ERDC/TEC SR-10-1

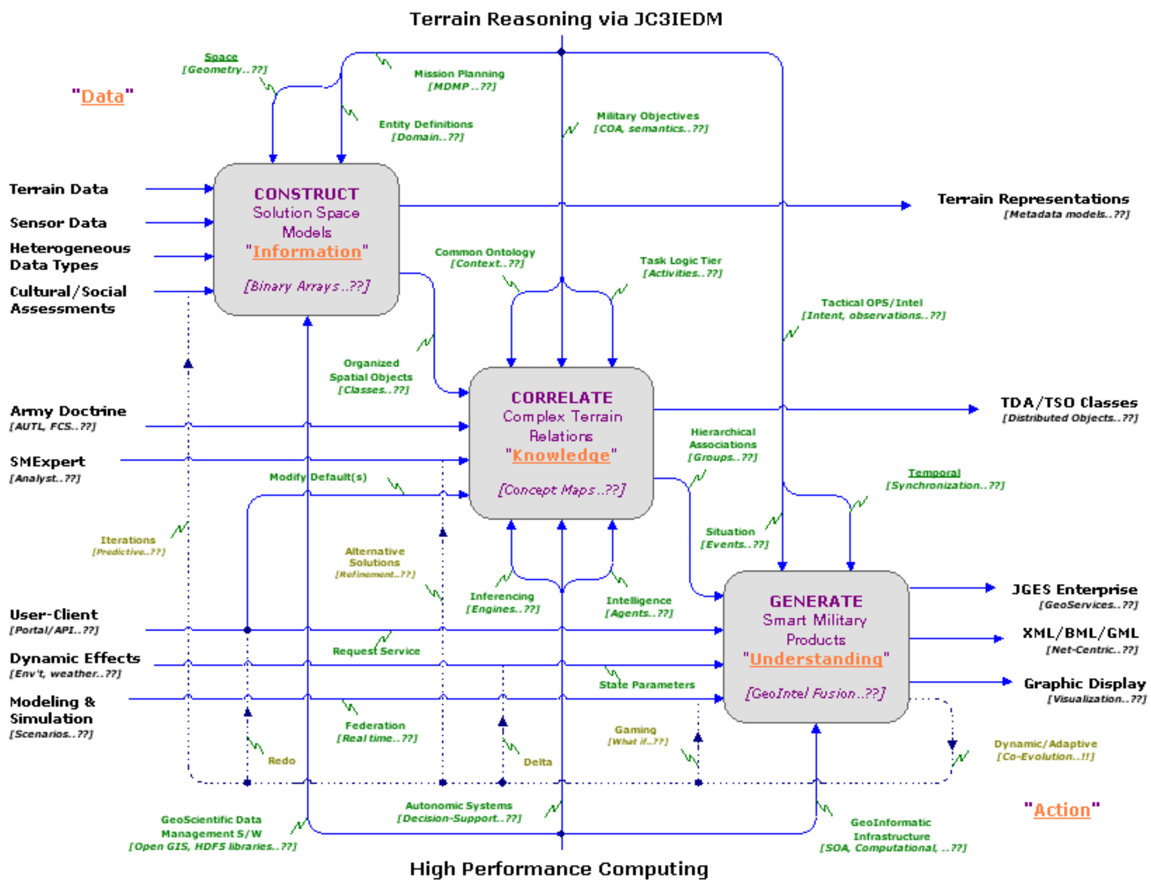
Topographic Engineering Research Center

US Army Corps of Engineers®
Engineer Research and Development Center

Transforming the Geocomputational Battlespace Framework with HDF5

Mike Folk, Peter Cao, and John A. Nedza

August 2010



(Image courtesy of Analytical Graphics, Inc.)

ERDC/TEC SR-10-1
August 2010

Transforming the Geocomputational Battlespace Framework with HDF5

John A. Nedza

Topographic Engineering Center
U.S. Army Engineer Research and Development Center
7701 Telegraph Road
Alexandria, VA 22315

Mike Folk and Peter Cao

The HDF Group
1901 S First Street, Suite C-2
Champaign, IL 61820

Final report

Approved for public release; distribution is unlimited.

Prepared for U.S. Army Corps of Engineers
Engineer Research and Development Center
Alexandria, VA 22315-3864

Under Broad Area Announcement (BAA) Contract Number W9132V-08-C-0028

Abstract: Success in the modern Battlefield depends on effective management of data. Every operation is unique, making it impossible to build systems that address all of the needs of a specific mission in advance. Information needs arise under intense time pressure, and the available information is often incomplete or uncertain. Data entities vary enormously in scale and resolution, and can exhibit a great deal of heterogeneity. These challenges will only intensify in the future, as networks of sensors increasingly collect and transmit huge amounts and varieties of valuable data, scaling from synoptic images to the vital signs of individuals.

The complexities inherent in mission operations make the information management task an immense challenge, one which must be addressed in part by focusing on how information is organized, integrated, accessed, and analyzed. Toward that end the goals of this investigation are to identify the role that HDF5 can play as a data management platform for Battlefield military operations, to demonstrate the use of HDF5 visualization tools to present operational data, and to identify a research and development plan to develop a prototype geoinformatic data management system based on HDF5.

Contents

Figures and Tables

Figures

Tables

Preface

The growth of networks, sensors, and other data sources has increased the variety and scale of data to be integrated and explored in Battlefield decision making. Traditional technologies, such as Geographic Information Systems (GIS), and Relational Database Management Systems (RDBMS), have proven inadequate to handle many of these new requirements. In recent years, the Army Engineer Research and Development Center (ERDC) Topographic Engineering Center (TEC) has explored novel approaches to these rapidly changing needs. This work led to the discovery of the HDF5, a technology that has proven effective for addressing many of these same needs in mission-critical applications in almost every scientific and engineering discipline, including some applications whose characteristics are very similar to military mission operations.

This paper explores the role that HDF5 can play as a platform for managing Battlefield data. Chapter 1 explains the need for a new approach to Battlefield data management, and provides an overview of the approach taken in the paper. Chapter 2 describes HDF5 and its applications at a sufficient level of detail to enable the reader to understand the capabilities HDF5 brings to Battlefield data management. Chapter 3 shows through an extended example how the variety of data encountered in a Battlefield can be readily accommodated by HDF5, how “concept maps” can be created to provide a clear framework for thinking about and working with the data, and how a common HDF5 viewing and editing tool can be readily adapted to provide a powerful interface to the underlying data. Chapter 4 identifies research areas requiring further investigation to adequately instantiate the HDF5 data structure for practical application. Chapter 5 explores the implications of incorporating a very different kind of data within the Battlefield information space, namely the types of data needed for the human, social, cultural, and behavior modeling needed to address “wicked” problems. Chapter 6 concludes the study by identifying benefits to the Army of this approach, identifying a path forward from GIS to “high performance GIS,” and recommending areas for future work.

The authors gratefully acknowledge many colleagues who participated in this study. Lloyd Hauck (ERDC-TEC), who guided the funding and management of the project, as well as providing essential insights and reviewing of the manuscript. Chapter 5 is based almost completely on the insightful work of Mike Stein (BAE Systems), who also gave generously of his time in explaining the socio-cultural dimension. Bill Meyer (ERDC-CERL) also provided valuable intellectual input for Chapter 5. Vanisha Taylor and Anne Jennings (The HDF Group) handled contractual matters with skill and timeliness, and Ruth Aydt (The HDF Group) was a valuable sounding board on the exposition of matters related to HDF5 technologies.

COL Gary E. Johnston was Commander and Executive Director of ERDC. Dr. Jeffery P. Holland was Director.

Acronyms and Abbreviations

BAA	Broad Area Announcement
AOI	Area Of Interest
API	Application Programming Interface
ASCI	Accelerated Strategic Computing Initiative
BC	Battle Command
BIM	Building Information Model
BML	Battle Management Language
BTRA	Battlespace Terrain Reasoning and Awareness
COA	Course of Actions
CERL	Construction Engineering Research Laboratory
EOS	Earth Observing System
ERDC	Engineer Research and Development Center
ESRI	Environmental Systems Research Institute
FGDC	Federal Geographic Data Committee
GDAL	Geospatial Data Abstraction Library
GIS	Geographic Information System
GeoPDF	Geographic Portable Document Format
GRED	Geospatial Research and Engineering
GUI	Graphical User Interface
HDF	Hierarchical Data Format
HDF5	Hierarchical Data Format Version 5
HPC	High Performance Computing
HPGIS	High Performance Geographic Information Systems
IDL	Interactive Data Language
IPB	Intelligence Preparation of Battlefield
MDMP	Military Decision Making Process
MIME	Multipurpose Internet Mail Extensions
NITF	National Imagery Transmission Format
OCOKA	Observation, Cover and/or Concealment, Obstacles, Key Terrain, Avenues of Approach
OSE	Open Systems Engineering
RECON	Reconnaissance
SMMP	Semantic Metadata Mapping Procedures
TEC	Topographic Engineering Center
UTP	Urban Tactical Planner
UAV	Unmanned Aerial Vehicle
XML	Extensible Markup Language

1 Technical Report and Discussion

A new perspective on battlefield data management

Background

Military operations require coordination among diverse groups and involve an increasing variety of data sources, data types, and applications in the field. Every mission is unique and requires novel combinations of information, making it impossible to anticipate and build integrated systems that address the needs of a specific mission in advance. Data access and integration frequently occur under intense time pressure. Information is often incomplete or uncertain. Images can vary enormously in scale and resolution. There can be a great deal of heterogeneity in the types of information. In the future, these information management challenges will be multiplied, as networks of sensors increasingly collect and transmit huge amounts of data, from images to the vital signs of individuals.

Battlespace Terrain Reasoning and Awareness

A prime example of the data challenge for military operations, and a focus of the proposed research, is the Battlespace Terrain Reasoning and Awareness – Battle Command (BTRA-BC). The functional mission of the BTRA-BC is

"to increase the effectiveness and agility of Battle Command (BC) and the Military Decision Making Process through the application of geo-environmental data, information and knowledge, across the greatest extent possible across of the force" (*http://www.agc.army.mil/btra/index.html)*.

BTRA depends fundamentally on our ability to effectively ingest, manage, , exploit, visualize, and disseminate a daunting volume and variety of digitally represented raw data, information, knowledge, and understanding. The process of going from raw data to battle ready information requires us to be able to access, organize, and integrate data occurring in a wide range of sizes and information density.

Figure 1 illustrates the process of converting raw data to battle ready data products. At the upper left (“Data”) are the maps, terrain data, sensor data, cultural assessments, and other data that is the raw material for the terrain reasoning process. These products come from many different sources and in different formats. Raw data products are collected and integrated to construct “Information” products, such as natural obstacles, roads, weather scenarios, and cultural features. Both the data and information products for a single military task can contain many gigabytes of data, or more.

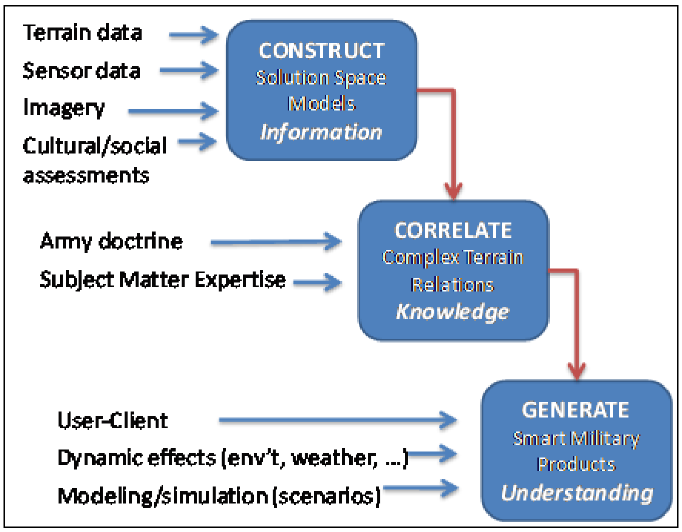


Figure 1. Converting raw data to battle ready data products for terrain reasoning.

Information products are combined and *correlated* based on Army Doctrine, the military decision-making process, gathered intelligence, and other subject matter expertise to create “Knowledge” products, or complex terrain relations, such as avenues of approach, battle positions, and potential routes. Knowledge products are modest in size, perhaps in the megabytes.

The final stage of information integration and fusion creates products suitable for the performance of specific military tasks or actions. The resultant “Understanding,” or “smart military products” require input from the user-client, real-time dynamic information such as weather effects, and scenarios produced by simulation models and gaming. These products are information rich but relatively small, measuring in the kilobytes.

Data heterogeneity – beyond traditional GIS

Figure 1 illustrates the wide assortment of data that needs to be managed in a Battlespace application. Because much of this data has spatio-temporal components, systems designed for these applications commonly are based largely on geographic information system (GIS) technologies. GIS technologies offer excellent tools for the queries and data analysis of geospatial data involved in Battlefield decision making. And yet, the explosion of data sources and data volumes is ushering in a new generation of expectations for GIS.

With traditional geographical information technologies, responses can be slow, and the ability to handle dramatic changes in scale, such as image sizes that vary by orders of magnitude, is limited. GIS implementations often are frequently tied to a particular computing platform. Traditional systems also do not handle very well many important types of data that are critical to military operations, such as audio, video, spreadsheets, real-time sensor data, acoustic data, chat transcripts, and weather scenarios. These different types of data typically are found in many different data formats, and the tools that work with the data are equally varied.

