

Comprehensive Large Array-data Stewardship System (CLASS)

Information Heterogeneity White Paper

November 2008



Prepared by:

**U.S. Department of Commerce
National Oceanic and Atmospheric Administration (NOAA)
National Environmental Satellite, Data, and Information Service (NESDIS)**

**Comprehensive Large Array-data
Stewardship System (CLASS)
Information Heterogeneity White Paper**

November 2008

Prepared by:

**U.S. Department of Commerce
National Oceanic and Atmospheric Administration (NOAA)
National Environmental Satellite, Data, and Information Service (NESDIS)**

Approval Page

Document Numbers:	
NOAA/NESDIS CLASS-1215-CLS-WHT-IHET	November 26, 2008 CLASS Information Heterogeneity White Paper
<p>Document Title Block:</p> <h2 style="margin: 0;">Comprehensive Large Array-data Stewardship System (CLASS)</h2> <h3 style="margin: 0;">Information Heterogeneity White Paper</h3>	
PROGRAM: CLASS	DOCUMENT RELEASE DATE: November 26, 2008
APPROVALS	
GROUP: CLASS Date	GROUP: CLASS Date
NAME: Rick Vizbulis, Project Manager	NAME: Khalid Alvi, COTR
GROUP: CLASS Date	GROUP: CLASS Date
NAME: Joe Hennessey, DGP Contract Project Manager	(If needed) NAME: Name, position
QMO REVIEW BY: Date	
NAME: Rae Ann Withers, CLASS QMO Manager	

Preface

This document comprises the initial National Oceanic and Atmospheric Administration (NOAA) National Environmental Satellite, Data, and Information Service (NESDIS) baseline publication of the Comprehensive Large Array-data Stewardship System (CLASS) Information Heterogeneity White Paper.

This document provides an introduction to information heterogeneity, discusses its impacts across NOAA, and offers recommendations regarding its reduction and management.

This version is the initial baseline document. Go to the CLASS Document Repository Web link @ <https://kt.nsof.class.noaa.gov> for the current version.

Table of Contents

Section 1.0	Introduction	1
1.1	Purpose	1
1.2	Audience.....	2
1.3	Scope	2
1.4	Document Organization	2
1.5	References and Related Documents.....	3
1.5.1	References	3
1.5.2	Related Documents.....	4
1.6	Document Maintenance.....	4
1.7	Terminology.....	4
1.8	Acronyms	5
Section 2.0	Information	6
2.1	Information Roles.....	6
2.2	Information Types	7
Section 3.0	Information Heterogeneity	9
3.1	Causes of Information Heterogeneity.....	9
3.1.1	NOAA Domain/Mission.....	9
3.1.2	NOAA Organizational Structure and Management	9
3.1.3	Environmental Science.....	10
3.1.4	Observing Systems and Observed Phenomena	10
3.1.5	IT Systems.....	10
3.1.6	Functionality.....	11
3.1.7	Users and Requirements.....	11
3.1.8	Lack of Comprehensive Standards.....	11
3.1.9	Local Customization.....	11
3.1.10	Schedule and Budgetary Constraints.....	12
3.2	Information Heterogeneity Impacts.....	12
3.2.1	Complexity	13
3.2.2	Functionality.....	14
3.2.3	Interoperability	15
3.2.4	Cost.....	17
3.3	Increasing Data Volumes and Information Heterogeneity	21
Section 4.0	Information Heterogeneity Across NOAA.....	23
4.1	Information Management Systems and Information Heterogeneity	23
4.1.1	Interest in Information Heterogeneity	23
4.1.2	The Potential Problem: CLASS as a Case Study	26
4.2	Information Heterogeneity and Information Stewards.....	28
4.3	Information Heterogeneity and NOAA Archives	28
4.4	Information Heterogeneity and GEO-IDE	28
4.5	Information Heterogeneity and NOAA.....	29

4.5.1	Significance across the Organization	29
4.5.2	Information Heterogeneity and NOAA Customers.....	30
4.5.3	NOAA’s Role in Larger “Systems of Systems”.....	30
4.5.4	Summary	30
Section 5.0	Something Has to Give!	32
5.1	Limit Functionality.....	32
5.2	Increase Resources	32
5.3	Mitigate Heterogeneity.....	33
Section 6.0	Information Heterogeneity Impact Mitigation	34
6.1	First Steps.....	34
6.1.1	Acknowledge the Problem	34
6.1.2	Provide Organizational Support.....	34
6.1.3	Speak the Same Language.....	34
6.1.4	Think Incrementally	35
6.1.5	Carefully Allocate Roles and Responsibilities.....	35
6.1.6	Apply Different Approaches	35
6.1.7	Establish a Framework	35
6.2	Information Heterogeneity Reduction.....	36
6.2.1	Elimination of Unnecessary Information Types	36
6.2.2	Standardization.....	37
6.3	Information Heterogeneity Impact Management	39
6.3.1	Roles and Responsibilities.....	39
6.3.2	Comprehensive Analysis.....	40
6.3.3	Leverage Industry Trends and Approaches.....	40
6.3.4	Generic IT Mechanisms	40
6.3.5	Information-Specific Mechanisms	41
6.3.6	Standardization.....	41
6.4	Summary	42
Section 7.0	Recommendations	43
7.1	NOAA	43
7.1.1	Promote Institutional Awareness and Commit to Mitigation Efforts.....	43
7.1.2	Emphasize the Importance of Information Governance.....	44
7.1.3	Commission an Information Heterogeneity Cost/Benefit Study.....	44
7.1.4	Develop an End-to-End Information Lifecycle Reference Model	44
7.1.5	Aggressively Push Standards Adoption	45
7.1.6	Make Producers Part of the Solution.....	45
7.1.7	Support Information-Related Role and Responsibility Allocation	45
7.1.8	Promote Information Management “Meta-Initiatives”	45
7.1.9	Make Hard Decisions Regarding Information Retention.....	46
7.1.10	Coordinate or Consolidate Ongoing Efforts.....	46
7.1.11	Engage the National and International Communities.....	46
7.2	GEO-IDE.....	46
7.3	NOAA Archives.....	47
7.4	NOAA Information Stewards.....	47

7.4.1	Be a Key Element of Standardization Efforts	47
7.4.2	Push for Role and Responsibility Transitions	47
7.4.3	Support Mechanism Specification and Development	47
7.4.4	Serve as a Bridge between Producers and Consumers.....	48
7.4.5	Coordinate Their Efforts	48
7.4.6	Take Advantage of the Information Lifecycle	48
7.4.7	Take a Much Stronger Role in the IT Requirements Process	48
7.5	IT Systems.....	49
7.5.1	Publicize Information Heterogeneity and Its Impacts	49
7.5.2	Develop Generic IT Mechanisms.....	49
7.5.3	Actively Support and Push for Role and Responsibility Transitions.....	49
7.5.4	Rigorously Pursue and Implement Standardization	50
7.6	Summary	50
Section 8.0	Conclusion.....	51

Appendices

Appendix A. Acronyms	A-1
----------------------------	-----

List of Figures

Figure 1 – Complexity Function.....	13
Figure 2 – NOAA Archive Volumes, 2004-2020 [NCDC unpublished presentation].....	21

List of Tables

Table 1 – Information Roles.....	6
Table 2 – Selected Information Type Examples.....	7
Table 3 – NOAA “Types” Stored in CLASS.....	26

Section 1.0 Introduction

1.1 Purpose

This paper is intended to provide a basic understanding of information heterogeneity, the problems it causes, and the importance of both reducing information heterogeneity and managing its impacts.

Specifically, this paper is intended to:

- Define “information heterogeneity” and describe some of its types and causes.
- Consider some of the most serious implications of information heterogeneity.
- Illustrate ways of thinking about information heterogeneity in terms of the systems, processes, and organizations impacted by it.
- Show that information heterogeneity is a NOAA-wide problem, and that its impacts—and the benefits of addressing them—are increasingly dramatic at higher levels in the organization.
- Emphasize the substantial obstacle information heterogeneity presents to the success of many of NOAA’s information-related initiatives and that addressing information heterogeneity is, thus, crucial to the success of these initiatives.
- Illustrate that information heterogeneity is a primary cost driver for NOAA information management activities.
- Make it clear that NOAA’s success in meeting organizational information management expectations will correlate strongly with its success in addressing information heterogeneity.
- Point out that appropriate allocation of organizational roles and responsibilities is a crucial first step to addressing the problems caused by information heterogeneity.
- Describe the roles of various NOAA information management-related entities in information heterogeneity reduction and/or impact management.
- Highlight the ways in which information heterogeneity reduction and impact management can yield substantial benefits across NOAA.

While this paper is written primarily from CLASS’s point of view, it is not primarily about CLASS. Information heterogeneity impacts all of NOAA’s information management initiatives, systems, and organizations, and even its customers.

Awareness of information heterogeneity and its impacts varies across NOAA. Within some circles, the issue is well-known and the subject of considerable thought and effort. In others, however, it is either unknown, unaddressed, or largely an afterthought. This inconsistency is one of the primary impediments to addressing the problem and one of the key obstacles for agency-wide information management initiatives and strategic objectives.

Ideally, this paper would be written by the DMC. However, there is not yet a unified, agency-wide focus on information heterogeneity that would naturally lead to the development of this document. Until there is sufficient awareness of the problem and impetus to solve it, it falls to those most affected by its impacts to do what they can to try and place it in the spotlight.

While CLASS does not cause information heterogeneity and is not in a position to reduce it, it does feel many of its impacts and is in a position to see some of the ramifications for the rest of NOAA. The serious consequences of information heterogeneity for all NOAA information management efforts—in terms of system development, staffing, resource utilization, O&M costs, interoperability, information utilization, and many others—have made it necessary for CLASS to try and heighten awareness of information heterogeneity and push for reduction and impact mitigation efforts through the preparation of this paper.

1.2 Audience

The general audience for this document includes anyone with a role, or interest, in NOAA information management.

The specific audience for this document includes:

- NOAA management
- NESDIS management
- NOAA information management-related bodies, including the NOSC, DMC, DMIT, CIO Council, and NAAT
- NOAA IT system developers and owners
- NOAA information producers
- Other bodies with influence regarding NOAA policies, for example the DAARWG
- GEO-IDE technical and management staff
- CLASS project management
- The COPB
- Information stewards and their management
- Parties interested in the CLASS target system architecture and its drivers

1.3 Scope

This paper applies to all NOAA information management systems and activities that have to deal with information heterogeneity.

1.4 Document Organization

This document is organized as follows:

- Section 1.0, *Introduction*, describes the purpose of this document and presents some general information useful to reading and understanding it.
- Section 2.0, *Information*, defines “information” and discusses information roles and types.
- Section 3.0, *Information Heterogeneity*, presents the main topic of this paper, information heterogeneity.
- Section 4.0, *Information Heterogeneity across NOAA*, discusses information heterogeneity in the context of selected NOAA organizational components, and NOAA generally.
- Section 5.0, *Something Has To Give!*, discusses alternatives for dealing with complexity

increases associated with information heterogeneity.

- Section 6.0, *Information Heterogeneity Impact Mitigation*, discusses approaches to mitigating the effects of information heterogeneity.
- Section 7.0, *Recommendations*, provides specific information heterogeneity reduction and impact mitigation recommendations for NOAA information management-related entities.
- Section 8.0, *Conclusion*, presents a brief summary of this paper.
- Appendix A, *Acronyms*, provides expansions for the acronyms used in this document.

1.5 References and Related Documents

The following are either referred to directly in this paper, or may be useful in obtaining additional information.

1.5.1 References

- Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, Blue Book, January 2002, ISO 14721:2003, <http://public.ccsds.org/publications/archive/650x0b1.pdf>. [OAIS-RM]
- Hankin, S. and the DMAC Steering Committee, *Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems: I. Interoperable Data Discovery, Access, and Archive*, Ocean. US, 2005, http://dmac.ocean.us/dacsc/imp_plan.jsp. [IOOS DMAC]
- Lautenbacher, Conrad C., *About NOAA*, accessed September 2008. <http://www.noaa.gov/about-noaa.html> [About NOAA]
- NASA Geospatial Interoperability Office, *Geospatial Interoperability Return on Investment Study*, April 2005. [NASA ROI]
- NASA, *NASA's Global Change Master Directory*, accessed January 2008, <http://gcmd.nasa.gov/>.
- NIST, *99-1 Planning Report Interoperability Cost Analysis of the U.S. Automotive Supply Chain*, March, 1999, <http://www.nist.gov/director/prog-ofc/report99-1.pdf>. [NIST99-1]
- NIST, *GCR 04-867 Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry*, August 2004. [NIST2004]
- NOAA DMIT, *Review of Standards Applicable to NOAA and Recommendations for Fast-track Consideration as Proposed Standards*, v. 4.2, February 2006. [DMIT2006]
- NOAA, *Global Earth Observation Integrated Data Environment (GEO-IDE) Concept of Operation*, version 3.3, September 13, 2006. [GICO]
- NOAA, *National Oceanic and Atmospheric Administration Report to Congress on Data and Information Management 2007*, 2007. [NOAA2007]
- NOAA, *NOAA Administrative Order 212-15*, December 22, 2003, http://www.corporateservices.noaa.gov/~ames/NAOs/Chap_212/naos_212_15.html. [NAO 212-15]
- NOAA/NESDIS, *CLASS Archive and Access System Requirements, version 2.2*, May 16, 2005. [A&ARv2.2]
- NOAA/NESDIS, *CLASS System Level 1 Requirements Document and Mission Success Criteria*, Preliminary, version 1.14, May 28, 2008. [CLASS-L1R]
- NOSA Action Group, *CasaNOSA Web Portal*, accessed January 2008,

<https://casanosa.noaa.gov/> (requires login).

