifying and extracting data from the various legacy systems.

Extraction, Transformation and Loading

Defines the set of capabilities that support the manipulation and change of data and which support the population of a data source with external data. Data Extraction-Transformation-Load (ETL) tools are used to extract data from data sources, cleanse the data, perform data transformations, and load the target data warehouse and then again to load the data marts. The ETL tool is also used to generate and maintain a central metadata repository and support data warehouse administration. The more robust ETL tools integrate with OLAP tools, data modeling tools and data cleansing tools at the metadata level.

Transforming data is generally performed as part of the preparation before data is loaded into the data warehouse and data marts. Understanding the business usage of this information and the specific business questions to be analyzed and answered are the keys to determining the transformations necessary to produce the target data mart. ETL tools are used to extract data from operational and external source systems, transform the data, and load the transformed data in a data warehouse. The same tool is used to extract and transform the data from the warehouse and distribute it to the data marts. When a schedule is defined for refreshing the data, the data extraction and transformation schedule must be carefully implemented so that it both meets the needs of the data warehouse and does not adversely impact the source systems that store the original data.

Extraction is a means of replicating data through a process of selection from one or more source databases. Extraction may or may not employ some form of transformation. Data extraction can be accomplished through custom-developed programs. However, the preferred method uses vendor-supported data extraction and transformation tools that can

be customized to address particular extraction and transformation needs as well as use an enterprise metadata repository which will document the business rules used to determine what data was extracted from the source systems.

Transformation of transaction level data into the data warehouse often involves several techniques: filtering, summarizing, merging, transposing, converting and deriving new values through mathematical and logical formulas. These all operate on one or more discrete data fields to produce a target result having more meaning from a decision support perspective than the source data. This process requires understanding the business focus, the information needs and the currently available sources.

Cleansing data is based on the principle of populating the data warehouse with quality data -- that is, data that is consistent, is of a known, recognized value and conforms to the business definition as expressed by the user. The cleansing operation is focused on determining those values which violate these rules and, either reject or, through a transformation process, bring the data into conformance. Data cleansing standardizes data according to specifically defined rules, eliminates redundancy to increase data query accuracy, reduces the cost associated with inaccurate, incomplete and redundant data, and reduces the risk of invalid decisions made against incorrect data.

The availability of products such as Oracle Warehouse Builder, IBM Warehouse Manager, and Microsoft DTS has definitively changed the landscape of the ETL market, providing a strong push to SQL as the industry standard language for ETL. Those offerings suggest that RDBMSs and SQL have the power to perform any type of ETL process and that SQL code generators are going to be the foundation for coming generations of ETL software solutions.

Recommended Practices:

INF-RP-36: Data quality – Ensure data entry quality is built into new and existing application systems. This will reduce the risk of inaccurate or misleading data in OLTP systems and reduce the need for data hygiene.

Rationale:

- This will reduce the risk of inaccurate or misleading data in OLTP systems and reduce the need for data cleansing.
- The system should be designed to reject invalid data elements and to assist the end user in correcting the entry.
- All updates to an authoritative source OLTP database should occur using the business rules that own the data, not by direct access to the database.
- Enforcement of data entry quality by OLTP systems is critical. If data related to business needs is not entered, or is entered incorrectly, then it is of little value when reported.

INF-RP-37: ETL documentation – Document data extraction and transformation information in the metadata repository.

Rationale:

Data extraction and transformation are important aspects of a data warehouse. This provides the information map connecting the data populating a data warehouse with its source operational databases.

Data Storage Structures

Relational On-Line Analytical Processing (ROLAP) tools extract analytical data from traditional relational databases structures. Using complex SQL statements against relational tables, ROLAP is able to create multidimensional views on the fly. ROLAP tends to be used on data that has a large number of attributes, where it cannot be easily placed into a cube structure.

Multidimensional On-Line Analytical Processing (MOLAP) is specially designed for the purpose of user understandability and high performance. A multi-dimensional database uses a dimensional model instead of a relational model. A dimensional model is a star schema characterized by a central ‘fact’ table. One fact table is surrounded by a series of ‘dimension’ tables. Data is joined from the dimension points to the center, providing a so-called “star”. The fact table contains all the pointers to its descriptive dimension tables plus a set of measurements of facts about this combination of dimensions.