Other observers have spoken to the need to re-examine our concepts about the scope and use of geospatial data. In “Process Models and Next-Generation Geographic Information Technology” Paul M. Torrens writes,

> “Much of the potential for advancing geographic information technology stems from the ability of GIS to interface with other processes and related informatics through complementary process modeling schemes. The early precursors of this interoperability are already beginning to take shape through the fusion of GIS and building information models (BIMs). BIMs offer the ability of urban GIS to focus attention on a much finer resolution than ever, to the scale of buildings’ structural parts and their mechanical systems. GIS allows BIMs to consider the role of the building in a larger urban, social, geological, and ecosystem context. When process models are added to the mix, the complementary functionality expands even farther. Consider, for example, the uses of a GIS that represents the building footprints of an entire city but can also connect to building information models to calculate the energy load of independent structures for hundreds of potential weather scenarios…” (Torrens 2009)

In an August 2007 column for *GeoWorld* titled "Innovation Drives GIS Evolution," Joseph K. Berry (2007) speaks to the need to manage new varieties of geospatial data:

> The bulk of the current state of geospatial analysis relies on "static coincidence modeling" using a stack of geo-registered map layers. But, the frontier of GIS research is shifting focus to "dynamic flows modeling" that tracks movement over space and time in three-dimensional (3D) geographic space. But a wholesale revamping of data structure is needed to make this leap.
>
> The new geo-referencing framework provides a needed foothold for solving complex spatial problems, such as intercepting a nuclear missile using supersonic evasive maneuvers or tracking the air, surface and groundwater flows and concentrations of a toxic release. While the advanced map analysis applications coming our way aren't the bread and butter of mass applications based on historical map usage (visualization and geo-query of data layers), they represent natural extensions of geospatial conceptualization and analysis …*built upon an entirely new set analytic tools, geo-referencing framework and a more realistic paradigm of geographic space.*

A new approach to Battlefield data management based on HDF5

Thus, a critical aspect of Battlefield data management is that current approaches can be inadequate in meeting the requirements of speed, scalability, platform portability, heterogeneity, and geoprocessing. This combination of requirements makes information management a massive task, which must be addressed in part by focusing on how information is organized, integrated, accessed, and analyzed.

The ERDC-TEC has for several years been looking for a unified approach to address this combination of requirements, including following the development of scalable data management software for scientific and engineering data, most notably exemplified by the Hierarchical Data Format (HDF) and supporting technologies. HDF5, the flagship HDF package developed by the National Center for Supercomputing Applications (NCSA), the Department of Energy's Accelerated Strategic Computing Initiative (ASCI), and NASA's Earth Observing System (EOS), was created first in 1998 to address precisely these same requirements.

Our investigations have convinced us that HDF5 has the potential to be the foundation upon which to build a comprehensive system for heterogeneous data management and analysis in urban mission operations. The goal of the proposed research is to test that idea.

In this paper, we describe the role that HDF5 can play as a data management platform for urban mission operations, demonstrate the use of HDF5 visualization tools to present operational data, and identify a research plan to develop a prototype Battlefield data management system based on HDF5.

Although the focus here is upon the Battlefield, HDF5 embodies the data structures and access software to efficiently organize, manage, and access virtually every type of information structure encountered in urban missions, and as such could prove to be of equal value in related areas, such as natural disasters.

Example: urban operations and data integration

Consider a simple example in which three types of tools and data are used together in an operation over urban terrain:

- The Urban Tactical Planner (UTP), which provides a quick and informative overview of city-scale terrain in the form of maps, imagery, and elevation data;
- Observation Cover/Concealment Obstacles Key terrain Avenues of Approach (OCOKA) based analytics, such as Battlespace Terrain Reasoning and Awareness (BTRA) engines, which process data from a Course of Action (COA) analysis;
- Weather simulations that use a variety of probable weather scenarios, plus current conditions, to assess the effects of different weather conditions on an operation.

Independent formats and operations. Typically, each of these tools operates independently with its own data and produces its own results, but ultimately, the findings from these tools need to be integrated to analyze the results, to decide on a course of action, and to act. Figure 2 illustrates this process. Each tool has its own data requirements, and data are converted (Figure 2 (a)) to whatever data structures and formats they are designed to work with. Often unique visualization and analysis tools (b) are implemented for each application. When common tools might be used, there is a need for those tools to adapt to the data structures and formats of the individual applications. Ultimately, decisions are made and actions taken (d) based on the integration (c) of the collective knowledge from the various tools.

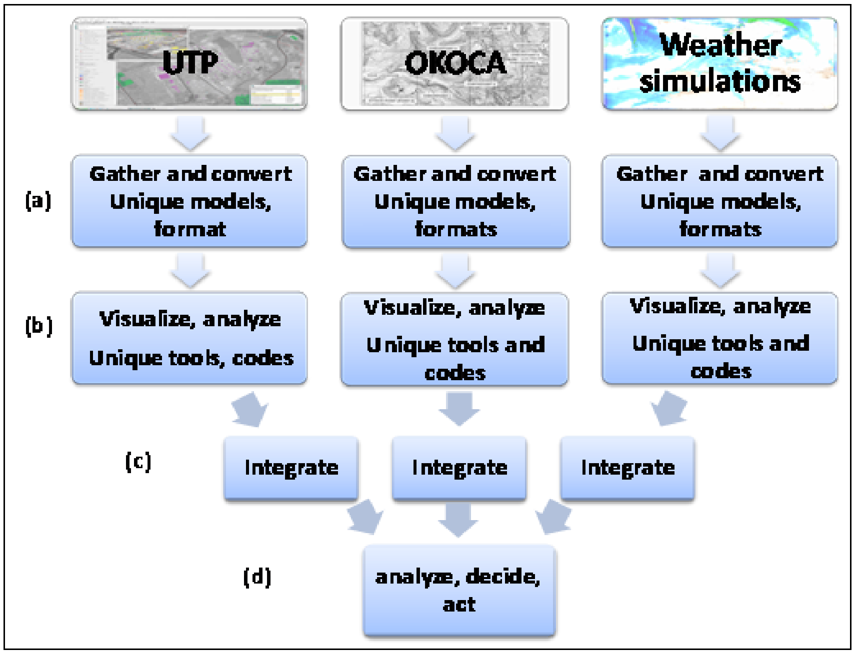


Figure 2. Separate "gather and convert" and "visualize, analyze" operations for different data sources.

Step © in Figure 2 is the "integration" step. *Data integration* allows data from multiple sources to be described in terms of a common conceptual view, which can make it easier for applications to operate on the diverse data. When applications normally act independently, as is the case in the scenario of Figure 2, each often reflects its own conceptual views, and data integration can be an ad hoc, often time-consuming operation.

Data integration facilitates *data fusion*, which is the process of combining information from heterogeneous sources into a single composite view that can then be used for decision making. Step (d) in the example can involve data fusion operations such as combining imagery and maps with terrain information and weather predictions to prepare a course of action.

Toward a unified model and format. In Figure 2, there are three completely separate pipelines and data sources. These pipelines typically would be developed independently, and have their own code base. Any operations that occur in these pipelines, such as data conversion, storage optimization, or data compression, would be developed independently, resulting in significant duplication and with no opportunity for one of the pipelines to take advantage of capabilities available in the others. Furthermore, each time that new applications are added to the process, many of those same duplications will occur over and over again.

Fortunately, there are ways to avoid this potentially costly duplication of tasks. One key to doing this is to address the problem of data heterogeneity earlier in the process, and to perform the data integration step before each of the applications actually works on the data.

Although the data for the different applications comes from many different sources and in many different forms, the real differences may be few. This can be exploited by developing a comprehensive view of the data that recognizes common meanings and structures among seemingly heterogeneous data, and then developing a conceptual model that encompasses as much of the data as possible. This model may be mapped to a unified set of data structures and a common format, and from this could be built a single system for heterogeneous data management and analysis that is adaptable to a wide range of scenarios.

Such a solution pushes the "data integration" step higher in the process, as illustrated in Figure 3. Figure 3 illustrates the benefits of having a unified set of data structures and a common format. Instead of a separate "gather and convert" process for each application, there is a common process in which a unified data model and single format are the targets for the conversions. Because there is a unified data model and single format, information does not need to be duplicated in different forms for the different tools. The output of the tools also conforms to the same data model and format, so the step that integrates the results from the three applications is simpler and faster, and results in a simpler view for the final analysis, decision, and action steps.

A higher level integration step should accommodate the full set of data types, facilitate data fusion, scale as needed, support high performance access, be platform independent, and provide a framework that supports the rapid creation of human interfaces. Such a system must work well with GIS, databases, imaging and analysis technologies, and high performance computing systems.

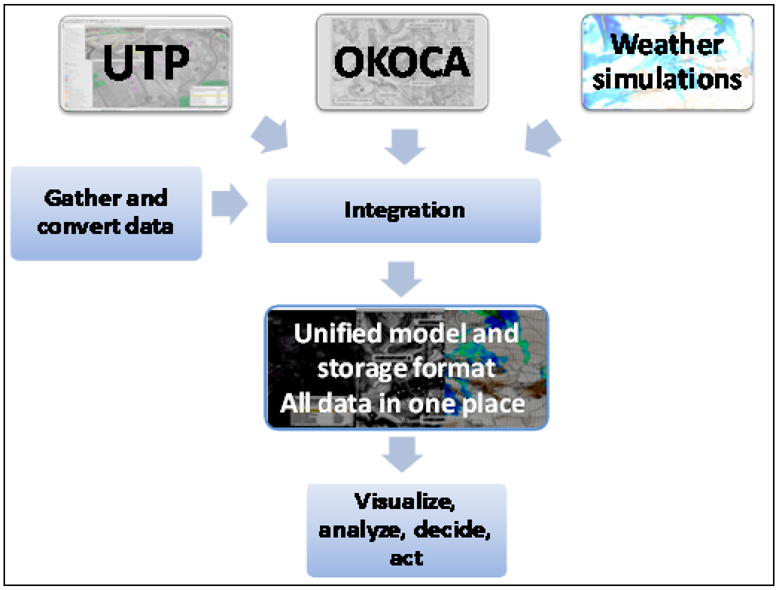


Figure 3. Unified model and format. Instead of a separate "gather and convert" process for each application, a unified set of data structures and a common format are the targets for the conversions.

In addition to simplifying the development and integration of applications, there are other important advantages to be gained by the use of common structures and a common open format. For instance, the new format can be optimized for the applications that will use the data. In a Battlefield application, the speed with which the data can be accessed and integrated can be critical, so the conversion can optimize the target format for speed.

Objective: a unified model and corresponding data structures

The proposed solution has two key components: (1) organizing the highly varied collection of data in ways that address the needs of the applications to store, access and operate on the data, and (2) finding some common ways to think about and describe this variety of data.

Component (1) is addressed by finding the right data structures and format. Component (2) is addressed by identifying common concepts that the different data embody, thus a *unified model* of the data.

Impact on the U.S. Army

Why is research necessary? The U.S. Army's 2008 ERDC Broad Area Announcement (BAA) describes the mission of the ERDC-TEC as;

> *To provide the Warfighter with a superior knowledge of the Battlefield, and to support the Nation's civil and environmental initiatives.*

In this role, the ERDC-TEC develops technologies "essential to the Army in accomplishing its global mission." Among the technology areas listed as essential are

1. Timely acquisition, fusing, analysis, display, and dissemination of remotely sensed, multisourced information depicting imagery, features, elevation, and other information essential to accurately describe the land warrior Battlespace;
2. The development of geographic information software that enables reliable, efficient, and secure information management, interoperability, and accessibility for various user communities operating globally, each with different needs;
3. The development of globally fielded applications and systems for acquiring, accessing, fusing, and delivering terrain and feature information to the soldier;
4. The development of accurate on-the-fly global positioning systems for use with inertial guidance as essential positioning engines for acquiring near-real-time, dynamic, highly accurate, remotely sensed 3D terrain and feature information;
5. The development of increasingly compact, more efficient, and more comprehensive applications and systems aimed at providing low echelon combat units with information in near-real-time, enabling rapid response to developing situations in any Battlespace;
6. The development of new and innovative techniques to understand and visualize terrain and Battlespace information in all dimensions, and to accommodate reasoning within analytical results;
7. The development of accurate and efficient survey and mapping systems for use by both military and civil communities;
8. Capabilities in acquisition, testing and fielding of topographic systems; advanced and engineering development of imagery systems; and research and development in the areas of imagery and intelligence data exploitation;
9. Operational capabilities in geospatial information and imagery requirements development; terrain, hydrologic, and environmental analysis; and information services.

Eight of these nine essential areas (all but number 7) address the need to be able to manage complex, high volume, heterogeneous data at high speed, and in a scalable manner that can adapt to growing volumes and changing types of data. These are effectively the challenge areas this proposed research will address.

The current BAA FY2010 research topic areas as proposed in the solicitation are

- Data Representation (TEC-16)
- Geospatial Information Exploitation (TEC-11)
- Data Manipulation (TEC-10)
- Spatial Data Bases (TEC-9)

The importance of research in these areas is also recognized in the National Academy of Science's study on Network Science, which lists "fusion of multiple sensors and sensor types across the network for real-time decision making" among the challenges associated with present-day military information networks at the tactical, operational, and strategic levels (NRC 2005).

Research methodology

Our research goal is to better understand the role that HDF5 can play in support of urban-based military operations through experimentation of alternative representations, so that we can work toward adapting HDF5 to generate scientific and engineering solutions for key data management problem areas.

Layers of specialization

Fundamental to adapting HDF5 is the understanding that HDF5 is a platform for storing and accessing data, and is just one of several conceptual layers that need to be considered in basing an application on HDF5. Table 1 describes these layers and shows how they are related.

HDF5 itself is represented by layer (d). It is the layer at which data is stored and accessed, and as such provides the fundamental building blocks for all of the layers above it. HDF5 does not embody entities or concepts from scientific or engineering domains, such as geographic features, physical relationships, variables, or coordinate systems. HDF5 provides data types and structures with which one can instantiate those entities and concepts, and the decision about how to do this in HDF5 depends on many factors.

Table 1. Layers of specialization of objects supported by data formats. Each layer describes a conceptual layer that is built upon the layers below.