- USGS, *Geospatial One-Stop Web Portal*, accessed January 2008, <http://gos2.geodata.gov/>.
- Walker, Jan, Eric Pan, Douglas Johnston, Julia Adler-Milstein, David W. Bates, and Blackford Middleton, *The Value Of Health Care Information Exchange And Interoperability*, Health Affairs, 19 January 2005, DOI: 10.1377/hlthaff.w5.10. [Walker2005]

1.5.2 Related Documents

- ESA, *Socio-Economic Benefits Analysis of GMES*, ESA Contract Number 18868/05, October 2006.
- http://en.wikipedia.org/wiki/Semantic_interoperability provides a useful, accessible introduction to semantic interoperability and related issues.
- Lowry, Roy, *Vocabulary management: a foundation for semantic interoperability through ontology development*, presentation at the 1997 GO-ESSP conference.

1.6 Document Maintenance

This document is subject to review and approval by CLASS project management. It may be periodically updated during the course of the CLASS project. Any changes will be submitted to CLASS project management for approval.

1.7 Terminology

Following are selected terms and concepts essential to a full understanding of this document:

Archive: An organization of people and systems dedicated to information preservation as defined by the OAIS-RM. A NOAA archive is a virtual entity, composed of an information steward responsible for the information-specific aspects of archival, CLASS—whose role is the provision of supporting IT capabilities—and possibly other supporting systems and/or organizations.

Data: The means by which information is represented. For the purposes of this paper, data are sequences of bits.

Information: Knowledge in a form that can be exchanged. Information is a combination of data and the structural and semantic context necessary to make the data understandable and usable.

Information heterogeneity: Variations in the ways in which information roles or types are represented, interpreted, and/or used.

Information preservation: The act of preserving information over the long term, where “long term” is a period of sufficient duration to span technology changes.

Information role: An information classification based on the OAIS-RM information taxonomy. For example, content information, representation information, descriptive information, etc., are all information roles.

Information-specific: Pertaining exclusively to a particular information type or role. This term is generally used to describe those artifacts of systems, processes, or organizations that apply to a single information role or type.

Information specificity: The characteristic of being information-specific. This term is also used to describe the degree to which a system, process, or organization is exclusive to an information role or type.

Information steward: An entity with responsibility for managing information throughout the information's lifecycle. In the archive context, the information steward is responsible for the information-specific aspects of information preservation.

Information type: An information classification based on non-role-related characteristics, including representation details, origin, purpose, and many other attributes.

Metadata: *data about other data*. [OAIS-RM] Metadata describe, characterize, or otherwise document other information.

1.8 Acronyms

In some cases (lists, for example), in-line acronym expansion would severely impair the readability of the text. To avoid this problem, no in-line acronym expansions are provided in the body of this paper. Expansions for all acronyms are provided in Appendix A.

Section 2.0 Information

“Information” is “*knowledge in a form that can be exchanged.*” [OAIS-RM] Clearly, knowledge exchange requires the recipient to be able to understand the received information. Thus, semantics are an essential part of information. In fact, semantics are **the** essential aspect of information. The importance of semantics will be a consistent theme throughout this paper, and successfully managing semantics is one of the key challenges for NOAA in addressing the information heterogeneity problem.

In contrast, “data” refers specifically to the means by which information is represented. For the purposes of this paper, data are simply sequences of bits; they have no specific semantics. The structure of the bits must be interpreted, and semantics applied to the structural artifacts, in order for data to be interpreted as information.

Information can be classified in many ways, but it will suffice for the discussion that follows to distinguish information “roles” and “types.”

2.1 Information Roles

The following table summarizes information roles as identified in the OAIS-RM information model. While these are intended specifically for long-term information preservation, they can be applied to many other information management activities as well.

Table 1 – Information Roles

OAIS-RM Information Role	Description
Content information	Original target of preservation (content data object) plus the representation information necessary to make it understandable and usable
Representation information	Describes how to interpret a data object
Structural information	Describes the structure of a data object
Semantic information	Describes the semantics of a data object and its components
Descriptive information	Provides descriptions that support search activities
Preservation description information	Describes the preservation-related aspects of information
Provenance information	Documents information history
Fixity information	Ensures information authenticity
Reference information	Uniquely identifies information
Context information	Provides context for information
Packaging information	Binds and/or identifies information

The characteristics that differentiate these roles constitute a form of information heterogeneity; information fulfilling the different roles is used differently, is often represented and stored differently, and may well exhibit other differences that must be taken into account. However, these roles also comprise a key part of a framework within which other, much higher impact, forms of information heterogeneity can be managed. While variations across information roles do contribute to the overall level of information heterogeneity, the framework provided by the roles is an essential tool for helping manage the problem. Thus, it is important that information roles be acknowledged and understood, but the wide range of information types within each role is the primary concern of this paper.

2.2 Information Types

“Information types,” as used in this paper, are distinguished by differences between instances of information within a given information role.

For example, the role of descriptive information can be fulfilled by browse images, spatial coverages, environmental parameter names, and many others. The differences between browse images, spatial coverages, environmental parameter names, and the like, are instances of information heterogeneity, and serve to distinguish these as information types.

At the same time, information types within a particular role often vary across campaigns and sensors. For example, browse image representations for observations from a particular sensor often differ significantly from those of similar sensors on different platforms. DMSP OLS and POES AVHRR are similar instruments in many regards, but the information they record is represented in different ways; thus, browse images for each may differ significantly as well. This example highlights the information heterogeneity that serves to distinguish information types. It is categorically different than the heterogeneity across information roles, and is the most problematic kind of information heterogeneity.

The following table provides a few additional examples of the heterogeneity present within each OAIS-RM information role.

Table 2 – Selected Information Type Examples

OAIS-RM Information Role	Examples of NOAA Archive Information Types
Content information	AVHRR imagery, DMSP SSM/I brightness temperatures, CoastWatch sea surface temperature grids, GOES imagery, GOES soundings
Representation information	
Structural information	<ul style="list-style-type: none"> • File formats: XML, HDF, NetCDF, JPEG, GRiB, BUFR, blob • Record and field formats • Structural data types: grid, swath, radial, profile, trajectory, point, image, video, audio

OAIS-RM Information Role	Examples of NOAA Archive Information Types
Semantic information	<ul style="list-style-type: none"> • Terminology, definitions, parameter names, units, coordinate systems, time standards • Semantic standards: CF, COARDS
Descriptive information	Browse images, quality indicators, text descriptions, percent cloud cover, spatial coverages, temporal coverages
Preservation description information	
Provenance information	Processing steps applied, software versions used, transformations applied during preservation
Fixity information	Checksums, CRCs, message digests, digital signatures
Reference information	URLs, database keys, filenames, accession numbers
Context information	Source code, validation data, quality assessment reports, orbital attitude data, sensor calibration, text descriptions
Packaging information	Manifest files, directory structures, “tar” file tables of contents, HDF “wrappers” for other representations

It is important to note that the number of information types that may appear in certain information roles is, or may become, extremely large. For example, almost any information might provide context for or describe other information, thus creating the potential for almost unlimited information heterogeneity. Moreover, some information may fulfill multiple roles. Finally, it should be noted that the above are insignificant in number when compared to all examples of information heterogeneity present across NOAA. As will be discussed, some sources estimate that NOAA currently holds thousands of types of content information alone. The total number of discrete information types, however, is dramatically greater than this, as many information-specific kinds of metadata are associated with each content information type.

Section 3.0 Information Heterogeneity

Generally, “heterogeneity” refers to variations among items belonging to a common class. For the purposes of this paper, “heterogeneity” refers specifically to variations that must be taken into account by the systems, processes, and organizations in which the variations appear. Narrowing the definition in this way places the focus on “significant” variations, or those variations with identifiable and tangible impacts. It also emphasizes that some heterogeneity may be irrelevant; an understanding of both the variations and the context in which they appear is essential.

Heterogeneity appears throughout NOAA, in a remarkable number of forms and places. Among the most prominent of these is NOAA’s information holdings; for reasons that will be presented in the material that follows, NOAA environmental information exhibits extraordinary diversity.

“Information heterogeneity” refers to **information** variations that must be accounted for by systems, processes, and organizations. Specifically, it refers to differences in the ways various kinds of information are represented, interpreted, and used. In an organization like NOAA, a certain amount of information heterogeneity is both necessary and beneficial. Too much information heterogeneity, however, dramatically increases the resources required to manage information and greatly limits the functionality that can be provided.

“Information specificity” refers to system, process, or organizational customizations necessary to address information heterogeneity. Information specificity is the certain consequence of information heterogeneity, and it is in information specificity that the tangible ramifications of information heterogeneity are seen.

3.1 Causes of Information Heterogeneity

Information heterogeneity has a wide variety of causes. The following sub-sections discuss some of the most important causes of information heterogeneity within NOAA.

3.1.1 NOAA Domain/Mission

The breadth of NOAA’s domain, a realm that encompasses both oceans and atmosphere, engenders a great deal of diversity. Underscoring this point, its Web site notes that NOAA’s “*reach goes from the surface of the sun to the depths of the ocean floor.*” [About NOAA] NOAA’s information holdings span this domain and inherit much of the variety within it.

3.1.2 NOAA Organizational Structure and Management

The many areas of interest that naturally emerge from NOAA’s broad mission result in an “organization of organizations” where sub-organizations are often highly specific and exhibit a great deal of variety. These sub-organizations differ in a wide variety of ways, including the types of information they produce and/or consume.

Also, NOAA organizational elements, programs, and projects are often managed differently, leading to dissimilar (and sometimes conflicting) agendas, needs, and priorities. The information generated or used by these efforts often exhibits diversity as a result.

3.1.3 Environmental Science

Environmental science is a broad and diverse field that includes study of “...*the physical, chemical, biological, geological, or geophysical properties or conditions of the oceans, atmosphere, space environment, sun, and solid earth...*” [GICO].

NOAA Administrative Order 212-15 describes this diversity similarly, characterizing the domain to include “... *recorded observations and measurements of the physical, chemical, biological, geological, or geophysical properties or conditions of the oceans, atmosphere, space environment, sun, and solid earth, as well as correlative data and related documentation or metadata. Media, including voice recordings and photographs, may be included.*” [NAO 212-15]

This diversity propagates into the information gathered through and for environmental research.

It is important to note also that environmental science is constantly evolving, resulting in new methods for characterizing phenomena and allowing more complicated and detailed measurements to be taken. The evolutionary process introduces significant information heterogeneity as well.

3.1.4 Observing Systems and Observed Phenomena

NOAA operates observing systems that range from people counting tree rings to sophisticated satellite constellations that make and record observations of Earth and its environment in minute detail. Its wide variety of observing systems tends to introduce a great deal of heterogeneity to the information NOAA produces.

Moreover the observations themselves tend to exhibit great variety: “The measurements are highly heterogeneous, originating from surveys (e.g., fish stock assessments), cruise measurements, laboratory measurements, satellites, and automated inputs from in situ and remotely sensed sources ...” [IOOS DMAC]. Heterogeneous measurements are, for obvious reasons, a direct contributor to information heterogeneity.

3.1.5 IT Systems

Computer hardware, operating systems, programming languages, interfaces, protocols, and application software all exhibit significant variety. This diversity is often reflected in different ways of representing, storing, and using information.

Moreover, information technology—like environmental science—is constantly evolving. This process leads to a profusion of file formats, progressively more sophisticated encoding methods, and a wide variety of approaches to data and information management. These all contribute to information heterogeneity, and their impacts compound over time.

3.1.6 Functionality

Different information types often require different operations. Even in cases where operations are semantically similar, variations in their implementation and execution are often present across information types, or even within a type. The impact of these functional variations is substantial, particularly since functional requirements are increasing at the same time that NOAA's information holdings are both growing and becoming more diverse.

As with some of the other sources of information heterogeneity, functionality is not static; new types of functionality continually emerge and new types of information are necessary to support them. For example, robust spatial capabilities require certain types of spatial information that differ categorically from NOAA's environmental observations, specifically: "*... information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the Earth. This information may be derived from, among other things, remote sensing, mapping, and surveying technologies. Statistical data may be included ...*" [NAO 212-15]. The types of information necessary to support new functionality often add to information heterogeneity.

3.1.7 Users and Requirements

Users of NOAA services and information encompass a broad variety of interests, needs, institutional associations, usage patterns, and technical sophistication. As is easily imagined, this results in an extraordinary diversity of information needs. Moreover, users tend to use different tools and to apply information to different problems. The result is wide variation in the types of information needed by different user communities.

3.1.8 Lack of Comprehensive Standards

An ongoing lack of inclusive, pervasive standardization efforts contributes significantly to information heterogeneity. Low levels of structural standardization (e.g., file formats) and semantic standardization (e.g., terminology, units) are substantial contributors. Lack of standardization can have many causes, most of which can be traced back to the point at which information was created. These include: 1) unavailability of appropriate or applicable standards; 2) failure to recognize the need for, or benefits of, the application of standards; 3) lack of organizational backing or mandates for the use of standards; and 4) the perception that standards application would slow current efforts. The last example is particularly troublesome; while expedience is compelling in the short term, failure to standardize leads to cost and complexity increases that, over the long term, are overwhelming for an organization like NOAA.

3.1.9 Local Customization

Customizing systems and products for specific purposes or communities causes a considerable amount of information heterogeneity. "One size fits all" solutions often do not exist, and it is frequently necessary to tailor information in highly specific ways. While information heterogeneity is the desired goal in cases like this, it is also a very costly achievement when considered across the

organization and over the long term.

3.1.10 Schedule and Budgetary Constraints

Schedule constraints and the need for expedience often cause information heterogeneity. Standards can be perceived as “heavyweight” and time-consuming to understand and implement. It is often argued that it is easier and faster (and therefore, less expensive) to get things done if problems are simplified and the involvement of others is minimized. Moreover, mitigating the impacts of information heterogeneity—as will be shown in the material that follows—requires significant effort, often involves cooperation with others, and really needs to be part of an overall organizational initiative. Thus, it is often both expedient and appealing for information management efforts to sidestep the information heterogeneity issue altogether.