Hybrid On-Line Analytical Processing (HOLAP) tools use the best features of multidimensional and relational databases. Relational databases are best known for their flexibility. Until recently relational databases were weak in their ability to perform the same kind of multidimensional analysis that the multidimensional databases are specifically optimized for. The introduction of hybrid relational systems with enhanced abilities to manipulate star schemas has increased the OLAP capabilities to the relational world. Hybrid tools provide high performance for both general-purpose end users and power users.

In a multidimensional database, a dimensional model is a Cube. It holds data more like a 3-D spreadsheet rather than a traditional relational database. A cube allows different views of the data to be quickly displayed. The ability to quickly switch between one slice of data and another allows users to analyze their information in smaller meaningful chunks, at the speed of thought. Use of cubes allows the user to look at data in several dimensions; for example, attendance by Agency, attendance by attendance codes and attendance by date, etc. Use of a cube can be a far better solution than reviewing a giant report that can be confusing and contains unnecessary additional information or is formatted in a manner that requires additional manual thought or reorganization.

Data Mining

Data Mining tools are a class of products that apply artificial intelligence techniques to the analysis of data. They are about making predictions, not navigating through the data. The promise of data mining tools is that given access to the raw data, the tool can dip through the data looking for patterns and discovering relationships that the user might never have suspected.

Demand Forecasting and Management

Tools and procedures to facilitate the prediction of sufficient production to meet an agency's sales or delivery of a product or service.

Balanced Scorecard

Tools and procedures to support the listing and analyzing of both positive and negative impacts associated with a decision. One example of a balanced scorecard software solution is **Microsoft Office Business Scorecard Manager 2005**:

Microsoft® Office Business Scorecard Manager 2005 is the name of Microsoft Corp.'s new server-based business performance management scorecard application. The new application leverages the power of the Microsoft Office System to help organizations build, manage and use scorecards and key performance indicators (KPIs), and perform deep analysis — all within an intuitive, collaborative environment. Microsoft Office Business Scorecard Manager 2005 gives companies a broad view of business opportunities, through which employees can better understand business challenges, effectively shape solutions and quickly execute on their objectives. Decision-making happens across all levels of an organization, and empowering employees with powerful business intelligence (BI) solutions can help them have greater impact.⁸

Decision Support and Planning

The decision support and planning component includes tools and procedures to support the analysis of information and predict the impact of decisions before they are made.

Business Analytics Suites

Business analytics is a term used for more sophisticated forms of business data analysis.

Analytics closely resembles statistical analysis and data mining, but tends to be based on physics modeling involving extensive computation.

⁸ Microsoft PressPass – Information for Journalist. Retrieved, December 12, 2005 from <http://www.microsoft.com/presspass/newsroom/office/factsheets/07-19BSM2005FS.msp>

Example: A common application of business analytics is portfolio analysis. In this [example], a bank or lending agency has a collection of accounts, some from wealthy people, some from middle class people, and some from poor people. The question is how to evaluate the whole portfolio. The bank can make money by lending to wealthy people, but there are only so many wealthy people. The bank can make more money by also lending to middle class people. The bank can make even more money by lending to poor people. Note that poorer people are usually at greater risk of default. Note too, that some poor people are excellent borrowers. Note too, that a few poor people may eventually become rich, and will reward the bank for loyalty. The bank wants to maximize its income, while minimizing its risk, which makes the portfolio hard to understand. The analytics solution may combine time series analysis, with many other issues in order to make decisions on when to lend money to these different borrower segments, or decisions on the interest rate charged to members of a portfolio segment to cover any losses among members in that segment.⁹

Dashboards

A dashboard (or cockpit or monitor) displays key performance indicators on a screen or report, enabling them to be examined at a glance, before drilling down to detail in a business intelligence (BI) tool. A scorecard assumes a more structured approach and framework than a dashboard, making use of a methodology such as BSC, EFQM, value-based management or Six Sigma.¹⁰

⁹ Analytics. (2005, November 2). *Wikipedia, The Free Encyclopedia*. Retrieved 21:36, December 12, 2005 from <http://en.wikipedia.org/>.