Layers of specialization	Data types, objects, features
(a) Problem-specific information	Building footprints, roads, cultural zones, line-of-sight, and ground cover. Metadata about buildings, roads, cultural zones, etc.
(b) Domain-specific information	Elevation models, satellite imagery, projections, geo-referenced features, coordinate systems, spatial metadata
© General application data	Raster image, value at a location, date/time, time series, finite-element (FE) mesh, vector, multi-resolution grid, index
(d) Basic data	Number (integer, real), record, array, group, attribute, storage structures

To understand how to best organize and access data in HDF5, it is advisable to start at the top, layer (a): problem-specific information. What are the problems to solve, what information is needed to solve them, and what operations should one be able to perform on that information? Data and the data operations need to be described in those terms, using the vocabulary and concepts that are natural to the problem space. Whenever possible, tools should be built that reflect that same layer of thinking as well. In other words, ideally there should be no burden on an application to understand data types, objects, or features in terms of layers (b), (c) or especially (d). Applications need to be able to focus on their problems and their information in their terms.

As long as there is only one problem to be solved, it may make sense to build layer (a) out of the components of layer (c) or (d). However, it is often the case that a community has many problems that are different in specifics but are similar in terms of the types of information with which they deal, and also are similar in terms of what they do with that information. For instance, a groundwater modeling group may work with hydrological and elevation models to understand groundwater processes. Another group may work with the very same data plugged into a model that predicts flooding along specific roadways. For cases like these, layer (b) represents an opportunity to develop information structures and tools that can serve a wide range of communities.

A good example of layer (b) is HDF-EOS, which is a software package that instantiates the "earth science data types" that constitute NASA's Earth Observing System (EOS), a system of satellites with over a dozen instruments, and hundreds of different data products. EOS earth science data types include, for instance, a "grid" data type, for storing data according to any of several map projections. A large portion of EOS data products are represented as EOS grid types.

HDF-EOS also includes tools and a library for common operations on its earth science data types. For grids, for instance, there are application programming interfaces (API) and tools that convert from one projection type to another, and others that extract data within a given rectangular area on the earth. In addition, a number of general tools have been adapted to support HDF-EOS data, including MATLAB, IDL, and HDFView.

Layer © describes data objects that are used widely across many domains. Some, such as raster images, occur in almost every scientific and engineering discipline and many others as well. Layer (b) applications can benefit by using these structures to create domain-specific information objects, instead of having to reinvent them.

Research steps

The research approach in this study adheres to the layers perspective, and takes a similar path to that of the development of HDF-EOS. In developing HDF-EOS, a number of different problem spaces were described by earth scientists, and those scientists and their teams developed prototypes using HDF. Out of those experiences, it was possible to synthesize a unified model of earth science data that covered a large portion of the expected data products anticipated from the EOS project. Lessons were learned, and the process was iterated a number of times, until the first version of HDF-EOS was defined and implemented.

In the case studied in this paper, the research begins with a representative example developed at ERDC-TEC involving a sample urban Battlespace. The example includes data that typifies Battlefield data in terms of data types, granule sizes, and heterogeneity. This paper goes into some detail in describing this example, using it to show how the layers of information are identified, described, and ultimately how they suggest particular instantiations in HDF5. In summary, the steps may be described as follows:

1. Identify the problem, the data, and the information to be used in solving the problem;
2. Identify data operational requirements, such as
   a. Expected operations on the data, such as geolocation, orthorectification (translating data to a common grid), layering, zooming, querying;
   b. Data representation characteristics (e.g. raster, vector, relational tables, free text);
   c. Other characteristics of importance, such as dataset size, data types, metadata needs;
   d. Constraints, such as limits on data volume and data accretion and access speed requirements;
3. Develop a conceptual model that encompasses as much of the data and information as possible, at the same time holding the number of data types to a minimum;
4. Based on these requirements, determine how to represent the data in HDF5;
5. Build prototypes to test the results;
6. Iterate to address lessons learned and to expand capabilities, evolving both the application and underlying HDF5 technologies accordingly.

From this investigation, the paper identifies possible short, medium, and long term research and development activities directed toward achieving scalable Battlefield information management.

2 An introduction to HDF5

HDF5 is a suite of technologies built around the HDF5 data format and HDF5 access library. The HDF5 format and supporting software provide a platform upon which to build applications and tools to address some of today's most critical challenges in organizing and accessing data, especially high volume, complex and heterogeneous data. HDF5 was designed to manage, access, analyze, share, and preserve every kind of digital data, regardless of origin or size.

The HDF5 community

A brief history of HDF. In 1988, the Hierarchical Data Format (HDF) was created at the National Center for Supercomputing Applications (NCSA) to provide a software library and file format addressing the need to move scientific and complex data among disparate computing systems. In the early 1990s, the HDF group began working with the National Aeronautics and Space Administration (NASA) to employ HDF as the standard format for the Earth Observing System (EOS), the data collection system supporting research on global climate change.

In 1998, a similar collaboration with the Department of Energy's (DOE) Accelerated Strategic Computing Initiative (ASCI) produced HDF5, a simpler yet more powerful successor to the original HDF. The ASCI program was aimed at transitioning nuclear stockpile stewardship from testing to computer simulation; and in HDF5, it needed a data technology capable of handling complex, metadata-rich, terabyte-sized datasets and parallel file processing on the world's largest computer systems. It continues to be used heavily at the National Laboratories, particularly for large scale simulations, but also in other applications that challenge conventional data management technologies.

HDF5 applications and users. Today, HDF5-based applications address some of the world's most critical data challenges, including the need to capture and organize complex heterogeneous data collections, to manage very large and very complex data, and to manage data across a wide variety of computing platforms and continuously evolving computing, storage, and network environments. As a universal platform for managing data, HDF5 has found acceptance in almost every kind of scientific and engineering application, and many others as well.

More than 600 organizations, more than 200 types of applications, and millions of individuals from more than 100 countries are now using HDF5. Applications as disparate as meteorology, flight testing, film making, and bioinformatics, and the data management challenges they bring, have enabled The HDF Group to build a team with a comprehensive and deep understanding of most aspects of scientific data acquisition, storage, and access.

The EOS project alone estimates more than 1.6 million users of HDF, including the global climate research community, and dozens of other applications such as atmospheric sciences, agriculture, fire detection, and land use. EOS stores three terabytes of satellite data per day in HDF5 and its predecessor HDF4. EOS data repositories manage several petabytes of remote sensed data, representing more than six hundred different data products. These products serve the needs of millions of users.

The National Polar Orbiting Environmental Satellite System (NPOESS) will succeed EOS and will, in addition, distribute instant weather data in HDF5 to the Army, Air Force, Navy, and US weather services.

A growing number of federal agencies are adopting HDF5 for data storage, exchange, and distribution, including many for military applications. A small sampling of HDF5 applications includes

- The Aberdeen Test Center's VISION project (VISION 2008), which uses HDF5 to store, query, and access data from nearly a million test runs. HDF5 technologies provide a unique platform on which to address many of the challenges described above;
- The use of HDF5 by major aerospace companies to acquire, query, and archive flight test data used in the development of several aircraft;
- A Naval weapon systems research program, which uses HDF5 as the data storage and retrieval structure for technical data, facilitating data sharing and interoperability across multiple facilities and projects;
- The U.S. Army Research Laboratory Multimodal Signatures Database, a centralized collection of data signatures including ground and air vehicles, personnel, mortar, artillery, and many other high value targets (Bennett 2007).

How HDF is supported. Because the HDF formats and basic software are open and free, the success of HDF depends on support by organizations in both the public and private sectors that rely on HDF. This support is channeled primarily through The HDF Group, a non-profit organization dedicated to stewardship of HDF and support for its users.

The Department of Energy (DOE) ASCI project sponsored the development of HDF5, and projects in DOE labs continue to support HDF5 maintenance, as well as development. Because HDF is a mission critical technology for the Earth Observing System, NASA sponsors a range of HDF activities by The HDF Group, including software maintenance and development, and direct support for users, vendors and applications developers. These and other organizations also invest in research activities by the HDF Group and its partners that help evolve and adapt HDF to address new data challenges.

These supporting activities help to insure the viability of the HDF software, and also guarantee that HDF will meet its sponsors' specific needs now and into the future.

HDF5 model and format

An HDF5 file consists of a collection of data objects with very flexible organizing structures. The basic HDF5 object model is relatively simple, yet extremely versatile in terms of the types of data that it can store. The model contains two primary objects: groups, and datasets. Groups provide the organizing structures, and datasets are the basic storage structures. An HDF5 dataset is essentially a uniform multidimensional array of elements of a certain datatype. HDF5 supports a rich variety of datatypes, so that virtually any kind of data can be conveniently represented by an HDF5 dataset or combination of datasets. HDF5 groups and datasets may also have associated attributes, which are small data objects for storing metadata that are defined by applications.

Groups, datasets, and links. An HDF5 file can be viewed as a container, in which data objects are organized in ways that are meaningful and convenient to an application. An HDF5 dataset is similar to a file in a computer file system. An HDF5 group is similar to a directory, or folder, in a computer file system. An HDF5 group contains groups or datasets, together with supporting metadata. Figure 4 shows the structure of an HDF5 file using HDFView.

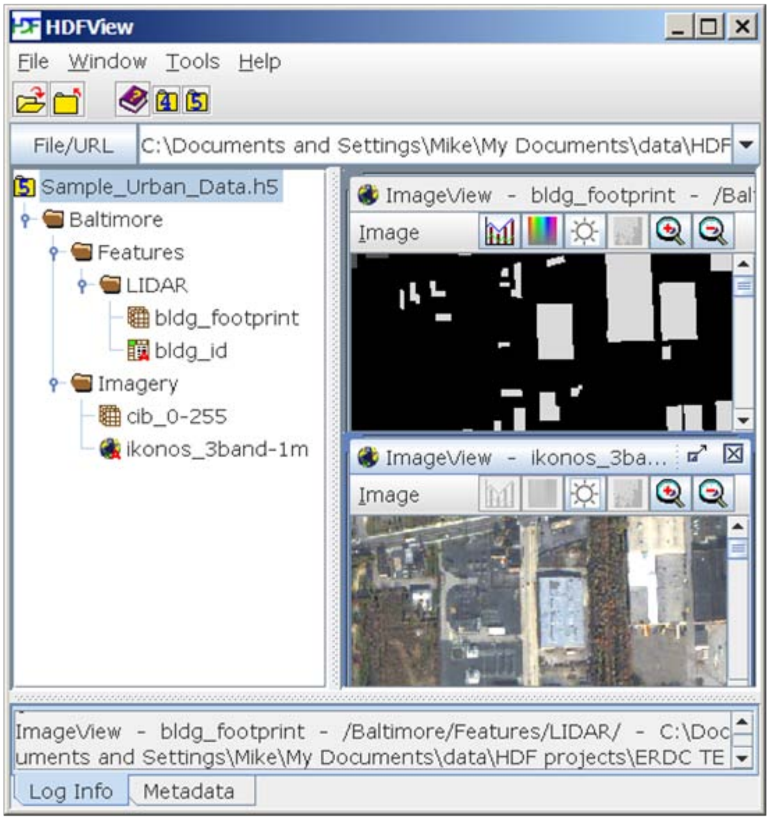


Figure 4. HDF5 file showing HDF5 grouping structure. The group *Baltimore* contains two groups: *Features* and *Imagery*. *Features* include the group *Lidar*, which contains two datasets: *bldg_footprint* and *bld_id*. On the right, images are displayed of the datasets *bldg_footprint* and *ikonos+3band-1m*. The illustration is created with HDFView, a general purpose HDF file viewer.

The contents of a group are designated using a HDF5 structure called a *link*, so that the organization of an HDF5 file can also be described as a directed graph structure in which groups and datasets are nodes, and links are edges. Links are important in this study because they provide a convenient way to show relationships among different information objects. HDF5 groups normally contain objects that are in a single file, but HDF5 links can also point to external objects. This feature is important because there will be times when an information object may need to be stored separately from an HDF5 file.

Attributes and other metadata. Any HDF5 group or dataset may have an associated attribute list. *Attributes* are small data entities used to describe the nature and/or the intended usage of a dataset or group. An attribute has two parts: (1) a *name* and (2) a *value*. The value part contains one or more data entries of the same datatype.

Metadata is also often stored in an HDF5 file using HDF5 structures. For instance, the dataset bldg_id in Figure 4 is actually a table in which each row contains information about a building in the corresponding dataset bldg_footprint.

Storage format. The HDF5 format specifies how HDF5 objects are stored. The way objects are organized can often have a profound effect on how efficiently they can be stored and accessed. For instance, if a format permits large numeric arrays to be compressed, redundancy can often be reduced, saving space. Similarly, if a format can accommodate indexes to data records in a table, the time it takes to randomly access a given record can often be much less than would be the case if the same table had to be searched sequentially for the same record. At the same time, no single storage structure is best for all types of data storage and access.

Recognizing this need, the HDF5 format and model offer a variety of ways to store objects on disk. The rich set of HDF5 datatypes makes it possible to choose an appropriate datatype for a particular dataset array. For instance, if the integers in a dataset will never exceed 255, then a one-byte integer may be chosen to store a dataset. HDF5 offers a variety of options that compress datasets, as well as options that allow applications to select storage structures that can improve the efficiency of storing data.

HDF5 also addresses the need to improve the speed of data access in a number of ways. The flexibility of the HDF5 grouping structure makes it possible to add information that can inform and speed up access. For example, metadata can be added to help find objects or portions of objects. This approach is often taken by adding indexes to the HDF5 file for rapid lookup. At the data layout level, dataset arrays can be stored in chunks or tiles, enabling fast subsetting of large datasets, including compressed datasets.

HDF software

Virtually all users of HDF5 access it through HDF5 software. Figure 5 shows the different layers of HDF5 software. The HDF5 I/O library and API (middle layer) provide access to all of HDF5's capabilities. This open source library is used to create, write, read, query, and delete objects in HDF5 files. It is also the interface for invoking other capabilities of HDF5, such as specifying the disk layout of HDF5 objects, or instructing the library to write data in parallel.