While these arguments are frequently true in the short term, the cost of accepting them is overwhelming over longer durations, particularly across a broad and complicated organization like NOAA. Standardization and other mitigation approaches can be made more efficient over time. Organizations and people learn to think and work in terms of information heterogeneity management, and it becomes second nature. The benefits of these progressions are enormous and the cost, relative to the ever-escalating expense associated with information heterogeneity and its consequences, is negligible.

One of the key points in the preceding sub-sections is that information heterogeneity results from active choices. Often, these choices are made carefully and thoughtfully, after significant analysis and consideration. Unfortunately, these decisions are often made within a very narrow context, and give limited weight to long-term organizational impacts.

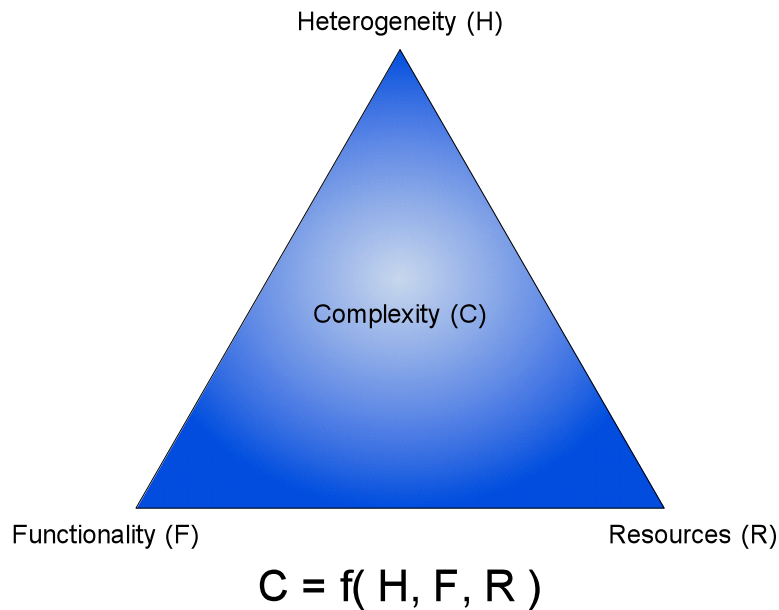
Of course, the short term must sometimes be prioritized over the long. Nonetheless, decisions regarding information heterogeneity must be made with full recognition of the ramifications of the various alternatives. Recognition and understanding of the actual costs and organizational impacts of these choices are essential ingredients to making choices that serve local and/or short-term needs, remain economical over time, and support overall organizational needs. Fortunately, the impacts of information heterogeneity can be mitigated by making more informed and longer-looking choices that give due weight to organizational considerations.

3.2 Information Heterogeneity Impacts

Information heterogeneity is often subtle, and frequently has unforeseen and profound long-term ramifications. Many volumes could be written about its impacts; this paper, however, considers only the most important: complexity, functionality, interoperability, and cost. Each of these areas applies to NOAA generally and its information management efforts, organizations, and systems specifically, and each is a key consideration for evolving NOAA’s information management approaches and achieving its overall objectives. The following sub-sections discuss these key areas of impact.

3.2.1 Complexity

Simply put, heterogeneity tends to increase the complexity of the systems, processes, and organizations that have to deal with it. In the case of IT systems, complexity can be thought of as a function of heterogeneity, functionality, and resources, where increases in heterogeneity and functionality tend to increase complexity and resources tend to bound it. The catch, of course, is that complexity is expensive; complexity increases are tightly linked to cost increases. The following figure illustrates this way of thinking about complexity.



Where:

Increasing heterogeneity tends to increase complexity
Increasing functionality tends to increase complexity
Resources bound complexity

Figure 1 – Complexity Function

The complexity triangle expresses relationships analogous to those among cost, schedule, and performance. These are often presented with an offer to “pick any two,” along with observations like the following:

- Aggressive schedules and high performance requirements increase cost.
- Aggressive schedules and cost reductions decrease performance.
- High performance expectations and cost reductions negatively impact schedules.

While the general consensus is that two of the qualities may be achieved through compromise, in many cases it is more likely that only one variable can be fully optimized, as maximization of one aspect requires compromise virtually everywhere else.

Figure 1 applies this way of thinking to heterogeneity, and leads to observations like the following:

- Resource constraints limit the amount of heterogeneity that can be managed and/or the functionality that can be provided.
- Similarly, high levels of heterogeneity and/or robust functional requirements necessitate the

availability of substantial resources.

- Heterogeneity and functionality are a “trade space” in the context of bounded resources.

It is worth noting that, even in the presence of abundant resources, it is extremely difficult to sustain high levels of both heterogeneity and functionality. The resulting complexity is simply too much to effectively manage.

As one of the most prominent and highest impact forms of heterogeneity in NOAA, information heterogeneity exhibits all of the characteristics and effects ascribed to heterogeneity in the preceding discussion.

The GEO-IDE concept of operations describes information heterogeneity, functionality, and resource consumption similarly:

There are many functions included within end-to-end data management. It is clear that developing and maintaining the tools required to support these functions represent a significant investment of resources. Many of these tools depend critically on the format of the data and associated metadata. The amount of software that needs to be created and maintained, and the resources required to accomplish this, therefore, depend critically on the number of formats used for similar or identical data types. Developing and maintaining multiple tools for visualizing multi-dimensional grids, for example, is not an effective application of resources. [GICO]

The unfortunate predicament for NOAA is that it has extraordinary levels of heterogeneity, very high expectations for functionality, and significant resource constraints. In many ways, this a “perfect storm,” with great potential for unmet expectations, cost and schedule overruns, and failed information management initiatives.

3.2.2 Functionality

The software coding process is often one of the most expensive aspects of the development phase of information management initiatives. For systems where similar functionality is required across a large number of information types, information heterogeneity has profound cost and complexity impacts. Even when emphasis is placed on modular and generic designs, information-specific customizations are often required. Of course, information-specific implementations of the same functionality demand resources that could otherwise be used for new functionality.

For example, consider the generally straightforward operation of browse image generation. Heterogeneity across information types for which browse images must be generated often results in separate software for each type. (Note that this implies two separate sets of “knowledge” as well, in that information-specific knowledge is required to understand each underlying information type, and IT-specific knowledge is required for the implementation of functionality for each type.) The additional complexity may well be manageable for a small number of types, but support for even relatively simple operations like browse image generation becomes problematic as the number of types increases. Of course, the potential for serious problems increases dramatically when the number of supported operations is large and/or increasing.

The preceding discussion points out a particularly troublesome aspect of information heterogeneity.

Often, decisions regarding heterogeneity are made early-on, when the full picture of functional requirements and supported information types is still developing. It may seem expedient to simply deal with the variety given that complexity is predicted to be relatively manageable. Over time, however, expanding functional requirements and mandates to handle an increasing variety of information types often result in unanticipated levels of complexity, rendering functional enhancement difficult at best.

The impacts of functional scope creep and increasing information heterogeneity compound most dramatically in efforts with long-term missions, like information preservation. If digital information preservation entailed only the provision of sufficient physical resources like storage and bandwidth, then it would be relatively straightforward to accomplish. Digital information preservation is significantly more complex than this, however, due mainly to two factors: 1) a substantial breadth of functionality is required to support archive user requirements; and 2) archives have long-term missions, necessitating that not only information, but functionality as well, must be maintained over significant periods.

These notions are echoed in the *National Oceanic and Atmospheric Administration Report to Congress on Data and Information Management 2007*:

NOAA invests substantially in the acquisition of quality data, resulting in a national asset that must be protected for current and future generations. Maintaining a digital data archive means more than simply storing files on an appropriate digital medium. The data must remain usable over time, which means, among other things, that the data must be described thoroughly and accurately. [NOAA2007]

It is clear that large amounts of information heterogeneity in the context of limited resources necessarily results in functionality limitations. This is particularly the case for digital preservation initiatives, given their long-term mission.

3.2.3 Interoperability

In simple terms, interoperability is the ability for entities to interact efficiently and effectively. While interoperability can apply to people, processes, organizations, and many other kinds of entities, it is typically used in regard to IT systems. In an organization like NOAA, where numerous, diverse IT systems are combined with broad requirements for collaboration and exchange, interoperability is a key organizational objective.

The following excerpt from NOAA's 2007 report to Congress emphasizes the importance of interoperability to NOAA:

The ability to integrate data from multiple sources enhances all of NOAA's data intensive endeavors, from multidisciplinary, team-oriented research to the development of data products serving the Nation's aviation, transportation, and marine communities. Data integration builds upon standards for data and metadata and requires interoperability among independently operated data systems. NOAA is making progress toward interoperability, beginning with the development of GEO-IDE. GEO-IDE is a recently initiated effort to plan for data system interoperability across NOAA and beyond. It is a framework that enables the effective and efficient integration of NOAA's many existing

systems and will serve as a guideline for future data system development efforts.
[NOAA2007]

Unfortunately, information heterogeneity may be the largest single impediment to interoperability. The following sub-sections discuss various aspects of interoperability and its relationship with information heterogeneity.

3.2.3.1 Types of Interoperability

Interoperability can be characterized in a variety of ways. In the context of information heterogeneity, two types of interoperability, “syntactic” and “semantic,” are fundamental. Syntactic interoperability involves the correct exchange of the structural aspects of information. This is typically achieved through the adoption of standard representations, transfer protocols, and so on. As an example, transferring a file between IT systems requires a bare minimum of syntactic interoperability. While transferring a file between systems is often necessary, in most cases it is not sufficient. Unless the sole purpose is to store it as a “blob” on the receiving system, or perhaps to simply forward it along to another system, the contents of the file will have to be accessed and utilized in some manner specific to what they “represent” or “mean.”

Semantic interoperability builds upon syntactic interoperability to add “meaning” to the exchanged information. Using the previous example, in most cases it is of little use exchanging the file unless the receiving system can make use of it in some way. In order to do so, the two systems have to be in agreement about the meaning—or semantics—of the exchanged information. Both systems must not only agree on the information being exchanged (what it is), but they must also agree to greater or lesser extent on its semantics (what it means), how it may and may not be used, and so on. As might be imagined, semantic interoperability is a great deal more complicated and difficult to achieve. It is, however, the gateway to true interoperability, and to many of the benefits organizations like NOAA seek from integration efforts.

3.2.3.2 Interoperability and Information Heterogeneity

Interoperability and information heterogeneity are closely related, and generally correlate inversely. Information heterogeneity severely impedes interoperability, as it increases not only the complexity of participating systems, processes, and organizations, but also the amount of effort needed to integrate them. Interoperability can be seen as a “second order” functionality problem, in the sense that information heterogeneity tends to drive the functionality common denominator downward for participating systems, resulting in a reduced common denominator across systems and, thus, a lower level of potential interoperability.

3.2.3.3 Interoperability and “System Integration”

It is a common, seductive, and potentially extremely expensive misconception that interoperability and system integration are synonymous. In fact, getting systems to “talk” is quite straightforward in today’s technological environment. The interesting—and much more difficult—part is enabling them to “say” something useful to one another and, in so doing, to provide increasingly useful and

powerful capabilities to their users.

For example, the telephone system is quite well integrated. It is possible to dial virtually any phone on the planet from another, have a connection established, and so on. Unfortunately, this technological achievement is of little value unless it can be put to good use by phone system users. For the phone system to be useful, parties on both ends need to be able to speak to one another in a common language. Even cultural context can greatly influence successful communication. Absent the ability for the parties to effectively communicate, the benefits of an integrated phone system are reduced a great deal.

So it goes with information systems. It is relatively easy to agree on communication standards and protocols for all of NOAA's systems, and to connect them such that bits can be exchanged. It is a much more difficult problem to ensure that these systems provide the foundation for an environment in which organizational and customer needs can be fulfilled easily and economically. Unsurprisingly, the biggest impediment to the development of such an integrated and capable environment is making sure that the information exchanged is understandable and usable by the systems and users who need it. Information interoperability, like language and cultural context in the world of voice communication, is the key ingredient to providing highly-desired "integrated environments."

GEO-IDE provides a good example of these concepts. GEO-IDE's mission is often characterized as "systems integration" and assumed to encompass primarily syntactic interoperability. However, as previously discussed, the key to information exchange (a fundamental underpinning of a truly integrated information environment) is not simple system integration, but information interoperability. If the exchanged information is usable across the various systems and people that need to use it, linking the systems together to transfer and store the information is the "easy part." Significantly, information interoperability is dependent on high levels of semantic interoperability. Thus, the initial question for interoperability initiatives like GEO-IDE should not be: "how do we make systems communicate with one another?" but rather: "how do we ensure information interoperability?" One of the key first steps in establishing information interoperability is the reduction and management of information heterogeneity.

In short, system integration without information interoperability is a hollow victory at best; even in the best case, it amounts to an expenditure of effort with very little in return.

3.2.4 Cost

This section discusses some of the ways in which information heterogeneity impacts the cost of information management efforts.

3.2.4.1 Direct Costs

As Figure 1 illustrates, direct costs for information management can be seen as a function of heterogeneity and functionality. In simple terms, providing a given level of functionality is more expensive when more information heterogeneity is present. The obvious corollary is that, for a fixed amount of resources, more information heterogeneity means less functionality.

The reason is straightforward: it takes effort to specialize systems, processes, and organizations in order to accommodate information heterogeneity. These specializations result in additional complexity, which drives cost increases; complex entities are more difficult and costly to develop than simple ones. Moreover, complex systems are more expensive to operate and maintain. Finally, as noted previously, time is a compounding factor as well; these impacts are more dramatic over long periods.