¹⁰ Gartner Group, Management Update: Just Give Me a CPM Dashboard (15 June 2005), Frank Buytendijk and Bill Gassman.

Knowledge Management

Knowledge Management seeks to make the best use of the knowledge that is available to an organization, creating new knowledge, increasing awareness and understanding in the process. Knowledge Management can also be defined as the capturing, organizing, and storing of knowledge and experiences of individual workers and groups within an organization and making this information available to others in the organization.

The Knowledge Management (KM) topic within the Information Domain includes the components of Information Retrieval, Information Mapping/Taxonomy, Information Sharing, Categorization, Knowledge Engineering, Knowledge Capture, Knowledge Discovery, and Knowledge Distribution and Delivery.

Although an important topic within the Information Domain, KM will not be addressed in this report any further than identifying and defining the basic key components as noted below:

Information Retrieval

Information retrieval defines the set of capabilities that allow access to data and information for use by an organization and its stakeholders.

Information Mapping/Taxonomy

Information mapping/taxonomy defines the set of capabilities that support the creation and maintenance of relationships between data entities, naming standards and categorization.

Information Sharing

Information sharing defines the set of capabilities that support the use of documents and data in a multi-user environment for use by an organization and its stakeholders.

Categorization

Categorization defines the set of capabilities that allow classification of data and information into specific layers or types to support an organization.

Knowledge Engineering

Knowledge engineering defines the set of capabilities that support the translation of knowledge from an expert into the knowledge base of an expert system.

Knowledge Capture

Knowledge capture defines the set of capabilities that facilitate collection of data and information.

Knowledge Discovery

Knowledge discovery defines the set of capabilities that facilitate the identification of useful information from data.

Knowledge Distribution and Delivery

Knowledge distribution and delivery defines the set of capabilities that support the transfer of knowledge to the end customer.

Glossary

Following are Glossary entries pertaining to the Information Domain and required to support this document. Additional glossary definitions can be found in the ITRM Technology Management Glossary located on the VITA website here:

<http://www.vita.virginia.gov/projects/cpm/glossary.cfm>.

Some useful public glossaries can also be found at:

Wikipedia, the free encyclopedia at http://en.wikipedia.org/wiki/Main_Page

Loosely Coupled Glossary at <http://looselycoupled.com/glossary/azindex.html>

Another excellent glossary can be found at: <http://www.matisse.net/files/glossary.html>

Agency	Any agency, institution, board, bureau, commission, council, or instrumentality of state government in the executive branch listed in the appropriation act. ETA requirements/standards identified in this report are applicable to all agencies including the administrative functions (does not include instructional or research functions) of institutions of higher education, unless exempted by language contained in a specific requirement/standard.
Component	A readily accessible and observable aspect of a technology topic, such as Ad-hoc is a component of the Reporting topic and Data Warehouse is part of the Business Intelligence topic. A component is not the individual pieces such as tables, SQL scripts, etc. and other many similar pieces which make up the component.
Data	Data is the plural of <i>datum</i> . A datum is a <i>statement accepted at face value</i> (a "given") or a large class of practically important statements that are measurements or observations of a variable. Such statements may comprise numbers, words, or images. ¹¹
Domain	The Enterprise Technical Architecture (ETA) is typically divided into logical groups of related technologies and components, referred to as "domains". The purpose of a Domain Architecture is to provide a combination of domain principles, best practices, reusable methods, products, and configurations that represent "reusable building blocks". Thus, the Domain Architecture provides the technical components within the Enterprise Architecture that enable the business strategies and functions. Note, the Conceptual Architecture serves as the foundation for the Domain Architectures, and ensures that they are aligned and compatible with one another. ¹²

¹¹ Data. *Wikipedia, The Free Encyclopedia*. Retrieved 20:38, January 25, 2006 from <http://en.wikipedia.org>

¹² COTS Enterprise Architecture Workgroup, "Commonwealth of Virginia Enterprise Architecture – Common Requirements Vision", v1.1, December 5, 2000, p 26.