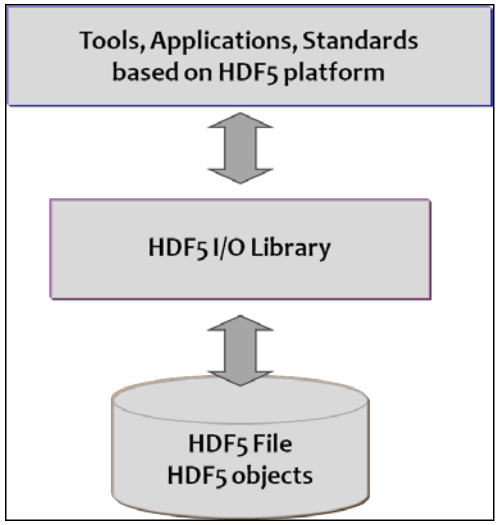


Figure 5. HDF5 software layers.

Virtually all tools and applications that use HDF5 (top layer) do so through the HDF5 I/O library. Tools and applications provide a conceptual buffer between pure HDF objects and the view of the data that users need to make sense of their data.

Many users access HDF5 files with tools. These include tools that are delivered with the HDF5 package, including command line packages such as h5dump (for dumping the contents of an HDF5 file), or HDFView, a graphical user interface (GUI) illustrated in Figure 4. Many third-party tools also provide access to HDF5 data. These include commercial tools such as MATLAB, IDL, ParaView, Vis5D, and Mathematica, as well as a large number of freely available open source tools developed by individuals and organizations that rely on HDF5.

Because HDF5 is used heavily in the earth sciences, there are many tools for working with earth science data in HDF5. MATLAB and IDL both display HDF-EOS files, for instance, and IDL has a large number of interfaces for specific EOS data products. Other geospatial tools, such as ERDAS Imagine, are able to import geospatial data from HDF.

Data integration and data sharing with HDF5

As noted above, a number of elements go into achieving data integration with HDF5 (Figure 6). First, there should be a common conceptual view of the various types of data that are to be integrated, a so-called unified data model. Second, because HDF5 offers countless different ways to organize any given collection of data, there should be an agreement and specification about how the data might be stored in HDF5. Third, it is useful to have an API for building applications to store, retrieve, and query the data, together with one or more implementations of the API in the form of software libraries. These three elements, when instantiated in HDF5, are sometimes referred to as a "profile." Finally, to facilitate access and use of the data by end users, an HDF5 profile may be supplemented with tools of various kinds.

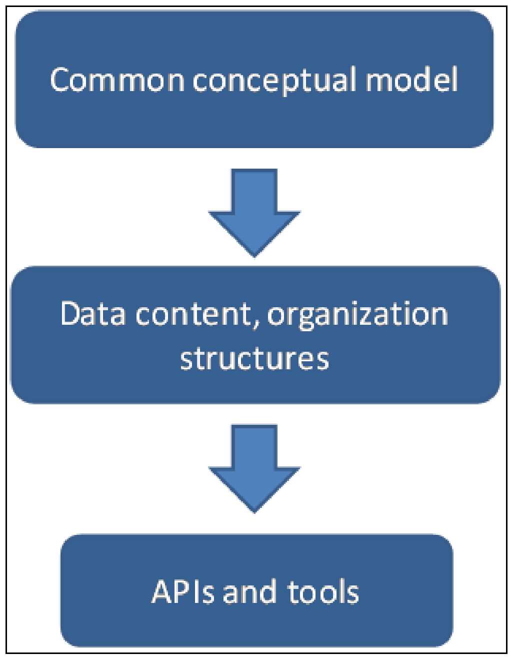


Figure 6. Achieving data integration.

A number of profiles have been developed by organizations or communities to integrate data, with HDF5 as their format platform. Examples are

- **HDF-EOS.** NASA EOS data comes from many instruments, and includes large granules of remotely sensed satellite data, in-situ data, and other geospatial data. The HDF-EOS profile conceptual model includes a small number of "earth science datatypes," such as map projections. An HDF-EOS API and library exist for developing applications to use the data, and a number of tools, including several commercial tools, are available for working with HDF-EOS data. EOS serves millions of users and countless applications from agriculture to climate science.
- **CGNS.** CGNS data, normally associated with computational fluid dynamics (CFD), can be large and varied. The CFD General Notation

System model consists of structured and unstructured grids, elements (bar, triangle, etc.), and other objects and metadata. CGNS has two formats, including an HDF5 instantiation that specifies how these objects are to be stored in HDF5, and includes an API, library, and tools. CGNS applications exist throughout government, private industry, and academia.

- **NetCDF**. NetCDF (Network Common Data Form) is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. NetCDF has an HDF5 instantiation (netCDF4). The essence of the netCDF model is a coordinate system, definition of variables within the coordinate system, and attributes for metadata. NetCDF is the format of choice for atmospheric sciences, and is commonly used in climatology and meteorology applications, as well as GIS.
- **NeXus**. NeXus is a common data format for neutron, x-ray, and muon science. The conceptual model includes the concept of experiments, with an experiment consisting of an instrument, data, samples, and other information. These are translated to certain HDF5 entities. There exists a NeXus API and NeXus utilities. NeXus serves a worldwide community of users involved in a wide variety of research and industrial applications.
- **BioHDF.** The rapid growth of genomic sciences, coupled with an explosion in genomic data, has presented life science applications from basic research to medicine with significant data challenges. Many in this community are turning to new data management methods. The NIH-funded “BioHDF” project is addressing some of these challenges by developing a standard data model, API, and tools based on the HDF5 platform. Begun in early 2009, the project has already demonstrated clear gains in access time and storage efficiency over traditional text-based methods.

3 Extended example: an urban Battlespace

Complex military operations require “wicked” problem solving methods. Commanders must devise ways to resolve a wide variety of highly complex and unique problem situations spanning the entire spectrum of military operations. Known and practiced solutions of doctrine will not suffice in this dynamic, unconventional environment. Innovative strategies and methods must be employed to meet the challenges of providing solutions to these ‘wicked problems’ (Schmitt 2006).

Solving wicked problems requires a systems approach. This means treating the problem domain as an integrated whole. One component of a system is information about the system: facts, conditions, and relationships. Information fills a critical need in solving wicked problems, and a wide variety of information must be integrated and readily accessible. Because wicked problems are essentially unique, solving them cannot rely on conventional information sources or tools. Custom-made information management approaches are needed to address unique operational situations.

This chapter describes an example in which conventional information sources and technologies are inadequate for military operations occurring in a modern urban Battlespace. Managing the large variety and volume of geospatial information to adequately represent the tactical terrain component of these operations in an integrated, readily available way is a considerable task, and one for which established geospatial technologies fall short. Furthermore, traditional categories of geospatial data are inadequate, as the need to incorporate socio-cultural information is now well known.

The example illustrates an alternative approach to addressing some of these issues within the context of the HDF5 data management paradigm, and thus demonstrates a promising approach to transitioning current GIS methods to meet the ‘wicked problem’ challenge.

The key steps to developing this complex model and computational data structure are:

1. Identify the problem and available data/information sources used in the analysis;
2. Identify the required data management operations;
3. Develop a conceptual model that encompasses the fundamental scope of data/information workflows and functional tasks, while optimizing the data types;
4. Based on these actions, determine how to best represent and organize the data in the HDF5 solution space;
5. Build a prototype to test, evaluate, and verify the technical approach and solutions.

Identifying the problem and data sources used in the analysis

Modern urban military operations require a rich mixture of geospatial information that enable terrain data reasoning, including geographic information (elevation, buildings, roads), terrain information (OCOKA), area usage information (human institutions), imagery, and weather. Irregular urban warfare makes this a “wicked” problem, adding a sociocultural dimension, namely the need to integrate temporal, social, and cultural concepts. All of these information sources must be unified, providing a clear and simple conceptual view, and making it possible to query, access, and combine the data objects.

The following example demonstrates some ideas on a technical approach to modeling, organizing, storing, and viewing some of this data. The example is restricted to terrain data reasoning, but the approach it exemplifies is one with the potential to evolve to accommodate the sociocultural domain as well. This latter domain area is explored in a later chapter.

For this example, the urban demonstration test area is a 4000 x 6000 dataspace terrain reasoning array at 1 m horizontal sampling resolution representing 24 million atomic spatial terrain objects (Figure 7). The experimental model includes geometry (x, y, z dimensions), imagery, city level feature data, and value-added topographic stacks.

Figure 7. Geographic area covered by test data.

The challenge problem in this example is to provide multiresolution geo-coincident information sources and operations based constructs for terrain reasoning over a sample urban scale Area of Interest (AOI). The sample combines a hybrid of data sources and types, including traditional maps, socio-cultural data, physical terrain data/features, spatial geometry parameters, and high fidelity aerial and satellite imagery. This would include feature data from sources such as the Urban Tactical Planner (UTP), LIDAR, NAVTEQ, OCOKA derived data, and simulated weather data tables.

Six distinct types of information are included:

1. **Features:**
    a. Urban objects:
        i. Buildings: footprint and ID for every building
        ii. Roads: map showing the locations of all roads
    b. Region usage information ("BTZone" group): footprints of areas involving human institutions and activities, together with identifying information, including separate datasets showing commercial, cultural, industrial, institutional, and residential areas.
2. **Geometry:**
    a. The UTM Northing, Easting, and Elevation (x, y, z) for each data point.
    b. Units in meters using WGS84 datum.
3. **Imagery**
    a. An 8-bit Controlled Image Base (CIB) image of the AOI: an orthophoto made from rectified grayscale aerial images
    b. An IKONOS satellite image made up of 3 spectral bands at 1 meter resolution.

4. **Terrain (OCOKA) information:**
   a. Omni-Directional Line of sight (LOS)
   b. Ground Cover/Concealment
   c. Obstacles
5. **Metadata** for all of the above, which can be quite varied, but consists typically of per-object (e.g. per-building) and per-collection (e.g. per-building footprint collection) attribute records.
6. **Weather**
   a. Templates of potential atmospheric conditions as captured from archived local weather stations
   b. Predictions of suitability for several UAV reconnaissance platforms.

Table 2 lists some of the source data used in the example, together with characteristics of the data.

Table 2. Examples of data objects used in prototype and their characteristics.

Source objects	Source file type	Informa-tion type	Data structure representing object	Size
Building footprint	LIDAR	Feature	Polygon	4468 polygons
Building footprint metadata	LIDAR	Metadata	Tuple w/ 12 attributes	4468 tuples
Building collection metadata	XML	Metadata	XML structures	Small
Cultural zone	UTP	Feature	Polygon	52 polygons
Cultural zone ID	UTP	Metadata	Tuple w/ 18 attributes	52 tuples
Dimensions (easting, northing, elev)	Coords	Geometry	32-bit 2D array	4Kx6K 32-bit values each
Ikonos 3 band image	Imagery	Image	8-bit 2D array	4Kx6K bytes
Weather scenarios		Weather	Tuple w/ 32 attributes	9 tables/23-62 rows

Enabling geospatial data operations in HDF5

Fundamental areas of geospatial data management operations:

1. Input – translation of data sources into the specified format;
2. Analysis – processing of the data to generate solutions;
3. Output – export of data to external applications or services.

Input – translation of data sources into the specified format

Input is the process of importing data into a physical storage space. In this instance, our interest focuses on conversion of data into the HDF5 format from a wide variety of relevant sources and formats with minimal loss of information. In our example, data sources are primarily geospatial, for example geometry, imagery, and digital feature data. Input can come from such varied sources as standard ERDAS and ESRI formatted files, imagery, relational databases, and on-the-ground real time observations. Because some of this data may arrive in real time, it may be important to be able to import data at a high rate of speed.

Each data source intended to be utilized in the system needs to be converted to HDF5, which means (a) mapping each type of data into appropriate HDF5 data types and structures, (b) coincident georeferencing in terms of coordinate geometry, and (c) creating tools to convert data from the format into HDF5.

HDF5 choices for the mappings should also be made in view of the input, output, and analysis requirements. Examples of data types that are defined in the urban Battlespace study are imagery (e.g. IKONOS 3 band image), which can be represented as two-dimensional (2D) HDF5 datasets, and relational tables, which can be converted to one-dimensional (1D) HDF5 datasets with compound datatypes. This process is illustrated in detail in section 4.

Definition of the input data geometry consists of linking appropriate geographic coordinate systems (e.g. WGS-84, UTM) to the internal data array structures to spatially link the internal data representations to a real-world frame of reference.

Conversion tools for importing data are needed to enable fast, accurate, and consistent conversion. Ideally, there would also be APIs and software libraries corresponding to these tools, so that HDFView and other applications could easily be extended to support the same import operations. Many of the tools should be scriptable, to enable implementation of complex workflows for importing combinations of data. Similarly, the APIs should be designed to enable complex workflows to be constructed by high level scripting languages.

Analysis– processing of the data to generate solutions

"Analysis" refers to processing data in memory. The number of possible analytical operations on the data is large and varied. Examples include performing queries about the nature of the phenomena represented in the data, such as the height of a building or the elevation of a particular position, determining a solution to an operational requirement, such as a position of advantage, line of sight, combining data in ways that increase understanding via creation of composite maps, and executing algorithms to determine alternative outcomes such as Military Course of Actions (COAs) and Intelligence Preparation of the Battlefield (IPB).

Figure 8 Illustrates how HDF5 formatted terrain data can be correlated to increase understanding of a particular scenario.

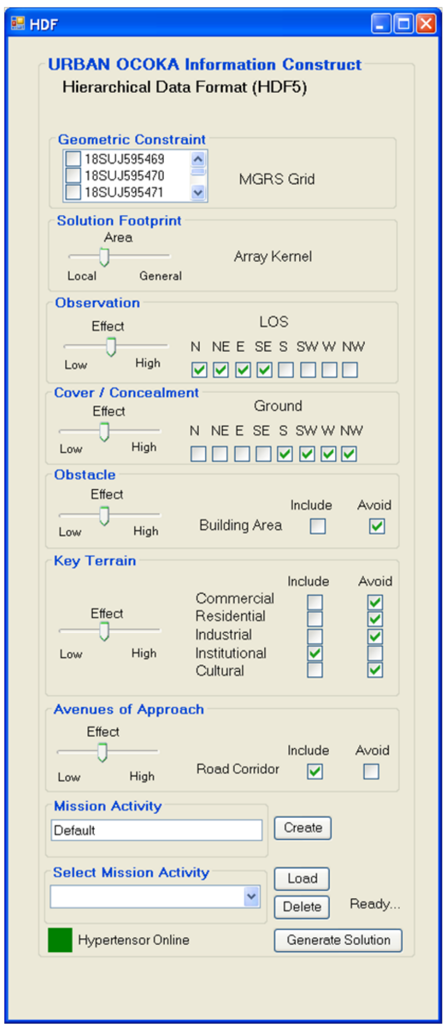


Figure 8. Generation of an Urban OCOKA information construct.