3.2.4.2 Opportunity Costs

The direct costs of information heterogeneity are compelling, but they pale in comparison to the opportunity costs. Every resource expended to account for information heterogeneity is unavailable for the development or provision of new products or services. In the context of the NOAA mission, opportunity cost is potentially staggering.

For example, the development of sophisticated products frequently requires the integration of multiple information types, often from different sources. In environments with a great deal of information heterogeneity, the difficulty and cost of integrating these information types may limit or preclude entirely the development of key products. The following excerpt from the GEO-IDE concept of operations echoes this theme:

Application of environmental data to multi-disciplinary problems is hampered by lack of agreed-upon and implemented standards needed to effectively identify, acquire, and correctly use all of the relevant data. [GICO]

For NOAA, where products often provide earlier and/or more accurate natural disaster predictions, clearer views of long-term environmental impacts, and so on, the cost of not being able to provide a crucial product can be immense.

To the extent that information heterogeneity consumes resources that could otherwise be applied to research and product development efforts or makes the development of new products infeasible due to the difficulty of integrating multiple information sources or types, it has a substantial—potentially overwhelming—opportunity cost.

3.2.4.3 Costs of Interoperability Deficits

As previously discussed, information heterogeneity and interoperability are tightly coupled, and they vary inversely. Interoperability generally tends to reduce complexity, increase efficiency, and provide an environment in which functionality is easier to develop, provide, and maintain. Information heterogeneity, as noted previously, tends to increase complexity, decrease efficiency, and foster an environment in which functionality increases are more difficult and costly.

Thus, interoperability improvements are a kind of information heterogeneity opportunity cost. Though there have been very few attempts to quantify the cost of information heterogeneity, there have been efforts to quantify the costs of interoperability deficits. For example:

- A 1999 NIST report estimated “*at least a billion dollars*” of annual cost to the automotive supply chain due to “*imperfect interoperability.*” [NIST99-1] Interestingly, this study

focused on exchanges of items like CAD files which, while heterogeneous across platforms and software packages, exhibit substantially less heterogeneity than NOAA environmental information. Moreover, there are a relatively small number of CAD purveyors to the automotive industry. Finally, the automotive industry generally exhibits a fairly high degree of standardization, as many parts are commodity items and engineering is a relatively mature field.

- A 2004 NIST study of the U.S. capital facilities industry projected an annual cost of \$15.8 billion due to “*inadequate interoperability among computer-aided design, engineering, and software systems.*” [NIST2004] In this industry, too, there are relatively few suppliers of the kinds of systems evaluated in the study, and—relative to NOAA—a very limited amount information heterogeneity.
- A 2005 study of the costs associated with information exchange in the health care industry projected a potential annual benefit of \$77.8 billion if current interoperability limitations were addressed. [Walker2005]

While these analyses do not apply directly to NOAA’s information heterogeneity and interoperability concerns, they highlight several important points:

- There is an increasing focus on the costs associated with interoperability deficiencies (thus, information heterogeneity) across a variety of industries.
- It is possible to perform studies to quantify these costs, and there is expertise and experience within the Department of Commerce in this area.
- Available studies consistently conclude that interoperability deficiencies have substantial cost consequences.
- These costs are significant even in industries where there is already a fairly high degree of interoperability. This suggests that relatively small interoperability improvements can yield significant benefits and, similarly, that even relatively small interoperability deficits can be significant cost drivers. In short, interoperability deficiencies are high-impact or high-leverage cost factors.

It is important to note that in each of the examples cited, information is a necessary part of doing business, but not necessarily the *object* of the business. Automotive manufacturers use information to make cars, the health care industry uses information to care for patients, and the capital facilities industry uses information to develop and maintain physical plants. For NOAA, however, there is a much more direct relationship between the information it uses to do business and the object of its business. In fact, to a large degree NOAA’s business *is* information. This strongly suggests that interoperability deficiencies and information heterogeneity are of much greater significance to NOAA than to the other industries cited, and that the potential benefits of improvement are much greater.

3.2.4.4 Return on Investment Associated with Standards Adoption

Interestingly, NASA performed a cost/benefit analysis regarding the return on investment for open geospatial standards adoption [NASA ROI]. Two projects were compared, one that freely adopted open geospatial standards (Case Study 1), and another that relied primarily on proprietary approaches (Case Study 2). Following are selected findings from the study:

- ... *the project that adopted and implemented geospatial interoperability standards had a*

risk-adjusted ROI of 119.0%. This ROI is a “Savings to Investment” ratio. This can be interpreted as for every \$1.00 spent on investment, \$1.19 is saved on Operations and Maintenance costs.

- *Overall, the project that adopted and implemented geospatial interoperability standards saved 26.2% compared to the project that relied upon a proprietary standard. One way to interpret this result is that for every \$4.00 spent on projects based on proprietary platforms, the same value could be achieved with \$3.00 if the project were based on open standards.*
- *Standards lower transaction costs for sharing geospatial data when semantic agreement can be reached between parties. The cost of achieving semantic agreement can be high, especially for data models. This cost is reflected in the higher implementation costs for Case Study 1. However, these costs are more than recouped in lower operations and maintenance (O&M) costs. In this study, risk-adjusted costs for Case Study 1 was 30.3% lower than those for Case Study 2.*
- *Case Study 2 had almost double the risk premium in Planning and Development costs. Moreover, Case Study 2 had a roughly 50% increase in risk in Acquisition and Implementation costs than did Case Study 1. Case Study 2 also had almost double the risk premium in M&O. Part of this increase in risk was due to the original cost structure developed by Case Study 2, where the majority of its costs (89%) were M&O costs. Because Case Study 2 had most of its costs in this category, and this category is exposed to the greatest risk over time, Case Study 2 had the largest increase in risk-adjusted M&O costs.*

Of particular interest is the following observation from the study (emphasis added):

- *The results indicate Case Study 2 Planning & Development costs increase 27.4% over the base year, Acquisition & Implementation increase 33.2%, and Maintenance & Operation (M&O) increases 59.6%. **Most notable is the increase in M&O costs, which suggests that the use of proprietary models limits the flexibility and adaptability of the program over time.** As M&O is the largest single contributor, this risk is the primary driver in the program’s 56.6% total increase in cost. For Case Study 1, the total cost increase due to risk is only 24.6%, which is significantly lower than Case Study 2. More important to Case Study 1 is the fact that M&O costs are a lower percentage of total costs than in Case Study 2.*

The comments about recurring costs have obvious implications for any information management initiative with a long-term mission, as the potential impacts are enormous over time.

The NASA study concludes that investments in standards adoption have very high potential rates of return. Since standardization and information heterogeneity are, in an important sense, opposite ends of a spectrum, it is not difficult to see that rampant information heterogeneity is likely to have an equally dramatic cost.

It is worth noting also that the NASA study identified substantial returns for standardization investments despite the relatively brief lifetimes of the two case study projects. Applying this concept across NOAA, especially considering the long-term nature of NOAA’s mission and many of its NOAA information management activities, the potential return is compounded immensely.

3.2.4.5 Summary

Information heterogeneity is tremendously expensive for large-scale information management efforts

like NOAA’s, where it is very often coupled with high functional expectations. Unfortunately, both the problem and the cost only compound with time.

While the direct costs of information heterogeneity are compelling, the opportunity costs are potentially overwhelming. Industry studies show that interoperability improvements—a key NOAA objective greatly hampered by information heterogeneity—have the potential for substantial cost savings, and that standardization—a primary means of addressing the information heterogeneity problem—can provide significant returns on investment.

Information heterogeneity is simply too expensive to ignore. New NOAA products and initiatives are essential to society, and entirely dependent on resource availability. Currently, significant resources that could be invested in these pursuits are being expended coping with information heterogeneity. Fortunately, there are known solutions to the information heterogeneity problem, and investments in these solutions will both begin to pay off quickly and yield compound benefits over the long term.

3.3 Increasing Data Volumes and Information Heterogeneity

NOAA archive data volumes are increasing dramatically, as indicated in the following figure.

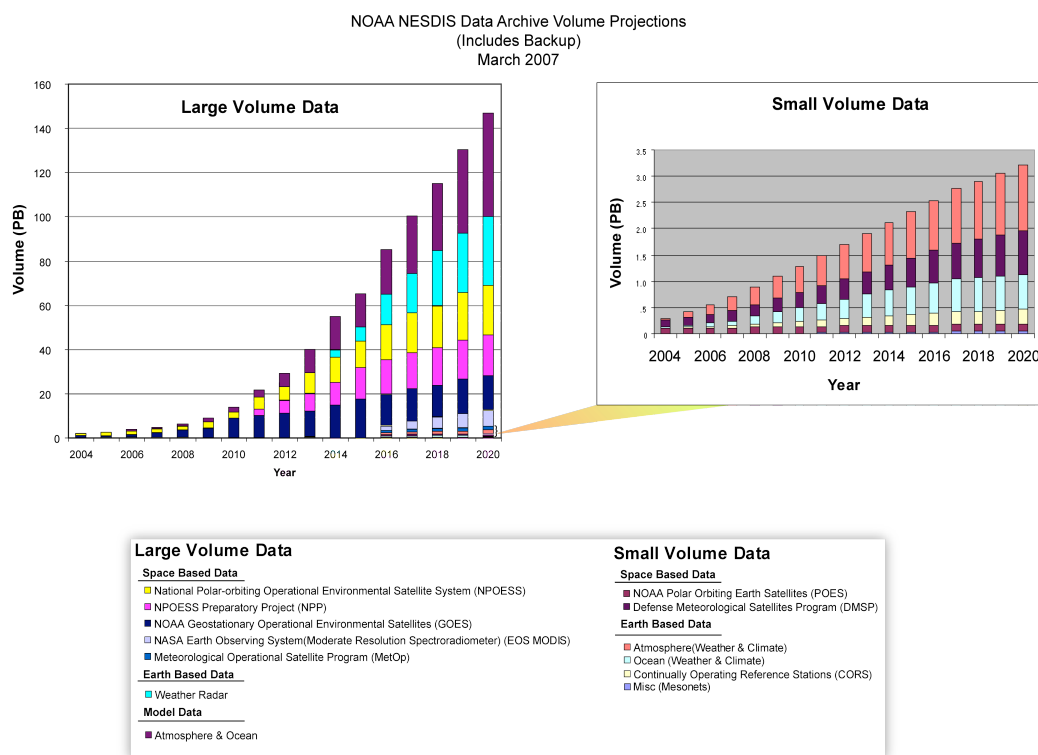


Figure 2 – NOAA Archive Volumes, 2004-2020 [NCDC unpublished presentation]

These increases, in and of themselves, have significant ramifications for information management across NOAA, as evidenced in the following excerpt from NOAA’s 2007 report to Congress:

NOAA’s data management technology must continue to evolve in parallel with its

environmental sensing technology or face being overwhelmed. [NOAA2007]

Addressing the increasing data volumes problem is generally straightforward, though certainly not easy. It typically combines the acquisition and integration of sufficient commodity resources like storage, bandwidth, and so on, with software design and development that allow these underlying resources to be easily scaled. Significant efforts are being expended to plan for and accommodate anticipated data volume growth.

One frequently overlooked impact of increasing data volumes, however, is the degree to which they combine with information heterogeneity to increase risk. Interestingly, the impacts of large data volumes and information heterogeneity are addressed quite differently: large data volumes are best addressed with commodity solutions, while information heterogeneity by definition necessitates custom solutions.

Increasing data volumes and information heterogeneity combine to greatly increase overall risk to information management initiatives. As previously discussed, information heterogeneity tends to increase complexity which, in turn, tends to increase the likelihood of a wide variety of problems. Large data volumes tend to reduce the margin for error, as they drive resource consumption upward. With complexity increasing the likelihood of problems and large data volumes decreasing the margin for error, the increase in risk is obvious.

For example, consider the responsibility of transforming archive holdings to keep pace with technological changes and to maintain long-term understandability and usability. As archive information holdings grow, the number and extent of transformation efforts grow, as do the physical resources required to execute these efforts. In an environment with a great deal of information heterogeneity, a large number of software components will be needed to perform the transformations. The more of these there are, and the more complex each is, the more likely it is that problems (e.g., software bugs) will necessitate rework. However, large data volumes increase the time required to perform the transformation such that it may not be possible to perform it multiple times. Together, increasing data volumes and information heterogeneity create an environment in which rework is potentially very expensive (driven by large volumes), much more likely (driven by complexity arising from information heterogeneity), and may not be practical (due to time or other resource limitations) even when it is necessary.

Section 4.0 Information Heterogeneity Across NOAA

Information heterogeneity is having profound impacts across NOAA, and these impacts will only be magnified in the future. IT systems and other information management entities have long experienced these impacts, and the problems caused by information heterogeneity—if not addressed now—will soon make it impossible for NOAA to fulfill all of its requirements and meet its strategic objectives. This section discusses some of the ways that information heterogeneity impacts various NOAA entities and efforts.

4.1 Information Management Systems and Information Heterogeneity

4.1.1 Interest in Information Heterogeneity

Information heterogeneity is among a handful of the biggest challenges facing NOAA information management systems today. The sub-sections that follow discuss some of the reasons why information heterogeneity is such a significant driver for NOAA IT systems.

4.1.1.1 Key Challenge

Thus far, NOAA information management systems have generally attempted to manage information heterogeneity and its impacts through mechanical means: generic software, reusable modules, “configurable” functionality, application of standards where possible, etc. While these measures help mitigate some of the impacts of information heterogeneity, they have limited potential to address its most severe consequences. Moreover, these approaches have been applied in various ways and to various degrees by different systems.

Generally speaking, information heterogeneity substantially increases the difficulty of developing and maintaining common functionality across information types. The employment of some of the mechanisms mentioned previously has allowed NOAA information management systems to cope with the impacts of information heterogeneity with varying levels of success. However, as noted later in this section, the introduction of new information campaigns, requirements to preserve legacy campaigns, and increasing functional requirements all have potentially enormous impacts on NOAA information management systems.