Enterprise	As used in this document and generally when discussing Enterprise Architecture topics, the <i>enterprise</i> consist of all Commonwealth of Virginia agencies as defined above.
Information	Too often the words information and data are used interchangeably which leads to confusion. Data is unstructured, lacks context and may not be relevant to the recipient. When data is correctly organized, filtered and presented with context it can become information because it then has "value" to the recipient. ¹³
ORCA	<u>O</u> nline <u>R</u> eview and <u>C</u> omment <u>A</u> pplication is a web based application managed by VITA to allow public comment and review of proposed policies, standards, and guidelines. ORCA may be accessed through the Commonwealth Project Management Web page or by pointing your Web browser to the URL http://apps.vita.virginia.gov/publicORCA .
Principles	High-level fundamental truths, ideas or concepts that frame and contribute to the understanding of the Enterprise Architecture. They are derived from best practices that have been assessed for appropriateness to the Commonwealth Enterprise Architecture. ¹⁴
Product Standards	Are specifications for the use of specific hardware and software relative to the particular component.
Recommended Practices	Are activities which are normally considered leading edge or exceptional models for others to follow. They have been proven to be successful and sustainable and can be readily adopted by agencies. They may or may not be considered the ultimate “best practice” by all readers but for this place and time they are recommended practices and should be used and implemented wherever possible.
Requirements	Are activities that are considered strategic components of the Commonwealth’s Enterprise Technical Architecture. They are acceptable activities for current deployments and must be implemented and used for all future deployments.
Topic	A topic is simply a logical subdivision of the domain. All components relevant to the Commonwealth’s Technical Architecture are included within one of the identified topics. Within the Information domain topics include Reporting, Data Management, Business Intelligence and Knowledge Management.

¹³ Information is not data. *Wikipedia, The Free Encyclopedia*. Retrieved 21:33, January 25, 2006 from <http://en.wikipedia.org>.

¹⁴ COTS Enterprise Architecture Workgroup, “*Commonwealth of Virginia Enterprise Architecture – Conceptual Architecture*”, v1.0, February 15, 2001, p 5.

Appendices

References and Links

State and Federal Sites:

The domain team would like to publicly thank their counterparts in the many states and federal government agencies whose excellent work preceded this. We couldn't have completed this report as quickly as it was done without the tireless energies obviously expended to complete their Enterprise Architecture documents. We also hope that other states will find this document useful in the design and updating of their own Enterprise Architecture. Significant contributions, references, and insights were derived from the following documents and web sites.

Housing and Urban Development:

<http://www.hud.gov/offices/cio/ea/newea/index.cfm>

Connecticut:

Data Management and Data Warehouse Domain Technical Architecture :

http://www.ct.gov/doit/lib/doit/DATA_ARCHITECTURE_ver_20_6-6-2002.pdf

Application Development Domain Technical Architecture:

http://www.ct.gov/doit/lib/doit/Application_Architecture_5-8-2003_ver_2-5.pdf

Massachusetts:

ETRM Version 3.5 Information Domain:

http://www.mass.gov/Aitd/docs/policies_standards/etrm3dot5/etrmv3dot5informationdomain.pdf

ETRM Version 3.5 Application Domain:

http://www.mass.gov/Aitd/docs/policies_standards/etrm3dot5/etrmv3dot5applicationdomain.pdf

North Carolina:

Statewide Technical Architecture - Data Domain

<http://www.ncsta.gov/docs/Principles%20Practices%20Standards/Data.pdf>

Pennsylvania:

Database Management Systems: Production and Operational Standards: February 23, 2005

http://www.oit.state.pa.us/oaait/lib/oaait/STD_INF001B.doc

General Information References:

Gartner Group:

<http://www.gartner.com/>

The Open Group:

<http://www.opengroup.org/architecture/togaf8-doc/arch/toc.htm>

The Data Warehouse Institute:

<http://www.tdwi.org>

Microsoft:

http://www.microsoft.com/resources/sam/Implementing_Policy.mspx