In this prototype HDF5 Graphical User Interface (GUI) developed as an urban situational analysis tool, the user can select key parameters to generate potential products from the linked HDF5 based terrain service. The results retrieved from the server based HDF5 terrain reasoning application are then converted to a Shapefile format using an Open GIS Geographic data abstraction library and then transferred back to the client for display. For this particular effort, the ESRI ArcGIS ArcMAP product was used as the client interface.

Geospatial data management operations are central to this systems approach. The principle operation upon which most others depend is geolocation: it must be possible to determine and traverse, either explicitly or implicitly, the ground location of data values. In many cases, queries can be answered directly from descriptive metadata. For example the building footprint metadata table contains such information as the area, minimum height, and maximum height of each building.

In the example, there is no special processing of imagery other than to display the three image bands. In a general case however, it should also be possible to zoom in or out quickly, and pan over the data. It should also be possible to merge and stack multiple layers of all geospatially referenced data, such as buildings, roads, and zones to build more complex terrain data objects. Other higher order operations include determining Line-Of-Sight (LOS) and ground cover within the boundaries of the solution space.

The metadata, such as the “ID” tag associated with each zone, should permit easy browsing, so that an application or user can gain a quick understanding of the data sources without examining the data itself. Metadata should be represented in ways that make searching efficient. It should also be possible to browse metadata for a given type of feature (e.g. buildings) or an instance of a feature.

The capability should exist to support analysis with a variety of toolsets, such as MATLAB, IDL, and GIS tools (both proprietary and OGIS compliant), which themselves would access the HDF5 data structures through the library, but which would hide the low-level HDF5 interfaces and format from users.

Output– export of data to external applications or services

Output is the process of transferring data and derived solutions from storage to another medium, possibly in a different form. In this instance, our interest focuses on exporting data from HDF5 to applications that provide complimentary visualization and analysis capabilities. Output data operations including exporting the data to other user environments, such as serving the data to a geostatistical package or geospatial tools. For example, existing terrain analysis tools may exploit a particular HDF5 data structure (e.g. complex group) output to determine an appropriate area and/or position to conduct a specified mission or task.

Geospatial formats of particular interest include Shapefiles, ESRI grids, and GML. This export capability is key to effective integration of important geospatial applications, such as the Geospatial Data Abstraction Library (GDAL), ESRI tools, and similar applications. It is also important in enabling services on the web.

A common view to support workflows while optimizing the data space

In the example, a common conceptual view is achieved by mapping all relevant physical content such as features, imagery, and terrain information to the coverage area within the context of the intended functional areas of operations. Consistent portrayal of the data must resolve issues of native resolutions, scale and metadata standards. The specified layers should also coincide with any existing functional requirements captured in prior mission scenarios and/or use case studies. The resulting 'Concept Map' serves as the notional framework to portray operational (*physical + functional*) workflows which can then be used to develop an optimized data space.

A brief look at the datasets involved illustrates the variety of data types that need to be accommodated, and how they may be formatted for best usage. In our sample implementation, the common view has several particularly important characteristics:

- The Battlefield in question can be described as a geographical area of interest (AOI), and hence all data needs to fall within that AOI;
- To permit fast data access, merging, and stacking, most of the data are best represented by regular arrays having the same resolution;

- There must be sufficient attribution to allow fast merges and other operations across the datasets;
- The metadata should permit simple querying and browsing, showing what is in the collection;
- The objects within the view must be organized by logical groups, naming conventions, and hierarchical levels to maximize workflows.

To satisfy criteria #1 and #2, there needs to be a way to represent both vector and raster formatted data in a way that permits fast access and data fusion operations. One way to accomplish this is to represent every feature as a set of points on a common rectangular grid that spans the AOI. Thus all of the data from Table 2 with the information type "Feature" is mapped to a common grid, here referred to as the "AOI grid." This includes imagery, terrain data, geometry, urban objects, and area usage information.

The decision to use a common grid also means that data that does not conform to the uniform grid format will have to be converted. For example, features represented by polygons will need to be translated to regions within the array. Source input data with different resolutions or coverage extents will have to be transformed through appropriate subsetting, interpolation, aggregation, and/or geospatial processing methods.

Although these data types map to a common AOI grid, the meanings of the grid points differ greatly among the different information types. Whereas the grid points for CIB images consist of 8-bit picture elements, the grid points in the elevation array consist of elevation values relative to sea level, the road corridor grid is a bit map (1=road, 0=no road), and so forth.

In addition to the grids, sufficient metadata is needed to interpret that data. This comes in several forms, including relational tables, XML files, and simple attributes. It is important to note that these forms are also expressed in terms of GIS community standards, as much as possible. Thus, for example, ESRI profile FGDC metadata is included and adheres to the ESRI profile XML document type definition for digital data.

This process of developing a common conceptual view results in building a HDF5 dataspace that represents a 'best fit' to the particular military problem solving, decision-making domain. Traditional methods of constructing these baseline data solution space structures (e.g. geodata-base) for geospatial analysis sometimes lack the necessary levels of organization and optimization to adequately support the required operations on these increasingly complex information workflows. The usage of these high level object-oriented modeling methods such as UML and CmapTools[1] to visualize, construct and 'fine-tune' the conceptual HDF5 dataspace components has proven to be a critical step in designing the HDF5 Urban Battlespace example.

How to best represent and organize the HDF5 solution space

Having identified key enabling data management operations and a common conceptual view of the workflows, how should the solution space be represented in HDF5? What HDF5 data types should be used, what organizational structures, what disk layout options, and so forth are appropriate for implementation?

Data structures

The urban Battlefield example contains essentially six different types of data: features, geometry, imagery, value-added information layers (OCOKA), weather scenarios, and descriptive metadata. What HDF5 data structures should be used for these to meet the required operations that have been identified?

These data represent a range of size requirements, from relatively small tables (less than 100 rows and 25 columns) to potentially large, massive high resolution datasets. The actual ground sample resolution in the urban example is fixed at 1m (meter). So, every data point in the HDF5 array is representative of a 1m lattice center-point stepping distance in both a northing and easting direction. This results in a solution space with 24M 'atomic terrain array objects' available for processing per identified data stack. Advanced sensors and collection technologies will certainly impose much higher resolution requirements on future data management operations, and our research efforts need to anticipate that.

Since a great deal of the data is represented by the same 2D rectangular AOI, it makes practical sense to store all of the grid objects as HDF5 2D datasets. The element type for each of these datasets can be chosen based on the constraints imposed by the model. In the example, an AOI grid size

[1] Institute for Human and Machine Cognition (IHMC) Concept Map tools. http://cmap.ihmc.us/conceptmap.html.

of 4Kx6K (Note: 4 kilometer x 6 kilometer) elements has been selected, as it provides sufficient areal extent for the demonstration. HDF5 datasets can, and must, easily support datasets of this magnitude, as well as much higher resolution datasets.

Most of the grid objects are mapped to this 4Kx6K resolution, and hence a 4Kx6K HDF5 dataset is used for each grid object. For simplicity, the element types for these datasets is a 32-bit integer. An exception is the IKONOS 3-band image. The image is mapped to the same 4Kx6K space to make it permit easy and fast data integration, but because there are three spectral bands, each band is stored as a separate plane. Hence a 4Kx6Kx3 HDF5 dataset is used. The element type for this dataset is 8-bit integer, corresponding to the element type of the original IKONOS image pixels.

Table 3 describes some of the HDF5 structures used to represent these and other source objects.

Table 3. HDF5 dataset properties for example data.

Source	General type	element type	HDF5 dataset properties rank, dimensions, data type
Building footprint	Grid	Building ID	2D, 4K x 6K, 32-bit integer
Road corridors	Grid	bit map	2D, 4K x 6K, 32-bit integer
Commercial zone, cultural zone, etc.	Grid	Zone ID	2D, 4K x 6K, 32-bit integer
Building ID, commercial zone ID, etc.	Table	Metadata per ID	1D, 4468, 12 field compound type
Geometry-easting	Grid	Longitude	2D, 4K x 6K, 32-bit integer

The operations that the prototype needs to support can be executed efficiently with this HDF5 representation. Because these datasets can be fairly large, and subsetting operations are likely to be performed on them, HDF5's chunking and compression capabilities should be used for storage. These options will not only save space, but they can in many cases increase I/O speed, which in some cases will be an important requirement.

It must be noted that the original data mapped to the 4Kx6K grids did not all originate in the same resolution. Much of it needed to be conflated. In this experiment, the data pre-processing was done separately with standard GIS tools, but in a working system, the HDF5 import tools described earlier should be enhanced to handle these data operations.

It is also notable that some of the data described here did originate in a gridded, array based format. Some of the features, such as roads and building footprints, originated as vector data and were converted to array form. This approach facilitated the research effort to prototype our example, but may not always be the appropriate practice. In future work, it may be necessary to represent data in ways that enable I/O and analysis operations to be performed rapidly. This means that it should not be always necessary to perform time-consuming conversions of data, such as converting vector data to grids, or vice versa. HDF5 structures may need to be developed for a more native representation of vector data sources.

The metadata objects in Table 2 occur in two different forms: lookup tables and XML formatted documents. In HDF5, a table structure is usually organized as a 1D dataset, with each element of the dataset defined as a compound type, where each field of the compound type corresponds to a column in the original table structure. This is the approach taken in this case. Since the tables provide lookup by ID, they all have an ID field, and are sorted by that field, allowing fast searches to be performed.

The XML formatted documents are each a few hundred lines of variable length. Since they will always be accessed in their entirety, they can be represented either as HDF5 attributes or as HDF5 datasets for this instance.

Follow-on research efforts will investigate alternative strategies for representing XML based objects in the HDF5 data structure. This includes development of descriptive objects and the linking mechanisms to other groups and datasets within the internal solution space and external interfaces/applications. This is a very important step to integrating the HDF5 data structures with emerging geospatial reasoning enterprise network services and modeling languages (BML[1], J3CIEDM[2]) via ‘semantic tags’.

Organizational structure

The HDF5 grouping and linking structures make it possible to express logical relationships among data entities in a collection, enabling meaningful browsing and simple access. Since all of the information in this example has to do with a certain AOI, it is natural to create a group at the

[1] Battle Management Language.

[2] Joint Consultation, Command and Control Information Exchange Data Model.

top level in the file that identifies the particular AOI. In this case, the group will be called "Baltimore," as shown in Figure 9.

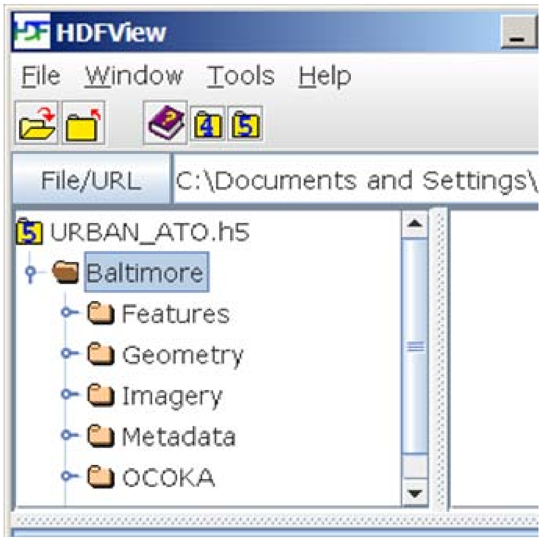


Figure 9. HDFView screen shot showing top level organization of HDF5 file reflecting information structures in example.

All associated information will be placed under the "Baltimore" group, where there are logical groupings, such as "features," and sub-groupings such as "UTP-specific features." There is also extensive metadata associated with most of the data granules.

In this instance, the various information objects fall into six previously outlined categories, and grouping them according to those categories can add meaning to the collection, facilitate browsing, and simplify the job of applications that need to find specific kinds of information. Those categories are illustrated in Figure 9 and include Features, Geometry, Imagery, Metadata, OCOKA (value-added terrain), and Weather scenarios.

Some of these major categories contain information that can also be grouped meaningfully into sub-categories, and some of those can be subdivided even further. The subdivision that was chosen is illustrated in Figure 10.

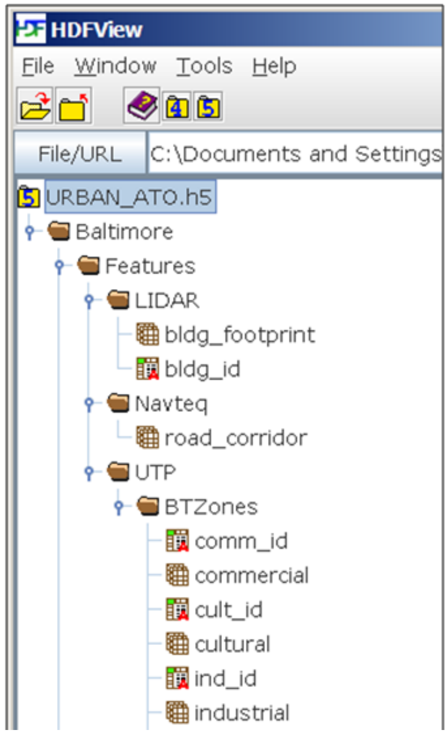


Figure 10, HDFView screen shot showing subgrouping of some sample data.

Build a prototype to demonstrate the technical approach

Military terrain reasoning and decision-making tasks may involve rather complex data management operations and may be expected to negotiate considerable heterogeneity in the types of information. This section explains a simplified use case of how the information can be presented in HDF5. The use case is concerned with capturing the structure and resources attached to a notional concept map. The concept map view is an abstraction and representation of the pattern represented by the structure and content of a set of nodes and of the resources associated with each node. There are three types of data files associated with the concept map: the raw data (example in Figure 10), the concept map layout file, and other heterogonous objects (in external files or links).