Current role and responsibility allocations make information heterogeneity a key challenge for IT system developers, maintainers, and operators. Absent successful implementation of NOAA-wide information heterogeneity reduction and impact mitigation recommendations discussed later in this paper, information heterogeneity will be a powerful and increasingly limiting factor in both the number of information campaigns NOAA IT systems can handle and the amount of functionality they can provide.

Each instance of information specificity increases information management system complexity, thus consuming additional resources per unit functionality and making it more difficult and costly to add

new information types. While NOAA IT systems can—and should—continue to implement generic mechanisms to support the addition of heterogeneous information types without greatly increasing information specificity, this approach can only go so far. Even if NOAA managed to achieve the significant goal of successfully eliminating all information specificity in its software, there would still be enormous complexity in terms of the configuration details required to describe the heterogeneity for generic software mechanisms. While transfer of the problem from development to configuration is a key objective and would represent a remarkable achievement, configuration complexity would still consume a great deal of resources and require a staffing mix inappropriate to IT system development, operations, and maintenance.

It is worth noting the importance of this last point: information heterogeneity has a human impact. IT personnel must currently develop and maintain information-specific expertise when their efforts would be much more efficiently applied strictly to IT concerns. Even in the previous scenario, where information specificity was relegated to system configuration, it would require a significant number of information specialists to manage the relevant details successfully.

Information heterogeneity is not only one of the most difficult problems facing NOAA information management systems, but also a driver for everything from their system architectures to their staffing mixes.

4.1.1.2 Subtlety of the Problem

Information heterogeneity and its impacts can be subtle, and are often difficult to quantify, making it harder to relate to than some of the other problems NOAA IT systems are facing. Take, for example, the problem of increasing data volumes. As depicted in Figure 2, data volumes are easy to quantify, can be measured directly in well-understood and commonplace units, and can be predicted fairly accurately. Moreover, the impacts of increasing data volumes can be readily modeled and accounted for in planning efforts. For all of these reasons, the increasing data volume problem is easy to grasp and communicate.

By contrast, information heterogeneity appears in many different forms, each of which has differing impacts at different times. There are no commonly understood information heterogeneity “units.” The effects of increasing information heterogeneity are often subtle: it can be difficult to see how the introduction/acceptance of some seemingly insignificant variation today ultimately impacts the cost and maintainability of a system, particularly if the system has a long-term mission. Finally, the problems arising from information heterogeneity must be addressed via software, policy, and management, all considerably less expeditious and more expensive than expanding hardware resources (the primary solution to the increasing data volumes problem).

An interesting example of the ease with which the data volume problem is characterized can be seen in the number of times NOAA’s 2007 report to Congress mentions volume increases as the driver behind lagging archive and access capabilities across the Department of Commerce. In stark contrast, information heterogeneity is never mentioned as a cause of these problems, though the need for—and efforts toward—data “integration” is discussed in other places in the document.

4.1.1.3 Key Role and Responsibility Driver

As will be discussed later in this paper, the proper allocation of roles and responsibilities is a key tactic for managing information heterogeneity. Roles and responsibilities are currently the subject of a great deal of discussion within NOAA, particularly with regard to NOAA archives, and a comprehensive understanding of information heterogeneity is essential to the establishment of optimal role and responsibility allocations.

4.1.1.4 Significant Complexity Driver

As described in some detail previously, information heterogeneity greatly increases the difficulty and scope of information management system development efforts and the complexity of the resulting systems. In addition to the ramifications discussed previously, one significant result is the introduction of considerable risk to IT system longevity and success.

4.1.1.5 Significant Interoperability Impediment

Information heterogeneity poses significant barriers to interoperability. As noted previously, and confirmed in cost studies, interoperability is a significant success factor for information management initiatives and interoperability deficiencies have substantial cost ramifications. A common understanding of information heterogeneity and its implications is crucial to information management system participation in NOAA's primary interoperability initiative, GEO-IDE.

4.1.1.6 Significant Aspect of IT Support

While information stewards are responsible for the information-specific aspects of the information management lifecycle, many information-related requirements flow down to IT systems. Moreover, IT systems have to provide the mechanisms that support information heterogeneity management, so information heterogeneity is at the center of some of information management systems' most important contributions to NOAA infrastructure.

4.1.1.7 Key Architectural Driver

Many NOAA IT systems' architectural artifacts and approaches will have to be designed specifically to provide a framework sophisticated enough to efficiently and economically address predicted levels of information heterogeneity. Further, these architectures must be very abstract because the wide spectrum of information contains few common factors. In many cases, NOAA IT systems must provide support for entities without knowing in advance what those entities might be. Finally, architectures will have to promote the encapsulation of information-specific functionality to help IT systems manage information heterogeneity and to support the proper allocation of information-related roles and responsibilities.

4.1.1.8 Enormous Cost Driver

As mentioned, information heterogeneity drives both direct and opportunity costs. These costs are

substantial for IT systems, and limit the amount of functionality they can provide. Reductions in resources necessary to accommodate information heterogeneity will enable NOAA information management systems to provide more and improved support for NOAA and its customers.

4.1.1.9 The Buck Stops with IT Systems

In some important ways, the “buck stops” with IT systems. If it is not reduced or managed “upstream,” IT systems become the recipient of information heterogeneity even though their focus is ideally on pure IT issues. It is, thus, crucial for IT system management to highlight the issues associated with information heterogeneity so all information management-related parties can participate in efforts to mitigate associated problems and ensure success.

4.1.2 The Potential Problem: CLASS as a Case Study

Up to this point, the focus of this paper has been primarily conceptual. This section places some of the concepts previously discussed in context by looking at the actual scope of information heterogeneity and its relevance with regard to a single IT system: CLASS. The issues highlighted in this sub-section and their implications are applicable to many of NOAA’s information management systems and efforts.

4.1.2.1 Increasing Information Heterogeneity Levels

The *CLASS System Level 1 Requirements Document and Mission Success Criteria, Preliminary, version 1.14* states:

The intent is to evolve CLASS into an enterprise solution supporting long-term, secure storage of and common web-based access to NOAA Archive maintained environmental data and information.

...
NOAA has directed that as new environmental data is identified, that the CLASS IT infrastructure be considered for satisfying storage and access requirements, [CLASS-LIR]

Increases in the breadth of CLASS holdings and the wide variety of NOAA information types are driving information heterogeneity within CLASS rapidly upward. In combination with increasing functional requirements, these forces are pushing CLASS’s complexity level beyond that which can be addressed with current resources. These forces are present across NOAA as well; functional requirements are increasing at the same time more sources and types of information must be supported—all in an environment of constrained resources.

Given current levels of NOAA information heterogeneity and archive functional expectations, the preservation of legacy collections will significantly increase levels of information specificity in CLASS. A survey of several catalogs that describe NOAA legacy holdings suggests that CLASS currently holds less than 5% of NOAA’s information “types” or “products,” as presented in Table 3.

Table 3 – NOAA “Types” Stored in CLASS

Catalog System	Number of CLASS "Types"	Number of NOAA "Types"	Percentage Currently Held by CLASS
CasaNOSA	37	~1,000	~4%
GCMD	43	~1,700	~3%
Geospatial One Stop	48	~1,000	~5%

None of these catalogs is fully populated with descriptions of NOAA’s legacy holdings, so these percentages are conservative. Ultimately, CLASS may see increases of 20 times—or more—in the number of information types it must manage. The number of new information types created by emerging and future campaigns is difficult to predict, but will certainly be considerable.

The extent to which these candidates for CLASS inclusion are heterogeneous is presently unknown. Based on previous experience, however, it can be assumed that they encompass a substantial amount of heterogeneity, and that some archive-related entity—whether CLASS or one or more information stewards—will have to deal with it directly. Indirectly, CLASS will have to continue to develop increasingly numerous and generic mechanisms to support NOAA information preservation efforts.

4.1.2.2 Increasing Functional Scope

As with many NOAA IT systems, CLASS functional requirements are increasing. Among CLASS’s new functional requirements are enhanced search capabilities, more robust spatial capabilities, support for provenance information, metadata subsetting, interoperability with a wide variety of NOAA and external systems, and many others. Implementing these capabilities across a substantial part of NOAA archive holdings is greatly complicated by the heterogeneity of those holdings.

As an example, *CLASS Archive and Access System Requirements version 2.2* includes the requirement that CLASS must be able to subset its metadata holdings to properly describe information to be disseminated, and also notes (emphasis added):

*The metadata model used by CLASS must include **all metadata provided to CLASS** ... CLASS must collect and maintain metadata detailed enough to meet the needs of the **most demanding users that can be imagined** ... [A&ARv2.2].*

While customizing metadata for dissemination is certainly a reasonable and important requirement, doing so regardless of the metadata’s types, forms, and levels of detail is an immense, complicated, and expensive undertaking given current levels of information heterogeneity. This is particularly true given the requirements to meet the needs of the “most demanding users” imaginable. Here, too, CLASS is a good barometer for many of NOAA’s systems, in that the reliance on, complexity of, and functional requirements regarding metadata are increasing across NOAA generally.

Many similar functional requirements—at the same time necessary, difficult, and complicated—must be met in order to fulfill NOAA archive objectives. Reductions in information heterogeneity, or improvements in its management, have great leverage as they serve to make the functionality necessary to fulfill these requirements easier and less expensive to develop, operate, and maintain.

4.2 Information Heterogeneity and Information Stewards

Information stewards have responsibilities throughout the information lifecycle. This requires detailed knowledge of each information type's representation, provenance, context, requirements, policies, producers, consumers, etc. The wider the range of information types and the more variation across types, the more difficult the information steward's job becomes.

Information stewards are also responsible for providing customized interfaces for their archive and other customers, and specifying or developing other information-specific software and products. While information stewards may elect to outsource the development of information-specific capabilities, they remain responsible for identifying and specifying all of the information-specific detail required for development. Information heterogeneity dramatically increases the resources information stewards need to support these efforts.

Long-term information preservation requires archives to update the representation of their holdings when current representations are becoming unusable. Each information type's representation must be tracked, periodically analyzed for usability, and then eventually transformed. In effect, heterogeneous information holdings form information "stovepipes" that must be managed and preserved distinctly. Since these "stovepipes" are all information-specific, information stewards inherit responsibility for all of this complexity and the resources necessary to support it.

4.3 Information Heterogeneity and NOAA Archives

Since, for the purposes of this paper, a NOAA archive is primarily a composition of one or more information stewards and IT systems, the NOAA archives inherit all of the information heterogeneity issues described previously for each of these entities. It goes further than that, however, as the complexity introduced by information heterogeneity demands additional organizational layers, processes, and communication that would otherwise be minimized.

4.4 Information Heterogeneity and GEO-IDE

GEO-IDE's mandate is to facilitate the creation of a NOAA "integrated data environment." Interoperability is one of the key means by which this is to be achieved. Information heterogeneity, though, is a significant impediment to this goal:

Data and products are available through incompatible interfaces and formats, and services from multiple centers cannot be easily combined. [GICO]

GEO-IDE's mission includes the establishment of interoperability for legacy information and systems, and the creation of a framework in which future information campaigns can evolve in a manner more consistent with organizational interoperability objectives:

Continuing to develop systems in an uncoordinated manner will lead to further incompatibilities and will further isolate NOAA programs from each other and from the wider environmental community. This will increase the difficulty in integrating information between programs.... [GICO]

These excerpts, taken together, provide an overview of GEO-IDE's task: establish information interoperability and then integrate systems. As noted in the interoperability discussion, system integration is a substantially easier task than establishing information interoperability. As a result, information heterogeneity is, in all likelihood, GEO-IDE's biggest obstacle. Conversely, any successes in addressing information heterogeneity and its impacts will make GEO-IDE's task much easier, less risky, and successful much more quickly.

4.5 Information Heterogeneity and NOAA

NOAA's responsibility for the full information lifecycle both broadens and amplifies the effects of information heterogeneity seen in its IT systems and individual information management activities. In many ways, these specific information management efforts are "microcosms" of the much greater information heterogeneity problem present across the organization. In reality, NOAA inherits not only the complexity of all of its individual information management efforts, but all of the difficulties associated with making them work together efficiently and effectively as well.

The sub-sections that follow focus on a few aspects of information heterogeneity and its impacts within NOAA that have not been discussed in detail previously.

4.5.1 Significance across the Organization

Following are five "themes" NOAA has identified in reviews of its information management challenges, and subsequently reported to Congress:

In viewing NOAA's data management challenges as a whole, the same five themes emerge from this assessment as from the 2005 assessment. These are:

- *managing the increasing volume and diversity of data*
- *extending and filling gaps in environmental data records*
- *improving access to the long-term archive*
- *enabling integration of quality observations and products*
- *improving descriptions of data, metadata, formats, and processing steps*

[NOAA2007]

It is interesting to note that information heterogeneity is directly mentioned in, or has a clear relationship to four of these:

- *managing the increasing volume and diversity of data*
- *improving access to the long-term archive*
- *enabling integration of quality observations and products*
- *improving descriptions of data, metadata, formats, and processing steps*

and is likely a complicating factor for the fifth:

- *extending and filling gaps in environmental data records*

The fact that information heterogeneity is a factor in each of these themes illustrates the breadth of

the problem across NOAA, and its significance to NOAA information management efforts. While this may seem daunting at first, there is a substantial upside: successes in mitigating information heterogeneity and its impacts will yield returns for all of NOAA's information management efforts. Thus, there is a great deal of leverage associated with incremental improvements in the problem. The potential return on investment for information heterogeneity management efforts is very high.

4.5.2 Information Heterogeneity and NOAA Customers

For the most part, this paper has focused thus far on impacts within NOAA. It is important to note that the effects of information heterogeneity are also felt by NOAA's customers. Information heterogeneity makes it harder for customers to identify and obtain information of interest, as they must frequently access multiple stovepipe systems, only to receive the information in a variety of formats, resolutions, coverages, units, etc. This concern is noted in the GEO-IDE concept of operations as well:

The extraction of data from data collections has often been a weak area in responding to user needs, primarily because of the multiplicity of formats presently produced by NOAA production systems. [GICO]

Often, consumers choose to use a single information type because they lack the resources (money, time, or even willingness) necessary to integrate multiple types. When this happens, it affects not only the quality of the research and generated products, but by extension the value of NOAA's information holdings, as the degree to which these holdings are utilized is one of the primary indicators of their value.

In fact, one of NOAA's stated objectives is to increase the utilization of its information holdings for just this reason: the more NOAA information holdings are used, the more benefit they provide to society, and the more value they have. In order to achieve this goal, NOAA must make it easier for customers to identify, obtain, and use NOAA information. Reduction and management of information heterogeneity is, thus, a crucial aspect of achieving the goal of increased utilization of NOAA information.

4.5.3 NOAA's Role in Larger "Systems of Systems"

NOAA is assuming a leadership role in the establishment of national and international "systems of systems" like USGEO and GEOSS. The success of these systems will hinge largely on the degree to which they achieve information interoperability. Accordingly, NOAA has a unique opportunity to both support its own internal information management efforts and to help chart the course for future environmental information initiatives by taking the lead in addressing information heterogeneity.

4.5.4 Summary

Information heterogeneity is an organization-wide problem. Each aspect of the information management lifecycle is affected by it, and the broader the activity or organization, the greater the impact. Fortunately, the benefits derived from information heterogeneity reductions are also realized

across the entire organization, thus presenting the opportunity for exceptional return on investment from information heterogeneity reduction and management efforts. Moreover, NOAA's customers can benefit greatly from reductions in information heterogeneity, as one of the biggest barriers to increasing information utilization is the difficulty of dealing with the great multiplicity of NOAA information types.

Section 5.0 Something Has to Give!

NOAA simply cannot fulfill functional expectations with current heterogeneity and resource levels. From the previous complexity discussion, the alternatives are clear: limit functionality, increase resources, or mitigate heterogeneity. The following sub-sections examine each of these alternatives.

5.1 Limit Functionality

Earth research is intensifying in both breadth and depth, driving demand for information and functionality dramatically upward. NOAA must keep pace with increased demands for both, or become less relevant.

Moreover, NOAA is expected to provide increased societal benefits in a variety of forms. These benefits can be realized only if NOAA improves and expands the ways in which it utilizes information and offers both the original information and derived products to its customers. Put more simply: increased societal benefits depend on increased functionality.

For these—and many other—reasons, functionality expectations are increasing and will continue to do so. Failure to meet these expectations will have significant negative impacts for NOAA. Moreover, functionality trends within the IT industry are always towards doing more, and users certainly expect NOAA information management system capabilities to follow suit.

While reducing overall functional scope would mitigate some of the impacts of information heterogeneity, it would only address part of the problem. Regardless, trying to limit or reduce functionality expectations is, for the most part, a non-starter as it makes it impossible for NOAA to meet its stated objectives and fulfill customer expectations.

5.2 Increase Resources

Resource increases have the potential to increase the amount of complexity that may be managed, thereby proving one means of mitigating the effects of information heterogeneity. However, estimating how much of an increase in resources is needed is difficult, and if levels of functionality and information heterogeneity continue to grow, it is inevitable that the complexity “ceiling” will eventually be reached again.

Thus, while it is conceivable that additional resources will become available in the future, it is extremely unlikely that resource increases would keep pace with the current rate of heterogeneity growth. More realistically, it can be expected that resources will remain relatively constant, and that NOAA will be challenged to meet functional expectations without significant funding increases. Moreover, the achievement of NOAA’s most important strategic goals must not ride on the potentially-unfulfilled promise of resource increases. It is essential that NOAA achieve its objectives within resource levels that can be reliably projected.

It is also important to note that there is a point of diminishing returns when trying to solve problems

with increased resources. Some problems do not respond well to additional resources, and in all cases there is a point at which additional resources result in decreased efficiency.

For all of these reasons, increasing resources is not anticipated to be a likely (or even complete) solution to the information heterogeneity problem.

5.3 Mitigate Heterogeneity

Attacking heterogeneity directly is the only viable solution to the problem. Where functionality reduction and resource increases have significant drawbacks—including low probabilities of acceptance and/or success—information heterogeneity reduction and management have the potential for dramatic and long-term success. Even incremental success in these efforts will have ongoing, definable, and substantial benefits for NOAA, its customers, and society generally. Simply put, addressing the problem directly will allow NOAA and its information management initiatives to do much more with their available resources, and the benefits will only increase with time.

Section 6.0 Information Heterogeneity Impact Mitigation

The breadth and complexity of the information heterogeneity problem, particularly in a large and diverse organization like NOAA, can make it difficult to know where and how to begin mitigation efforts. This section focuses on selected approaches to mitigating information heterogeneity impacts, and lays the groundwork for the specific recommendations in Section 7.0.

6.1 First Steps

The following sub-sections discuss essential first steps to creating an environment in which information heterogeneity and its impacts can be successfully addressed.

6.1.1 Acknowledge the Problem

The first step in dealing with information heterogeneity is to fully acknowledge the problem. It must be understood as readily and intuitively as other problems, like the “large data volume” problem. It is particularly important for the entire organization to understand that information heterogeneity is a substantial obstacle to achieving NOAA’s strategic objectives, and that successes in addressing it will contribute significantly to NOAA’s ability to meet expectations over the long term.

6.1.2 Provide Organizational Support

Once the problem is acknowledged and understood, it is essential for NOAA to strongly support information heterogeneity reduction and impact mitigation efforts. This support must be reflected in the establishment of information heterogeneity-related tactical objectives, and figure prominently in performance assessments, funding decisions, and other activities that tend to influence behavior. This point can not be stressed enough: **NOAA must change the way it does business in order for its various information management organizations and systems to deal successfully with information heterogeneity.**

Addressing information heterogeneity is more a matter of behavior modification than technological innovation or application. This observation is echoed in the following excerpt from the IOOS data management and communication plan:

The greatest challenge to enhancing marine data integration is one of coordination and cooperation among the members of IOOS and its user communities. [IOOS DMAC]

Thus, the degree to which NOAA backs information heterogeneity reduction and impact mitigation initiatives by making them an integral part of NOAA culture will largely dictate their success.

6.1.3 Speak the Same Language

A key first step to addressing information heterogeneity, particularly in a diverse organization like

NOAA, is to establish a common framework for communicating about the issue. The OAIS-RM does this for information preservation, but leaves unaddressed many aspects of the information lifecycle. It is imperative to establish a standard way of talking and thinking about all facets of the information management lifecycle as a precursor to substantive information heterogeneity reduction and impact mitigation efforts. Developing an end-to-end information lifecycle reference model will be a time-consuming and difficult activity. However, this effort also serves as a means to establish and test the processes by which related efforts can be accomplished.

6.1.4 Think Incrementally

Information heterogeneity is a large and multifaceted problem that has taken a long time to develop. Even under the best of circumstances, it will not be possible to completely—or even largely—eliminate information heterogeneity and its impacts. It is not possible to tackle the problem in large chunks, nor would it be advisable to do so even if it were possible. Thus, it is crucial to establish incremental objectives, and to find ways of measuring and motivating incremental progress. Fortunately, even modest improvements can have dramatic positive impacts, so the incremental approach has the potential to be as successful as it is necessary.

6.1.5 Carefully Allocate Roles and Responsibilities

The proper allocation of roles and responsibilities is another key element in the creation of an environment in which information heterogeneity can be successfully addressed. This is particularly true within NOAA, where specialization involves a great deal of knowledge that is often costly and inefficient to replicate, and overall diversity can lead to a great deal of overlap. It is essential to understand and acknowledge the importance of role and responsibility allocations, to be rigorous in assigning them clearly and unequivocally, and to operate consistently within these assignments.

6.1.6 Apply Different Approaches

Different approaches, in terms of both method and timing, are needed to address various aspects of the information heterogeneity problem. For example, prioritizing the adoption of standards for use in future campaigns can help minimize the creation of additional information heterogeneity, while dealing with information heterogeneity in legacy campaigns can be deferred for some time. Similarly, it is important to begin analyzing whether information from legacy campaigns can be sunsetted immediately, while the approach for upcoming campaigns should include consideration for long-term aspects of their lifecycle, like preservation. Planning for all lifecycle stages of future campaigns must be as comprehensive as possible, begin as early as practicable, and include consideration of legacy and other future campaigns, as well as overall NOAA information objectives. Information heterogeneity is a multi-faceted problem; addressing it requires a variety of approaches.

6.1.7 Establish a Framework

Establishment of a framework for dealing with information heterogeneity is an essential precursor to taking specific action. Within the framework, new systems can be contemplated, conceived, and implemented, and legacy systems can be evolved – all consistently and with careful consideration of information heterogeneity and its impacts.

For example, information heterogeneity reduction and impact mitigation efforts can become part of the information lifecycle. Information heterogeneity may be naturally—and relatively painlessly—reduced when campaigns are sunsetted. Incremental improvements can be made during the transformation processes for actively-preserved collections. Finally, as new systems are contemplated and implemented, they can evolve in a coherent, end-to-end context of information interoperability. Establishment of a framework that relies on and accounts for the information lifecycle is a key element to incremental, consistent progress in handling information heterogeneity. Significantly, establishment of the framework does not require the substantial effort needed to effect the overall transformation, but can start providing tangible returns nearly immediately.

6.2 Information Heterogeneity Reduction

Once an appropriate foundation is established, the information heterogeneity problem should be approached by: 1) reducing information heterogeneity where possible, and 2) managing the impacts of information heterogeneity that cannot be eliminated. This sub-section focuses on the first of these.

Both reduction and management are essential to any comprehensive mitigation effort. It is important to note that information heterogeneity reduction efforts are typically broader in scope and somewhat more difficult than those for managing information heterogeneity impacts. Just as important, however, is that reduction efforts are much higher leverage and provide returns over a much longer period.

Information heterogeneity reduction involves the elimination of unnecessary variation. This can be accomplished in two distinct ways: 1) eliminating unnecessary information types, and 2) application of standardization where possible to those information types that cannot be eliminated.

6.2.1 Elimination of Unnecessary Information Types

Elimination of unnecessary information types has applicability throughout the information lifecycle, but has particularly strong impacts for long-term preservation. It is clear that reducing the number of information types can directly reduce the complexity of the systems, processes, and organizations that have to deal with them. In the long-term preservation context, however, the benefits compound over lengthy periods to yield dramatic returns.

It can often be difficult to decide to “abandon” information that has been carefully gathered and stewarded, but careful and unsentimental cost/benefit analysis is mandatory for any information being considered for long-term preservation. Information preservation is simply too expensive to be considered for any information unessential to the long-term record.

Moreover, the difficulties of long-term preservation cannot be overlooked. It is not simply a question

of whether a particular information type is essential to the long-term record. It must be established that it is possible to preserve the information, and that the institutional resources—and even willpower—necessary for long-term preservation are available. Instances abound where records have been stored and over time—due to retirement of key personnel, changes in priority, funding limitations, etc.—associated descriptive and representation information have been lost, thus rendering the information unusable. Similarly, the passage of time can also involve technological changes that render information difficult to use even if sufficient descriptive and representation information have been maintained.

Each time information usability is “overcome by events,” significant portions of the effort that went into its preservation may have been invested unwisely. Thus, it is crucial to accurately project not only which information is essential for the long-term record, but also whether it can (and will) be preserved such that it can be successfully used in the (potentially long-term) future.

While the preceding discussion focused mostly on long-term preservation, the principles apply across the information lifecycle. In an environment of resource constraints and high functional expectations, it is essential to focus on the things that absolutely must be done and the information crucial to doing them. Eliminating unnecessary information and/or information types is an important aspect of reducing information heterogeneity and, thus, a key tool to achieving overall NOAA objectives.

6.2.2 Standardization

Standardization is often considered the “holy grail” for information heterogeneity reduction and interoperability facilitation. While standardization is a crucial factor in the success of these efforts, its limitations must also be considered and expectations set accordingly. The following sub-sections discuss both the importance of standardization and some of its limitations.

6.2.2.1 The Importance of Standardization

Standardization has the obvious benefit of reducing information heterogeneity at its source: the variation across information types. It addresses the root cause of the problem and, as a result, is a very high leverage opportunity.

While standardization is commonly associated with information representation (e.g., file formats), it can be effectively applied to a wide variety of information heterogeneity-related areas. Syntax, semantics, functionality, information and metadata models, terminology, and interfaces are just a handful of the areas in which standardization can be applied to greatly reduce information heterogeneity, facilitate information heterogeneity impact management, and improve information management efforts overall.

Standardization has potential benefits both within NOAA and externally. Inside NOAA, the benefits of standardization include complexity and cost reductions, facilitation of research and derived product development efforts, and greatly improved interoperability. Externally, standardization is essential to large-scale “system of systems” integration efforts and substantially simplifies NOAA

customer information utilization efforts. Moreover, standardization can pave the way for increased functional capabilities—at less cost—both inside NOAA and out.