The concept map (in an HDF5 file) captures the structure of information (nodes) and resources attached to each node. Figure 11 is a snapshot of sample concept map shown in the default HDFView.

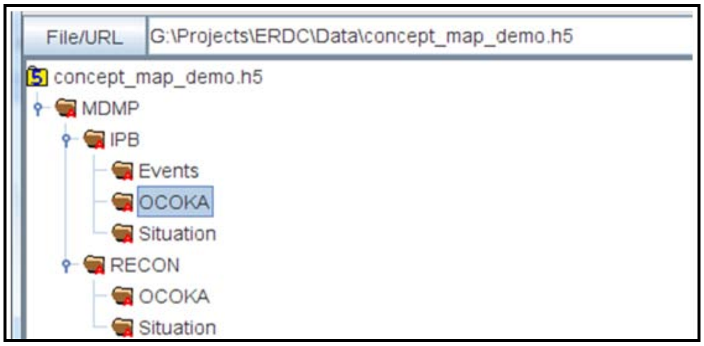


Figure 11. Concept map shown in default HDFView.

HDFView has a plug-in capability that enables specialized GUIs to be created to display a file's contents in a way that corresponds to a particular application domain. Thus, an HDFView plug-in GUI can be implemented to display HDF5 data in ways that show how data integration is achieved, and can also be adapted to perform simple fusion operations. A demonstration of these capabilities will help to illustrate the results of this study and will stimulate ideas for the next phase of work. For the purpose of this study, a simple ERDC plug-in was implemented to prove the feasibility of the concept.

When opening the concept map, the ERDC plug-in shows the concept map in a directed graph tree that represents a potential military decision making process. Instead of showing a simple group structure as in Figure 11, the ERDC plug-in shows the logical flow of information and relationships (links) among the data objects, shown in Figure 12. The labeled shapes represent concepts and the arrows represent relationships among the concepts. This sample concept map depicts the top-level of a scenario composition, where the shapes are icons representing associated resources that can take many forms: images, documents, websites, videos, executable software, etc.

In Figure 12, the rectangle, diamond, and circle shapes represent the groups (or collections) of information. The leaf nodes at the bottom (oval shaped) link to the datasets in the raw data file. For example, the bldg_footprint node links to the dataset, "/Baltimore/Features/LIDAR/ bldg_footprint," as shown in Figure 10. Other leaf nodes connected by dashed arrows are the links to other external files/objects that may be valuable to the MDMP process.

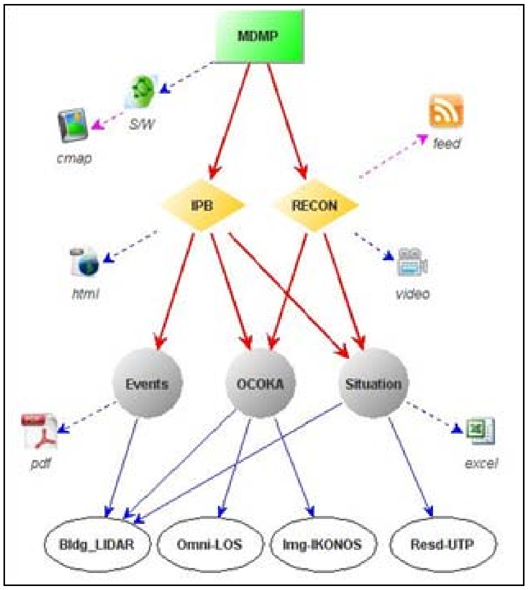


Figure 12. Concept map in ERDC plug-in.

Links to external files/objects are represented in attributes in the concept map file. The links work in <MIME, URI> pairs. For example, the <"MIME = application/vnd.ms-excel ", "URI = Situation_Weather-XLS.xls"> pair indicates that the URI is a link to an external Excel file. Table 4 shows all the links of the OCOKA concept group in the example file.

Table 4. Objects linked to OCOKA concept group.

MIME	URI
MIME = application/ vnd.ms-excel MIME 2 = application/ x-hdf MIME 3 = application/ x-hdf	URI = Situation_Weather-XLS.xls URI 2 = ATO.h5#///Baltimore/Features/UTP/BTZones/ residential URI 3 = URBAN_ATO.h5#///Baltimore/Features/LIDAR/ bldg_footprint

4 From example to prototype: next steps

The notional example outlined in Chapter 3 illustrates the foundation capability of a Battlefield geospatial data management system, but is still lacking in some key component areas to provide viable terrain reasoning solutions. Several related research areas require further investigation to adequately instantiate the HDF5 data structure for practical application.

These fall into three general areas of specificity, as illustrated in Figure 13: Battlefield conceptual model and problem space; content needs, object types, and semantics; APIs and tools.

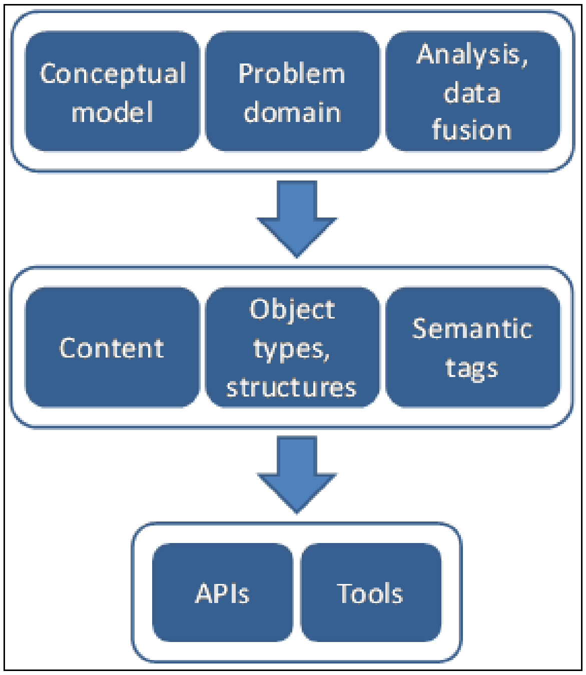


Figure 13. Battlefield conceptual model and problem space imply content needs, object types, and semantics, which are implementedby APIs and tools.

The conceptual/problem space includes three areas:

- **Conceptual Model.** A more complete model must utilize existing geospatial community standards in the development of digital data models and metadata;

- **Problem Domain.** In the example, the problem domain is limited in scope, and may need to consider the representation of many other types of complex data sources;
- **New analytic and data fusion approaches.** Geospatial Applications of the data structure were considered in the example, but more challenging, new analytic and data fusion approaches need to be applied.

The potential to further enhance and customize HDF5 technologies to support more advanced methods of geoprocessing, data fusion and services-oriented information management operations will be accomplished by additional Research & Development prototyping efforts directed towards the 'Extensibility' of the example content, data structures and semantics. Some current thoughts about these are as follows:

- **Content requirements.** Increased content and the need to handle massive data loads will occur as more sensors enter the Battlespace and resolution demands grow presents a unique challenge to data management operations;
- **Object Types and structures.** New object types must be customizable to handle large scale spatial indexing and multifaceted topologies (adjacency, containment, connectivity, networks);
- **Semantic Tags.** Semantic information embedded with objects (fields/procedures) can facilitate data and information exchange within and external to HDF5 solution space.

These capabilities and content are implemented through software in the form of APIs and tools, namely:

- **APIs (Application Programming Interfaces).** APIs can seamlessly connect with native geospatial formats (i.e. import, export) and be modifiable based upon behavior during runtime;
- **Tools.** Command line and GUI tools can automate common data management operations (e.g. query) and leverage capabilities from other existing applications such as MATLAB.

Conceptual model

The gridded data in the example has its own special profile, and as such is not interoperable with other applications. A good deal of work has gone into the development of models for geospatial data, and ultimately should not be ignored. Alternative models such as the HDF-EOS Grid profile used for EOS could be used for raster data, as well as models supported by other standards, such as Unidata's Common Data Access Model, and the ISO 19123 conceptual schema for the spatial characteristics of coverages (Nativi 2008). The conceptual model must utilize existing geospatial community standards in the development of digital data models and metadata, and should encompass the major object types in the geospatial domain, including raster and vector geospatial data, as well as sensor streams.

Problem domain

The problem domain in the example was limited in scope to a simplified situational assessment of OCOKA terrain parameters within an urban environment. Expanding the generation of terrain reasoning solutions to other more challenging, dynamically based environments will need to be considered. The representation of many other types of data sources such as utilities, hydrology, vegetation, administrative boundaries, and real time high resolution satellite imagery (e.g. micro terrain) also needs to be considered. Beyond these more traditional representations, the solution space domain may also need to manipulate data and information stacks from other more intrinsic problem areas such as socio-cultural relationships, political delineations, and economic analysis.

New analytic and data fusion approaches

A few geospatial applications of the data were considered in the example, but more rigorous, new analytic and data fusion approaches will need to be applied. We want the new data management system to interoperate with the powerful array of existing geospatial technologies, including GIS systems, such as ESRI and ERDAS, with general purpose tools such as MATLAB and IDL, and with specialized tools such as the OCOKA urban situational analysis tool described above. These new approaches will need to leverage emergent technologies in the area of High Performance Computing (HPC) to capitalize on the evolving capabilities in data processing such as parallel computing, clusters and virtual network operations.

Content requirements

Increased content and the need to handle massive data loads will occur as more sensors enter the Battlespace and resolution demands grow, presenting unique challenges to data management operations. Sensor data may originate from soldiers in the field, from low-flying aircraft, from bugeye satellite images, or any of a number of other sources. It should be anticipated that some of these datasets will become much larger, with greater spatial, temporal, and radiometric resolutions. It should be possible to stream sensor data in real time. For the time being, it is assumed that grids can accommodate most 2D data, including multi-layered and multi-resolution images. Over the long term, it may be important to support geo-referenced sensor data that does not map well to a 2D grid or projection, but this does not seem necessary at this time. Examples of such data are the so-called “swath” and “point” HDF-EOS data types.

Being able to import, integrate, and store such a variety of high volume data is a challenge, and the unique capabilities of HDF5 to manage large scale data at high accretion rates should be studied and applied.

Object types and structures

New native HDF5 object types are needed to support the conceptual model, problem domain, and analytic and fusion approaches. The next stage of the project should investigate object types and structures that are customizable, to handle large scale spatial indexing and multifaceted topologies (adjacency, containment, connectivity, networks).

Compatibility with common geospatial formats. As many of these capabilities are available in existing geospatial tools, interoperation with such tools is important. One aspect of achieving this is to assure that the HDF5 instantiation of geospatial images, maps, and other data and metadata mimics those of the other tools, and in particular, their formats. Some key examples of formats HDF5 should interoperate with include:

- NITF (National Imagery Transmission Format), a Department of Defense (DOD) suite of standards for the exchange, storage, and transmission of digital-imagery products and image-related products (NITF 2006);
- GeoPDF, an extension to the Adobe PDF format (GeoPDF 2007) used to store GIS and mapping data in a standard PDF container, including metadata to allow transformation of PDF coordinates to a projected Cartesian coordinate system;
- ESRI Shapefile, a vector format for storing geospatial features with points, polylines, and polygons;

- GeoTIFF, a metadata standard that describes georeferencing information to be embedded within a Tagged Image Format File (TIFF) file.

The Geospatial Data Abstraction Library (GDAL) and supporting OpenGIS Simple Features Reference Implementation (OGR) Simple Features Library provides abstract open source data models for encompassing many of these formats, and may be used as a basis for developing APIs and tools (described below) for managing these kinds of files.

Organizing for scalability. To allow scalability, it will be important to exploit certain storage capabilities, such as chunking and compression. In addition, storage and access for certain structures, such as images, may benefit by creating composite structures that include more than the usual raster linearization. For example, adaptive grid refinement techniques make it possible to represent in a single image certain regions at high resolution, along with other regions of lower resolution. Another beneficial technique for large image storage is to store multiresolution images, where different versions of an image are stored at different resolutions, enabling fast panning and zooming, yet preserving the information content. HDF5 can easily accommodate all of these techniques.

Another key scalability factor is control of the level at which the User is accessing the data structures and information layers. Multiple view points (i.e. scales) at which the User (and/or application) is working may be required. The proper organization of the solution space container by usage of various types of links between these levels (hierarchy) and appropriate data indexing schemes will enhance these types of scaling operations.

Semantic tags

To facilitate data and information exchange within and external to the HDF5 solution space, additional semantic information needs to be embedded within the objects and structures. These enhanced descriptive fields will provide the internal mechanism to add context to individual objects and groups to facilitate a more ‘meaningful’ structure for traversing the internal HDF5 layers and links and external associations (i.e. Symbolic Markup Language). Another relevant example pertaining to the exploitation of enhanced Semantic Tags may include the adoption of Semantic Metadata Mapping Procedures (SMMP) into the geospatial data 'mapping' process, essentially outlining the appropriate actions for improving the semantic interoperability of data (ISO/IEC 2008).

Specifying the meanings of numeric types. For example, the gridded data in the example can be categorized by these very different types of values: measurements, such as elevation and imagery; identifiers, such as building footprints; special bit-field encodings, such as line of sight. In all three cases, there should be descriptive information that enables an application to know which types of elements are represented. In the case of special bit-field encodings, there needs to be a way for applications to interpret the fields in these data types. Metadata should be included with important information about the fields, such as name, units, and field location and size.

Concept descriptions and structural information. The notional example file has a specific structure and content, but clearly both the structure and content can vary greatly, not just among problem domains, but even within a given problem domain from one day to another. Some means to describe individual file structures and content would be very useful. A concept map could be stored in an HDF5 file to show its conceptual interpretation. In addition, there should be a way for applications to link the concept map to the corresponding HDF5 groups, datasets, and links. This information would enable tools to decipher the file, from concept map to data. Applications should have information that allows them to locate the corresponding metadata or indexes to provide a user with that information on demand.