The importance of standardization for NOAA is clear, as evidenced by the following excerpt from NOAA’s 2007 report to Congress:

Historically, the development of NOAA’s data management systems was focused solely on a particular project, discipline, or observing system capability. While these legacy systems were individually efficient, many of the current and potential uses of NOAA data and services are interdisciplinary in nature. This requires access to a wide range of NOAA’s capabilities, and the disparate user interfaces, data models, protocols, formats, etc., of the legacy systems present a barrier to their effective access and use. The need for more common approaches to the development of data systems and services built upon data and information standards has been apparent for some time. [NOAA2007]

Significantly, standardization is both the first and the largest step on the road to system interoperability and, ultimately, GEO-IDE’s mission to create NOAA’s “integrated data environment”:

The lack of broad, uniform utilization of information technology (IT) standards ... is arguably the most acute factor contributing to the weakness of data integration within NOAA today. [GICO]

The success of standardization can be seen in many places, but perhaps none more clearly than the World Wide Web. Standard interfaces, protocols, languages, and—crucially—information types have enabled an incredible quantity and significant variety of information to be exchanged regardless of underlying differences in computer systems, user needs, or other variables.

While the Web example illustrates standardization within a particular environment, it also highlights another opportunity for standardization, namely bridging the gap between the highly specialized science community and the general Web user. For example, the ability to discover NOAA information using simple Web searches is a frequently mentioned goal. The application of standards, particularly semantic standards, is the avenue by which this and similar objectives may be achieved.

The importance of standardization is undeniable. It is one of the key tools available for addressing the information heterogeneity problem, and it has myriad indirect benefits in other areas.

6.2.2.2 Standardization Limitations

Without question, standardization is crucial to NOAA’s success. Nonetheless, it is important to understand its limitations.

Standardization is a simple concept, but it can be difficult to implement. Given the scope of NOAA information holdings, the standardization effort will require organization-wide collaboration and agreement, the participation and consensus of external parties, and a great deal of behavioral change. All of these constitute significant challenges.

Standardization will take time. Identifying and approving standards can be a lengthy process, and

implementing them involves even more time. A great deal of patience will be required—as will the ability to focus on long-term benefits and goals—in order to maintain the momentum of standards efforts.

There is also a point of diminishing returns with standards; it is possible to lose overall value by being too aggressive in the standardization effort. The proper balance between convention and diversity needs to be struck, as suggested in the DMIT's *Review of Standards Applicable to NOAA and Recommendations for Fast-track Consideration as Proposed Standards*:

Different standards are most suitable for different types of data. Furthermore, a variety of software is used by customers to analyze and visualize data. As a consequence, no single format could possibly satisfy the requirements of all users. However, a proliferation of formats is costly and inefficient for both NOAA and its users. To balance the needs for consistency and flexibility, users should be given the option to select (from a short list of standards) the format that best meets their requirements. [DMIT2006]

Finally, standards efforts can be easily derailed by a wide variety of tactical considerations, including schedules, funding considerations, performance expectations, changes in management priorities, and many others. The standardization process requires a level of stability in order to be successful. As noted previously, this stability has to be created from the top down, in the form of long-term objectives and support.

6.2.2.3 Summary

Standardization is a crucial tool for reducing information heterogeneity; it has the potential for significant reductions in cost and complexity, and substantial improvements in functionality, interoperability, and information utilization. Standardization has limitations, however, that must be acknowledged in order to properly set expectations and objectives. The ultimate success of any standardization effort depends on organization-wide buy-in, and widespread, top-down support.

6.3 Information Heterogeneity Impact Management

While information heterogeneity reduction is the highest-leverage approach to addressing the general problem, mitigating the impacts of information heterogeneity that cannot be eliminated is an essential ingredient of overall management efforts. This sub-section discusses the key elements of establishing an environment in which information heterogeneity can be managed as efficiently and effectively as possible. It is worth noting that these management efforts tend to interrelate and overlap. Each is important on its own, but together the benefits have the potential to be significantly compounded.

It is also important to emphasize that the items mentioned in Section 6.1 are all crucial to ongoing information heterogeneity impact mitigation efforts. These will not be repeated in this sub-section, though some similarities and overlap will be observed in the material that follows.

6.3.1 Roles and Responsibilities

The proper allocation of roles and responsibilities is a crucial element of effective, efficient information heterogeneity impact mitigation. In many cases, the question is not whether a particular capability should be provided or an activity performed, but rather identification of the right entity to discharge the responsibility.

6.3.2 Comprehensive Analysis

Once unnecessary information types have been identified and eliminated, it is imperative to do a comprehensive analysis of the remaining information, with an eye to identifying variations and commonalities, and identifying the significance of each within the various phases of the information lifecycle. For example, variations and commonalities may be relevant to some information management activities, but not to others.

This analysis should then be used as input to information heterogeneity impact management activities and future information heterogeneity reduction efforts. The analysis will have to become a “living document” that is updated with the addition of new information types (and potentially new information roles as well). Development and maintenance of this document should be the responsibility of an information governance board like the DMC.

6.3.3 Leverage Industry Trends and Approaches

The digital information management industry is evolving rapidly, particularly with regard to information preservation. Many opportunities exist for NOAA and its information management efforts to take advantage of the work being done by others. As an example, the European Space Agency is currently heavily involved with the standardization of archive holdings and the adoption of a single standard representation for information across a variety of uses. Much of this work is tied to international standards like the OAIS-RM, or conventions that may well become international standards in the future (for example, the Standard Archive Format for Europe - SAFE). The opportunity for NOAA to leverage these and similar efforts is substantial.

6.3.4 Generic IT Mechanisms

IT mechanisms to support information heterogeneity management include capabilities to identify specific information variants, encapsulate and abstract these variations when possible, support controlled vocabularies, transform information, and many others. The objective of most of these efforts is to transition system enhancement from an emphasis on software development to software configuration.

As a rule, IT systems should provide generic mechanisms that will be used by information specialists in their information management roles. Architecture and design artifacts to capture and manage semantics provide a convenient example. IT systems will implement these mechanisms, but their population and utilization should fall primarily to information stewards.

The development of “plug-in” architectures provides another useful example. Plug-in architectures

enable an “open source” model for information-specific software. In particular, this approach makes it possible to transition the information-specific aspects of software development to information specialists while focusing IT systems and their personnel on their appropriate role of providing IT support.

Many other examples could be cited, across the entire spectrum of information management activities. The important concept, however, is the separation of pure IT mechanisms and information-specific capabilities to support proper role and responsibility allocation and make it possible to enhance and evolve systems by configuration rather than development.

6.3.5 Information-Specific Mechanisms

A number of mechanisms that exhibit varying degrees of information specificity may be employed to help manage the impacts of information heterogeneity.

For example, “middleware” can be developed that provides common, generic interfaces to multiple information types. The portions of these software layers that interface directly with heterogeneous types are information-specific, but they provide a generic interface on the application side that can be used as if the underlying information were homogeneous.

Data description languages provide another means of managing information specificity, in that they provide the ability to describe information representations in a generic way. Tools can then be developed that operate on heterogeneous information types using these generic descriptions. Adding new types requires simply writing new descriptions – a much lower cost activity than software development.

Finally, structural data types provide a means to aggregate information types by shared structural characteristics. For example, swaths, grids, profiles, and the like tend to have common semantics, operations, and utilization patterns. By creating systems that deal with the aggregated type, simpler and more efficient software development can be achieved.

There are a great many alternatives for generic mechanisms. Together, they provide one of the most powerful approaches available to managing the impacts of information heterogeneity.

6.3.6 Standardization

Standardization is not only a key means of reducing information heterogeneity, but also a fundamental aspect of managing its impacts. For example, the establishment of an information lifecycle reference model is a form of semantic standardization that will greatly facilitate information heterogeneity management efforts. The OAIS-RM has been very effective in this regard for the preservation phase of the information lifecycle; an end-to-end reference model would likely have even greater benefit.

Standardization is the foundation of interoperability. As discussed, interoperability is much more than simple inter-system communication. Thus, standardization efforts are essential for not only

system interoperability, but also interoperability among people, organizations, processes, and information. Unsurprisingly, there is a significant amount of overlap across these categories. For example, controlled vocabularies are essential not only to support human communication, but also to support machine-encoded semantics efforts.

6.4 Summary

The information heterogeneity problem must be attacked by both reducing the amount of overall information heterogeneity and concurrently creating an environment in which remaining information heterogeneity can be managed. Standardization is a key element in both efforts, but has limitations that prevent it from being a complete, exclusive solution. A variety of means must be employed in both reduction and management efforts. The common threads among all of these are the importance of top-down organizational support for information heterogeneity management efforts, and changes in behavior and culture that prioritize information heterogeneity reduction and impact management.

Section 7.0 Recommendations

This section provides recommendations regarding information heterogeneity reduction and impact management. The following sub-sections include specific recommendations for NOAA, GEO-IDE, the NOAA archives, NOAA information stewards, and NOAA IT systems. The end-to-end information heterogeneity analysis recommended in the following material should include consideration for all NOAA information management entities, and provide recommendations for each.

These recommendations are not intended to be complete or exhaustive, but rather representative of the kinds of things NOAA and selected NOAA entities must do to address the information heterogeneity problem. A number of these recommendations include the execution of studies, development of plans, and the like. It should be a primary objective of these efforts to identify in detail the steps needed for information heterogeneity reduction and impact management.

It should be noted that efforts to acknowledge and address information heterogeneity and its impacts do exist within NOAA, and have for some time. Evidence of this can be seen in standards efforts, interoperability initiatives like GEO-IDE, emphasis on “systems of systems,” the existence and efforts of the DMIT, and many other places. While these efforts are necessary, they lack some of the key ingredients necessary to insure success. The most important of these is that NOAA has not yet made information heterogeneity reduction and impact management a part of its organizational culture. Until it does so, current efforts can only address the symptoms of the problem; they cannot attack the problem itself.

7.1 NOAA

NOAA has the primary responsibility for information heterogeneity reduction. While its organizational components can, and must, contribute in this area, NOAA is the only entity with the authority and scope to improve the situation across the agency.

It is important to note that—as NOAA’s “data and information service”—NESDIS has a key role in addressing the information heterogeneity problem. With responsibilities that include operation of the NNDCs, NESDIS is at the center of many of NOAA’s information-related issues. NESDIS must assume significant responsibility for, and be very active in addressing, the recommendations made explicitly for NOAA that follow.

NOAA should:

7.1.1 Promote Institutional Awareness and Commit to Mitigation Efforts

Chief among NOAA’s responsibilities is the promotion of awareness of information heterogeneity across the organization, and an unequivocal, cross-cutting commitment to reducing it and managing its impacts. This commitment must figure into strategic and tactical objectives, program and project

planning, performance assessments, funding decisions, and the various other mechanisms that influence behavior. Above all, NOAA has to establish an environment in which information heterogeneity is a highly visible issue, and in which mitigating its impacts is a high priority. This requires a level of cultural and behavioral change that demands consistent, long-term, top-down commitment and support.

As an example, it has often been easier for information stewards to obtain funding for IT system development, operation, and maintenance than information stewardship. This makes sense in a number of ways, since IT systems are a lot easier to describe, quantify, cost, and account for over time. Unfortunately, this has led to obvious and disadvantageous role and responsibility shifts. If the only way information stewards can achieve full funding is through IT projects, then they will clearly emphasize IT projects. In the end, both information stewardship and IT projects suffer. Only by funding information stewards **for** information stewardship will NOAA policy and objectives be aligned. This is, of course, just one example, but the concept applies quite broadly. People and organizations will behave in accordance with the ways in which they are supported and measured. It is essential for NOAA to support and measure its organizational components and efforts in complete consistency with its strategic objectives.

7.1.2 Emphasize the Importance of Information Governance

A variety of NOAA entities share responsibilities for various aspects of information governance, some with responsibility for information systems, some for integration of systems or information, and so on. The NOSC, DMC, DMIT, GEO-IDE, NAAT, and CIO Council are examples.

The DMC has primary responsibility for information governance across the entire information lifecycle. DMC efforts should be emphasized and expanded and, in particular, emphasis should be placed on information heterogeneity.

Moreover, information governance should be an active, integrated aspect of all information management efforts across NOAA. The importance of consistent involvement, strong leadership, and integrated efforts across the various information management entities can not be overemphasized.

7.1.3 Commission an Information Heterogeneity Cost/Benefit Study

Quantifying the costs and benefits of information heterogeneity reduction and impact management is an essential early step. Similar studies in other fields demonstrate their viability and usefulness, as well as their importance in establishing an understanding of the importance and scope of the task.

Since some of the prior studies were performed in connection with or on behalf of NIST, it may well be worthwhile to establish relationships with the appropriate NIST personnel as a means to better scope the activity, set objectives, and evaluate potential benefits.

7.1.4 Develop an End-to-End Information Lifecycle Reference Model

The establishment of a common language for the entire information lifecycle is a mandatory step in addressing the information heterogeneity problem. The OAIS-RM does this in the context of long-term preservation, and those who have adopted it are starting to see significant benefits. NOAA must extend the concept to include the entire information lifecycle so that all of the parties involved in NOAA information management can communicate succinctly, precisely, and unambiguously. The establishment of an end-to-end information lifecycle reference model is the primary means by which this can be achieved. A key element of the reference model should be the identification of roles and responsibilities.

7.1.5 Aggressively Push Standards Adoption

Standards adoption has to be pushed from both the top down and the bottom up. NOAA must ensure that standards adoption is a key aspect of its ongoing commitment to information heterogeneity reduction and impact mitigation efforts, and that all of its various sub-organizations, projects, and programs are properly motivated to make substantive progress toward standardization.

7.1.6 Make Producers Part of the Solution

NOAA has to aggressively push its information producers to be part of the solution to the information heterogeneity problem. Historically, producers have often had a limited view of their customer base. For example, so-called “real-time” users have been the sole focus of information production efforts, while long-term preservation was an afterthought – if it was considered at all. Of course, as with information stewards, motivating producers to change their behavior can only be accomplished in an environment that supports the desired changes. NOAA has to tie producer funding and other programmatic decisions to producer contributions to information heterogeneity management. In turn, producers have to start educating their customers about the overall benefits of information heterogeneity reduction and impact management, and encouraging and enlisting customer support. Involving the producers is a key first step to minimizing heterogeneity increases associated with future campaigns.

7.1.7 Support Information-Related Role and Responsibility Allocation

One of the most difficult aspects of role and responsibility reallocation is aligning funding and other organizational support with the reallocations necessary to support information heterogeneity reduction and impact mitigation. As noted in the example regarding funding for information stewardship, in order to get the “right people doing the right things,” NOAA must commit to changing the way it does business internally, and support this commitment consistently in its actions.

7.1.8 Promote Information Management “Meta-Initiatives”

NOAA should prioritize the extension of existing “inventory cataloging” and metadata management systems like CasaNOSA and NMMR to help identify the scope and depth of information heterogeneity. Systems like these are key tools in the ongoing effort to reduce information

heterogeneity and manage its impacts.

7.1.9 Make Hard Decisions Regarding Information Retention

As noted previously, one very clear way of reducing information heterogeneity is to eliminate unnecessary information types. While NOAA has pursued this to some extent via EDAPT, the outcome was the development of a process for identifying information that should be preserved. The process, however, has not been comprehensively applied to NOAA information holdings. The DMC's responsibilities should include overseeing the application of this process in an effort to eliminate unnecessary information types and associated costs and complexity.

7.1.10 Coordinate or Consolidate Ongoing Efforts

A number of ongoing efforts are doing excellent information heterogeneity-related work. These include IOOS DMAC/DIF, SNAAP, GEO-IDE, CLASS, DMIT, NAAT, and others. These efforts should be coordinated and/or consolidated to promote a more comprehensive set of solutions and eliminate potentially redundant effort.

7.1.11 Engage the National and International Communities

As significant as information heterogeneity is for NOAA, its impact on national and international "system of systems" efforts like USGEO and GEOSS will be even more profound. NOAA has the opportunity to take a lead role in championing the importance of information heterogeneity within these efforts, and should strongly consider doing so. That said, if NOAA is to take advantage of this opportunity it must first take strong steps and make demonstrable progress toward addressing information heterogeneity internally.

Moreover, NOAA should engage with others in the national and international communities regarding information heterogeneity and its management. Considerable effort is being invested, and there are significant opportunities for NOAA to leverage and extend the work that is being performed. As an example, both the EC and ESA have made consistent efforts to work on the information heterogeneity problem, with the former passing legislation called "INSPIRE" that mandates spatial data standardization, and the latter—among many other initiatives—standardizing on a single representation for preserved information (Standard Archive Format for Europe) and moving to have information produced natively in this representation.

7.2 GEO-IDE

GEO-IDE's interoperability mandate places it at the center of many information heterogeneity-related issues. Somewhat paradoxically, the effort is not scoped to include information heterogeneity reduction or impact management activities.

The most important recommendation for GEO-IDE is to identify and document the various obstacles

to interoperability, publicize these to NOAA management, and support the provided recommendations for NOAA that will help establish an environment in which interoperability can be achieved. Only after this environment is established can GEO-IDE have a reasonable chance at success.

7.3 NOAA Archives

Recommendations regarding archive responsibilities for information heterogeneity management apply primarily to information stewards and supporting systems like CLASS. It is worth noting, however, that the NOAA archives have a key role in that they are at the intersection of information-specific efforts (by information stewards) and generic IT efforts. The establishment of archive roles and responsibilities across information stewards and supporting systems and/or organizations will greatly impact the success of the entities and activities that support information preservation.

Also, NOAA archives should initiate interactions with other national and international archives to leverage their efforts, learn more about overall industry trends, and become active participants in international activities like standards initiatives.

7.4 NOAA Information Stewards

NOAA information stewards are in a unique position to help mitigate the information heterogeneity problem because they have the greatest concentration of information-specific expertise across the agency. NOAA information stewards should:

7.4.1 Be a Key Element of Standardization Efforts

Since information stewards are the repositories of much of the information-specific knowledge and expertise in the organization, they have a central role in standardization efforts. Information steward input and participation are essential to the development of syntactic and semantic standards for information types, the establishment of a NOAA-wide information lifecycle reference model, and to virtually all other information-related NOAA standardization efforts.

7.4.2 Push for Role and Responsibility Transitions

Information stewards should support the transition of information-specific roles and responsibilities. These roles and responsibilities need to be transitioned from IT projects and systems, whose focus should be solely on IT support, to the information stewards. There are many instances across NOAA where this transition would greatly improve efficiency and effectiveness.

7.4.3 Support Mechanism Specification and Development

Information stewards play a key role in the specification and development of information-specific

mechanisms to manage information heterogeneity impacts. The development of “middleware” that provides generic interfaces to selected heterogeneous information types is a good example, as is the specification and development of semantics management capabilities and “plug-ins” that tailor generic IT mechanisms for specific information types.

7.4.4 Serve as a Bridge between Producers and Consumers

Information stewards are uniquely positioned to bridge the needs of producers and consumers, and must use this opportunity to publicize, support, and execute information heterogeneity reduction and impact management initiatives. No group within NOAA has more direct relationships with those who produce and consume NOAA information.

7.4.5 Coordinate Their Efforts

Information stewards are driving, or contributing to, many ongoing and proposed efforts that affect information heterogeneity reduction and impact management; these efforts must be coordinated. Some progress has been made in this regard, but both cultural and organizational issues (e.g., funding) place information stewards in as much a competitive posture as a cooperative one. These barriers have to be eliminated so that motivations are aligned to support cooperative and integrated information management efforts by all of NOAA’s information stewards. NOAA management has a key responsibility to create an environment in which this is not only a possible, but certain, outcome.

7.4.6 Take Advantage of the Information Lifecycle

Information stewards are in a prime position to capitalize on the information lifecycle to help reduce information heterogeneity and mitigate its impacts.

For example, information stewards provide important inputs to the questions of what to archive and what to sunset. Aligning their inputs on these questions with information heterogeneity considerations has powerful potential impacts.

Also, information stewards will decide which archive holdings to transform, when to transform them, and the manner and details of transformation. Transformations are a natural part of the information preservation cycle, and provide great opportunities to reduce information heterogeneity. Information stewards can have substantial impact on information heterogeneity by making its reduction a key objective of the transformation process.

Many examples could be cited, but the point is that information stewards are well-positioned to positively impact information heterogeneity by simply considering it a primary goal at decision points throughout the information lifecycle.

7.4.7 Take a Much Stronger Role in the IT Requirements Process

One of the biggest hurdles for IT systems is that information-related requirements are often sparsely

and/or incompletely specified. Far too often, requirements focus on obvious considerations like data volumes, performance, and the like, while many others with significant long-term impacts are overlooked. The information stewards' expertise must be collected and translated into clear, complete requirements so that IT system developers can build the *right* generic IT tools and mechanisms to support information management objectives. Assuming responsibility for information-related requirements is a fundamental, essential aspect of the information-specific role and responsibility transition the information stewards must undergo.

7.5 IT Systems

IT system roles do not include direct participation in information heterogeneity reduction. However, IT system owners and developers can—and must—support information heterogeneity impact management efforts.

7.5.1 Publicize Information Heterogeneity and Its Impacts

As noted earlier, in some respects IT systems have been the place where the information heterogeneity “buck” stops. This situation has motivated much of the thinking about information heterogeneity reflected in this paper, including its prominent place in overall NOAA IT system design and implementation and all of the mechanisms and other artifacts important to mitigating its effects.

IT system owners and developers must continue to play a key role in the “bottom up” aspects of information heterogeneity management, to champion the issue throughout NOAA, and to identify conceptual approaches to mitigating the problem. IT system developers can contribute to these efforts by continuing to monitor industry research and best-practices on heterogeneity, and providing recommendations for addressing information heterogeneity. Further, IT system development efforts can also serve as “early warning systems” for potential increases in heterogeneity through earlier involvement in new campaign planning.

7.5.2 Develop Generic IT Mechanisms

IT system developers must continue to build generic IT mechanisms that facilitate the management of information heterogeneity that cannot be eliminated. These mechanisms must allow producers, information stewards, consumers, and other members of the information preservation community to manage the impacts of information heterogeneity as effectively and efficiently as possible.

7.5.3 Actively Support and Push for Role and Responsibility Transitions

Currently, NOAA IT systems include substantial amounts of information specificity. The mix of information-specific artifacts with a “generic IT” mission causes a variety of problems, from development priorities to staffing mix. It is essential for IT system owners and developers to transition information-specific responsibilities to information stewards so that each party can more

effectively contribute to information management activities.

This transition will take time, and is heavily dependent on IT systems providing the generic IT mechanisms that make it possible. While implementation of these mechanisms is crucial, the big picture—role and responsibility transition—must be kept in mind, and pursued, from the start.

7.5.4 Rigorously Pursue and Implement Standardization

Standardization must continue to be a primary focus for NOAA IT system development and evolution efforts. Information system developers must continue to aggressively implement IT standards, and to work with information stewards to understand information standards sufficiently that they can design and implement the generic mechanisms that will support them. IT system owners and developers must also continue to support and promote the importance of standardization throughout NOAA and participate in appropriate standards efforts.

7.6 Summary

Information heterogeneity is a NOAA-wide (world-wide, in fact) problem; addressing it within NOAA requires a wide variety of efforts by entities throughout the organization. Every participant in the information lifecycle has a role in reducing information heterogeneity, managing its impacts, or both. The first, most important, step is for NOAA to institutionalize an understanding of information heterogeneity and its impacts, and to commit fully to the establishment of an environment and a set of tactical and strategic objectives that facilitate its reduction and impact mitigation.

Section 8.0 Conclusion

Information heterogeneity and its impacts are ubiquitous across NOAA, increasing virtually unchecked with the addition of new and more complex information and increased functional requirements, and being compounded by other major drivers like increasing data volumes. Moreover, information heterogeneity is the key barrier to one of NOAA's most important strategic objectives, the creation of an "integrated data environment." Significantly, information heterogeneity is at the root of the interoperability question, as it is the reason interoperability is needed in the first place.

The impacts of information heterogeneity are increasing dramatically faster than they are currently being addressed. Absent aggressive mitigation efforts, the outcome is clear: information heterogeneity will soon make it impossible for NOAA and its information management initiatives to fulfill expectations with available resources; information heterogeneity in the current context will soon be a "showstopper."

Fortunately, information heterogeneity and its impacts can be mitigated through a variety of means, primarily by reducing the overall amount of information heterogeneity and establishing an environment in which the impacts of remaining information heterogeneity may be efficiently and effectively managed. It is essential, however, to understand that information heterogeneity is an organization-wide problem, solutions to which necessitate cross-cutting cultural and behavioral changes in addition to the judicious application of technology.

NOAA is in the highly desirable position of possessing a great wealth of information essential to the study of Earth and charting the course for future generations. This privilege, however, comes with great responsibility to ensure the usability and value of these precious resources. Vigorously addressing the information heterogeneity problem is an essential step for NOAA to continue providing the societal benefits needed now and in the future.

Appendix A. Acronyms

<u>Acronym</u>	<u>Definition</u>
AVHRR	Advanced Very High Resolution Radiometer
BUFR	Binary Universal Form for the Representation of meteorological data
CAD	Computer Aided Design
CF	Climate and Forecasting
CIO	Chief Information Officer
CLASS	Comprehensive Large Array-data Stewardship System
COARDS	Cooperative Ocean/Atmosphere Research Data Service
COPB	CLASS Operations Planning Board
CORS	Continually Operational Reference Stations
COTR	Contract Officer's Technical Representative
CRC	Cyclic Redundancy Check
DAARWG	Data Archive and Access Requirements Working Group
DGP	Diversified Global Partners
DIF	Data Integration Framework
DMAC	Data Management and Communications
DMC	Data Management Committee
DMIT	Data Management Integration Team
DMSP	Defense Meteorological Satellite Program
DOI	Digital Object Identifier
EC	European Commission
EDAPT	Environmental Data Archive Policy Team
EOS	Earth Observing System
ESA	European Space Agency
GCMD	Global Change Master Directory
GEO-IDE	Global Earth Observation Integrated Data Environment
GEOSS	Global Earth Observation System of Systems
GICO	GEO-IDE Concept of Operation
GMES	Global Monitoring for Environment and Security
GOES	Geostationary-orbiting Operational Environmental Satellites
GO-ESSP	Global Organization for Earth System Science Portal
GRiB	Gridded Binary
HDF	Hierarchical Data Format
IOOS	Integrated Ocean Observing Systems
INSPIRE	Infrastructure for Spatial Information in the European Community
IT	Information Technology
JPEG	Joint Photographic Experts Group
M&O	Maintenance and Operations (alternative form of "O&M")
MetOp	Meteorological Operational Satellite Program
MODIS	Moderate Resolution Imaging Spectroradiometer
NAAT	NOAA Archives Architecture Team
NAO	NOAA Administrative Order

NASA	National Aeronautics and Space Administration
NCDC	National Climatic Data Center
NESDIS	National Environmental Satellite, Data, and Information Service
NetCDF	Network Common Data Format
NIST	National Institute of Standards and Technology
NMMR	NOAA Metadata Manager and Repository
NNDC	NOAA National Data Center
NOAA	National Oceanic and Atmospheric Administration
NOSA	NOAA Observing System Architecture
NOSC	NOAA Observing System Council
NPOESS	National Polar-orbiting Operational Environmental Satellite System
NPP	NPOESS Preparatory Program
O&M	Operations and Maintenance
OAIS	Open Archival Information System
OAIS-RM	OAIS Reference Model
OLS	Operational Linescan System
PB	Petabyte
POES	Polar-orbiting Operational Environmental Satellite
QMO	Quality Management Organization
ROI	Return On Investment
SAFE	Standard Archive Format for Europe
SNAAP	Simple NOAA Archive Access Portal
SSM/I	Special Sensor Microwave Imager
URL	Universal Resource Locator
USGEO	U.S. Group on Earth Observations
USGS	United States Geological Survey
XML	Extensible Markup Language