Encapsulation of digital media formats. As the example outlined in chapter 3, HDF5 can encapsulate digital media along with the other data. The method used in the example was ad hoc, but a more robust, comprehensive method might be implemented. This might include protocols for links to support encapsulation and association of media objects in a standardized way. In addition, one could store metadata for these objects using the internet standard Multipurpose Internet Mail Extensions (MIME), a protocol for identifying the type of data in attachments to email that includes the vast majority of file formats in common use. MIME types include image, audio, video, text, multipurpose, and many other types. Examples include mp3 (audio), mp4 (video), html (text), and pdf (multipurpose).

APIs

A software implementation in the form of an API and library that can store, retrieve, and query the data will help to ensure consistency in how the data is organized, and to greatly reduce the effort to build and extend applications and tools. Development of additional API capabilities within HDF5 may include routines that can support the capabilities described above, for example:

- Store and interpret (i.e. translate) concept map information, including managing links and other relationships.
- Be able to acquire, store and access high volume sensor data, possibly with temporal referencing.
- Support internal structures for scalability, such as adaptive grid refinement and multiresolution grids.
- Seamlessly import and export native geospatial formats such as NITF, GeoPDF, Shapefiles, and GeoTIFF, which will assist with compatibility within the GIS community.
- Employ common geospatial access (e.g. query) methods, such as storage and retrieval of feature data.
- Store, retrieve and interpret metadata associated with special numeric types
- Encapsulate, store, and retrieve digital media, such as MIME types.
- Potentially modifiable APIs based upon behavior during runtime for profiling of externally linked data sources.

Tools

Where API's and libraries enable the building of applications that access HDF5 data, tools are existing applications that provide direct services. Software tools may be divided into two types:

- Command line tools that make it possible in a development environment to run applications, automate common data management operations, examine the contents of a file, run performance analyses, check correctness, and similar activities. Examples of such tools are
  - h5dump: dump contents from an HDF5 file.
  - h5repack: re-organize the storage of an HDF5 file for efficiency.
  - h5perf: run performance analyses on an HDF5 file.
- Interactive tools, especially GUI tools, that make it possible for an end user to view and query data, perform data analysis, and otherwise

interact with the data. Ideally, such tools can be extended with "plug-ins" or other techniques for working with special kinds of data. Examples include:

- HDFView, described above.
- Interactive Data Language (IDL) and MATLAB: scientific data analysis languages and GUIs, both of which support the HDF5 format, and also support scripting capabilities.

The next phase of the research work will need to adapt and extend existing tools to support many of the same set of capabilities that the APIs will make available, but with the end-user in mind, rather than the application developer. For example, HDFView might employ GDAL to read and write raster geospatial data formats. HDFView may also be adapted to invoke certain applications that give access to encapsulated files, such as launching a GeoPDF viewer when a GeoPDF file is within an HDF5 file.

In addition, certain data conversion tools will be needed, such as export/import tools for commonly used formats listed above: NITF, GeoPDF, Shapefiles, and GeoTIFF.

5 Beyond maps and images: battlefield geometry

A new conceptual framework.

Current and more importantly future military missions are generating increasing demands for new capabilities in adapting, understanding, deciding, and acting upon tactically significant information sources, according to the Defense Science Board "2006 Summer Study on 21st Century Strategic Technology Vectors" (Defense Science Board 2007):

> Counter-stealth has supplanted stealth as a critical need, since it is U.S. adversaries who are able to operate hidden underground and hidden in plain sight among civilians. The capabilities needed for such counter-stealth operations *are ubiquitous observation, recording, and archiving of difficult target data and being able to rapidly extract useful information hidden in massive clutter*. Precision has expanded from "hitting what you aim at" into tailoring effects to the circumstance, including minimizing counterproductive effects. Lastly, tactical ISR—seeing deep—can be viewed now as the much broader challenge of mapping the human terrain, including foes, ourselves, and others. [Italics added.]

Traditional approaches to Battlefield information management, focusing on geospatial information and modeling, fail to provide the technologies for effective human, social, cultural and behavior modeling that will be key to addressing "wicked" problems. The DSB study identifies four new critical capabilities to meet the demands of today's missions: "human terrain preparation, ubiquitous observation and recording, contextual exploitation, and rapidly tailored effects (with computational speed implicit in all)."

Against the backdrop of these capabilities, new data management requirements emerge, along with an expanding vision for organizing information to support Battlefield commanders and their subordinate staff elements. In the paper *Battlefield Geometry*, Dr. Michael Stein uses the term “battlefield geometry” to describe “the relationship among conceptual objects necessary to represent the modern battlefield.” (Stein 2009) The battlefield geometry concept envisions a semantic solution space that includes “not only geospatial concepts, but also other temporal, social, and cultural concepts.”

Figure 14 describes the data information types for describing socio-cultural problems, issues, and factors. These information types are rather unique and differ from what we traditionally encounter in geospatially-focused Battlespace terrain reasoning systems. The ability to concurrently assimilate these new information ‘constructs’ with traditional data sources is imperative to the achievement of a successful transition to the envisioned semantic solution space.

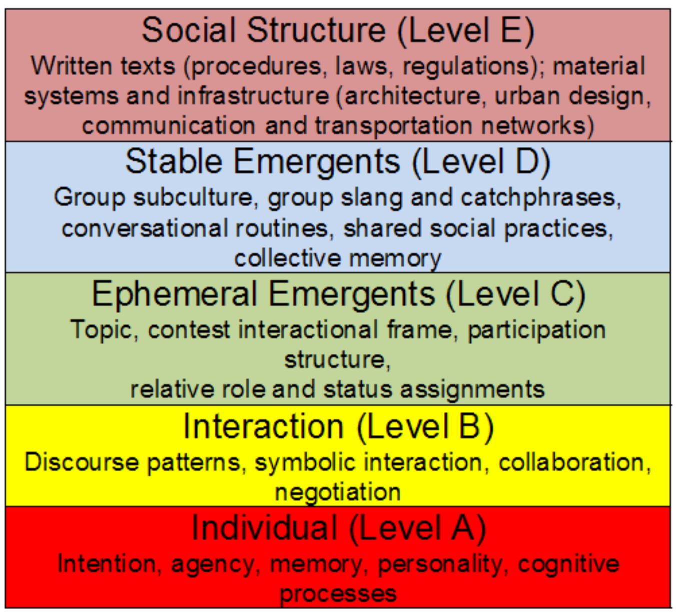


Figure 14. Data and information types for socio cultural representation. (Stein 2009)

Redefining maps

The concept maps described in earlier sections need to be extended to incorporate these new geoprocessing strategies. This is a key challenge area for ongoing and proposed work efforts within the ERDC-TEC research programs. This in turn raises the primary questions that we have been addressing throughout this paper, namely, “How do we represent these new kinds of information and how do we address the large-scale storage requirements to adequately support future military operations?”

Battlefield Geometry divides the information representation into three facets or views:

4. Geotemporal facet: "the attributes of objects and relationships that users normally think of as geospatial, with the addition of temporal attributes."
5. Social network facet: "attributes associated with various agents, groups and organizations and the beliefs, desires and intentions that they hold."
6. Events and artifacts facet: "events represent the sensed or inferred atomic elements that are composed into the actions and activities that comprise behaviors." Artifacts are "entities that allow, support, or are correlated with the events or their purposes and objectives."

The Battlespace geometry technical approach would endeavor to represent each of these facets in ways that enable sufficient understanding, open access, and efficient computation. However, representing an individual object in each facet in isolation (i.e. non-coincident data bins) is of course not enough to achieve the 'linking' necessary to perform analytic operations over facets. There must be an overarching mechanism (e.g. HDF5 solution space) that enables the information within and across facets to be integrated and coincident on both a semantic and geospatial level. Traditional approaches to data representation, which includes relational databases, spatial databases, and object-oriented databases, are insufficient for some of these more complex operations. Classical approaches to data representation "suffer from the problem of trying to construct, maintain, and operate on a data structure which does not express the relational and geometric structure of the data in a unified intrinsic framework."

To accommodate this expanded semantic coverage in a way that supports more rigorous mathematical and computational modeling, Dr. Stein proposes to represent the combination of spatial and socio-cultural information within a more robust topological space such as a "simplicial complex-based data structure." This approach gains several advantages over traditional approaches, including

- Enough expressive power to represent arbitrary relational information.
- A natural hierarchy for multi-source integration and multi-level representation.
- Processing that takes advantage of a computational framework rather than SQL based I/O intensive framework.

- Exploitation of the intrinsic geometry of spatio-temporal and other partially ordered data.

This innovative approach to representing the information and subsequent knowledge in a “concept container” that “can contain arbitrary content, has value-added attribute metadata to define the content and its characteristics, and can include links to other containers.”

Enterprise architecture and the HDF5 soup

A layered enterprise software architecture is proposed, as illustrated in Figure 15. The user interacts with the system through the visualization layer, where tools provide conceptually meaningful interfaces to the data. The prototype HDFView concept map outlined in Chapter 4 would be one such example of a visual interface. At the Enterprise Layer, Dr. Stein coins the phrase “HDF Data Soup,” which serves to represent the heterogeneous collection of data about a particular AOI, as in our example. There will certainly be other data sources, such as concept maps, and metadata catalogs, but our vision is that most data types can be either stored in HDF5 or referred to from HDF5 as externalized associations (i.e. links). In the latter case, for example, there may be data in a relational database management system that is requested from an HDF5 solution space, providing up-to-date information about a specific tactically significant instance, such as the number of persons inhabiting a building at any point in time.

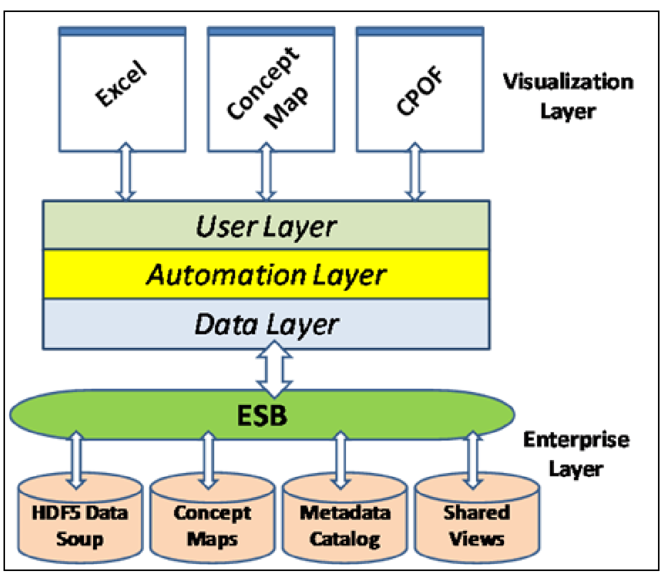


Figure 15. Enterprise Architecture Concept for the Computational Framework.

In Chapter 3, we saw examples of *geospatial* ingredients that may reside within the HDF5 soup. But what about the other two information facets: the *social network* and the *events and artifacts* facets? Each poses its own data management challenges and it needs to be further demonstrated through design and implementation that HDF5 can represent these information types in conjunction with the baseline geospatial datasets. In addition, we need to determine how HDF5 might best represent and organize the structures that integrate the three facets into one HDF5 solution space. How would HDF5, for example, represent a simplicial complex data structure? There are precedents for this, such as the Sets and Fields model (Miller 2001), but the best way to proceed with this approach remains an open question at this point in the research.

A socio-cultural analysis use case

Battlefield Geometry provides three use cases for representing heterogeneous data in HDF5: a structure for geospatial and socio-cultural analysis, a structure for encoding a hypothetical set of courses-of-action, and a top-level concept map structure of local weather knowledge.

The first use case exemplifies our interest in this chapter. It describes a collection of information entities to support socio-cultural analysis for the region of Colombia in South America. Figure 16 illustrates the variety of different formats, sources, granularities, and qualities of information that might be called upon to support socio-cultural analysis. In the example are a boundary map (Shapefile), media documents (unstructured text), extracted events and locations (relational table), and others.

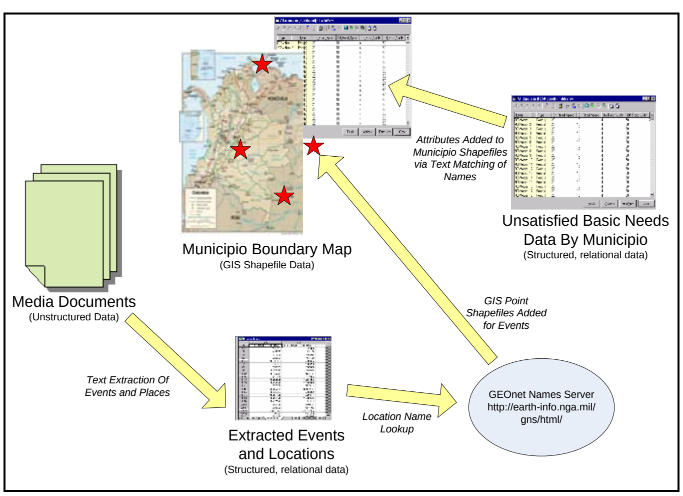


Figure 16. An example heterogeneous data structure for socio-cultural analysis.

It is not hard to imagine visualizing this collection of data sources and their interrelationships when stored in HDF5. An HDFView visual interface similar to the plug-in shown in Figure 12 could be constructed to match the workflow in Figure 16. An analyst could drill down (i.e. decompose) on any of the functional objects in the figure in one of two ways. (1) Any HDF5 objects could be investigated using the built-in viewing capabilities. (2) Objects for which there are external applications (e.g. a Shapefile) could be viewed by having HDFView launch the corresponding application.

The ability of HDF5 to deal with heterogeneity is particularly applicable in this instance, as is the availability of appropriate linking mechanisms that can relate both internal and external objects, an important capability that is demonstrated in Figure 12. The same approach may be used to describe key relationships among socio-cultural data objects.

However, as we have seen, another more comprehensive computational structure is fundamental to enabling the battlefield geometry vision. This proposed structure is suggested by the arrows in Figure 16, which "represent various transforms or relations between elements of the various datasets. Each arrow is a composite of text extraction or matching, search, query, or inclusions operations. These operations not only capture a workflow of software and tool operations but encode an analytic perspective and hypothesis about the relationship between the events and their locations and the attributes of the populations that are the agents or context of the events."

Obviously more will be needed than standard HDF5 links to represent this structure. How to accomplish this is a question that will require follow-on investigations. There have been promising efforts to model such complex relations in the physical domain, such as the Sets and Fields data model, and the "Fiber Bundle HDF5 format" developed by Werner Benger (Benger 2001, Buleu 2007) and used in a study of the Katrina disaster (Venkataraman 2006). These provide a plausible basis upon which to create models that incorporate ideas in the socio-cultural domain.

Another challenging aspect of this example is how the contents of a collection can change in response to dynamic conditions. Data structures must be found that can grow and adapt to varying semantics. The belief is that HDF5 container abstraction can support this in a sufficiently flexible way.

HDF5 itself will not embody all of the relevant semantics of a domain-specific application, much less those of socio-cultural structures and operations such as these. HDF5 provides many of computational and storage structures needed to describe the model and store the enormously varied and dynamic data and metadata, and that is its main contribution to this endeavor. Domain-specific structures and operations such as those we have been describing would be implemented by *using* these HDF5 structures and operations.

Even with HDF5's flexibility, it is likely that some enhancements to HDF5 may be needed to best accommodate socio-cultural models. Just as we saw the importance of developing the structures and operations in HDF5 needed to support geospatial concepts and data, it will be necessary to develop similar structures and operations corresponding to socio-cultural concepts and data. For example, the semantics of HDF5 link structures may need to be extended to assist in representing the relationships among data objects (i.e. Semantic Tags). In a similar vein, *Battlefield Geometry* identifies these operations that must be available for the data structures in the example:

- Reconstruction of the combined GIS file using new media sources or updated unsatisfied basic needs data;
- Comparison with other hypothesized models of the relationships between media events and unsatisfied basic needs;
- Abstraction and representation of the pattern of events and the municipio context, the pattern of the data sources and transform operators for this geographic region, or even the pattern of the overall workflow.

The road ahead

The extension of battlefield geometry to the social networks, events and artifacts facets is a clear necessity to address the previously discussed "wicked" problem space. The data and information management/technology challenges in doing this are not trivial, but there has been a ground-level foundation to build upon as a result of this work effort. The translation of these multifaceted representations into an HDF5 based format for the purpose of conducting data fusion for future military decision-making offers an important component to this foundation due to the technologies' native capabilities for organizing and managing a large-scale, highly complex computational environment.

6 Conclusions

In a press release from February 2009, the ERDC describes its mission as follows:

> The Army Geospatial Research and Engineering Division will continue TEC's legacy of providing geospatial support and products to Warfighters, but will expand its mission to support the Army's Battle Command Systems, facilitating dissemination of relevant geospatial information to every level across the dynamic Battlefield environment. Additionally, the center will coordinate, integrate, and synchronize geospatial information requirements and standards across the Army, as well as develop and field geospatial enterprise-enabled systems and capabilities to the Army and Department of Defense.[1]

Emerging technologies and the explosion of available information from a wide range of sources present opportunities to help achieve this mission. A key to exploiting these factors is *computational efficiency:* being able to get answers fast from complex, large scale, heterogeneous data, and over time being able to seamlessly collaborate within a net centric data landscape that includes data volumes and data sources that are constantly growing and changing.

The approach described in this paper uniquely addresses this combination of challenges, and strives to do so at an overall lower cost. The capabilities we have identified include:

- A software infrastructure and generic file structures capable of ingesting, integrating, and storing the wide variety of geospatial (e.g. feature classes, imagery, value-added) and socio-cultural data needed for Battlefield decision making.
- An Open Systems Engineering (OSE) approach to data integration designed to eliminate the profusion and confusion of proliferating data models and formats by mapping data to a single, all-purpose container model, format, and access technology, while at the same time providing alternative appropriate conceptual views of the data.

[1] US Army Corps of Engineers News Release, Release No. A-05-09, February 23, 2009.

Experience gained from this study has highlighted a number of important lessons to guide future work:

- **Need for standards.** Because of the great variety of data, there is no obvious common body of standards to follow in developing a unified data model. Further consideration needs to be given to community standards and practices as we expand the types of data to be included. Appropriate implementation of formats (grids, polys, clouds), attribution (descriptive fields) and exchange (XML, BML) standards will be imperative.
- **The value of concept maps.** Concept maps can be a very powerful tool in helping to understand a Battlefield information space, especially when used to tease out non-linear relationships and dynamic interactions. These maps will play a key role in developing future methods and tools for rapidly organizing and visualizing Battlefield data. Concept maps should remain a focus as this work continues.
- **Metadata requirements.** The capacity for managing heterogeneous collections of data is important, but the enhanced ability to store uniform geospatial metadata, as well as problem-specific metadata, is equally critical for a full, common understanding of the Battlefield information space and robust Battlefield decision analysis. This study also demonstrated a need for additional *structural* metadata to describe concepts and data organization. Future research should focus on these two metadata requirements.
- **Agile approach.** It is very challenging to implement a military terrain reasoning and decision-making system. Such a system involves highly complex data management operations and deals with very heterogeneous information. The system has uncertainty in design and the requirements may change throughout the lifecycle of the project. Agile software development methods can produce a small and workable subset of a system in a short period of time and allow customers/users to evaluate the system at every stage. Developers are able to change the design and direction as needed. The Phase 1 software development work of the ERDC-HDFView plug-in has demonstrated that the agile approach is very efficient and productive.

Benefits to the Army

This new approach to data organization and management gives the Warfighter immediate access to high volume Battlespace information from within an optimized solution space, where such information may otherwise be accessible only after costly delays caused by accessing multiple disjointed sources and/or services. Battlefield solutions should be both more accurate and timelier because of these new information management capabilities. The ability to manipulate the latest and highest resolution imagery, dynamic sensor data feeds, geocomputational data-sets, other key infrastructure variables, a richer mixture of geographic information (political, economic, social-cultural) and a wider variety of non-proprietary decision-support tools make for a rather unique solution space to meet future Army digital mapping requirements.

The Warfighter would encounter information in a single, integrated form, greatly shortening the time it takes to understand and act. Through an integrated interface, specially-adapted to the situation, a soldier for example would be able to visually assess Battlefield conditions, pan and zoom over the Battlefield using a variety of image modalities, make queries about the state of roads, buildings, and persons of interest, and combine these and other georeferenced information to produce an unified view of the Battlefield. The same solution space may also allow the soldier to investigate alternate scenarios by launching auxiliary processes, such as visualizations and weather simulations.

The new approach attempts to simplify the increasingly complex data and information management process. More informed decisions and solutions can be generated with less specialized training because the Warfighter has information in a form that reflects detailed Battlefield content vs. context in the most meaningful, accessible ways. General misinterpretations within the decision environment could be avoided because of the greater depth of potential analysis enabled. Military planning operations should be more streamlined and comprehensive because of the enhanced computational ability to consider alternative scenarios (i.e. predictive capability) along with migrating away from often stove-pipe GIS technologies.

Because of the widespread adoption and support for the underlying HDF5 technologies in the scientific community, as well as the high potential to integrate within emerging future GIS enterprise architectures, Battlefield information systems could be developed more rapidly, at a better cost-benefit ratio, and with richer (more realistic) information content than ever before, thereby reducing delays that can ultimately inhibit success and cost lives.

Migration of GIS to HPGIS

We have identified several ways in which the HDF5 technologies can be leveraged, and even facilitated from the evolution of current GIS to "High Performance" GIS (HPGIS) taking place in the community of practice today. The fundamental challenge of applying HPGIS methods to critical geo-spatial reasoning problem areas is the migration of traditional (i.e. legacy) GIS data structures and analytic processes into a High Performance Computing (HPC) compatible environment. These transitions of methods (GIS to HPGIS) focusing on the exploitation of new, large scale computing resources, will resolve limitations in handling high fidelity data, complex information types, and dynamic geoprocessing. As Berry states, this next generation solution space will be "built upon an entirely new set of analytic tools, geo-referencing framework and a more realistic paradigm of geographic space (Berry 2007)." Table 5 encapsulates how current GIS concepts may evolve to HPGIS via a HDF5 'computational' strategy.

Table 5. Current operations in GIS to Future state of HPGIS crosswalk.

Traditional GIS Concepts	Future *(5-10 years)* HPGIS Techniques
Database Access – Disk I/O	Memory resident, Real Time transactions
Table/Attribute SQL queries	Dynamic pointers, API level referencing
Native resampling resolutions	Scalable data representations (*global, local*)
Externalized spatial index	Embedded Geometry (*dimensions, coordinates*)
Proprietary / COTS - S/W, H/W	Platform independent, Interoperable architectures
Hard boundaries, abstractions	Implicit patterns, Non-linear solutions (Wx, *t*)
Discrete layers of data	Linked hierarchical stacks, Novel relationships
Points, lines, polygons	Organized groups, Complex topologic structures
Standard Digital Products	Non-Traditional Sensors/Sources

A critical path to HPGIS (sample use case)…

Some earlier foundational research efforts investigating this technical challenge area were focused on translating exploratory data analysis methods from the realm of BioComputation into the GeoComputation domain, specifically for military terrain reasoning applications. Gleaned from the background review of the BioComputational publications was a particular HPC life sciences modeling technique referred to as "In Silico - Biological experiments carried out entirely in a computer"; used in this case for the discovery of cellular interactions between certain pathogens and native immune defenses.

It is quite intriguing how the simulation algorithms would utilize massive complex datasets to predict movements, behaviors, and outcomes within an HPC environment. The overall implementation of the model was closely aligned with some preliminary ideas and concepts on how to develop a topographic sciences (vs. life sciences) 'framework' for enabling a higher order terrain reasoning capability utilizing HPC technologies, hence addressing the GIS to HPGIS technical challenge.

The result was a notional functional mapping (early concept map of sorts) of a terrain reasoning solution space based upon the 'In Silico' Bio-model. The Predictive GeoInformatic Science (PGIS) diagram below has served as a useful reference in crafting the HDF5 data management system approach (i.e. HPGIS) designed to bridge the GIS to HPGIS technical gap that we have been outlining in the aforementioned chapters.

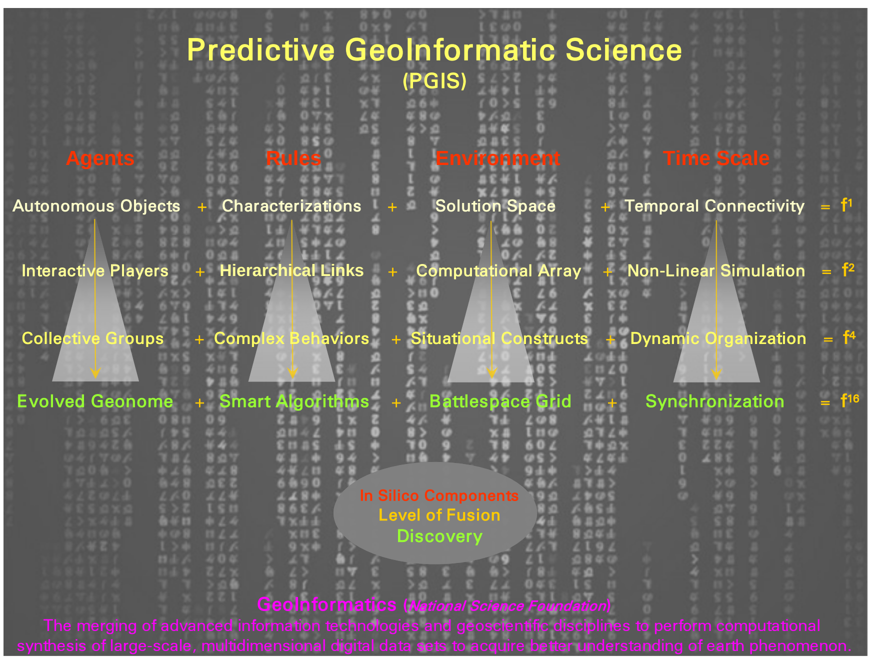


Figure 17. PGIS example using 'In Silico' components.

Future work areas (recommendations)

We have presented a new, innovative data management approach to how significant terrain information may be more effectively/efficiently organized, integrated, accessed, and analyzed. This approach addresses not only current but future challenges faced in the digitized Battlefield, in coincidence with any civil and environmental events that may require rapid and comprehensive mastery of complex, dynamic information spaces. To achieve success in adopting this approach, we will provide some applicable recommendations to address short, mid and long term Research and Development support areas.

Short term agenda

- Translators - Simply providing compatibility between common geospatial formats and HDF5 will provide important interoperability with existing geospatial tools. This could be achieved by instantiating the GDAL data model in HDF5, and creating APIs and data conversion utilities for managing these files, including tools to convert between HDF5 and common geospatial formats (e.g. Shapefiles). Encapsulation of non-geospatial formats will also be important and can be provided by defining storage and access protocols for such formats, including common media formats.
- Context - The sample application described in Chapter 3 demonstrates a general approach to data organization and management, but lacks much of the infrastructure needed to create a truly useful product. An important key to usability will be to add semantic information that will enable applications and users to understand the meanings of the data objects, as well as the intended conceptual views and relationships of the content that they are consuming.
- Interfaces - HDFView can be adapted readily to make these enhancements available in a visual environment that brings the underlying data to a user in a convenient and meaningful form. Examples of HDFView enhancements include the ability to launch applications, to export and import common geospatial formats, and to interpret semantic information for users through meaningful visual representations of concept maps and data. In addition to HDFView, other common geospatial APIs, tools, and GUIs may be adapted to support the interaction of these new structures and data (e.g. ESRI).

Midterm objectives

- BAA Phase 2 – Follow-on work efforts should investigate more robust Geo-Representation methods for creation of Spatio-temporal (geometric, dimensional, temporal, dynamic) enabled constructs. The research should include modification of object attributes and/or metadata to facilitate the linking and exchange of information via XML-like semantic level operations. It will also be important to experiment with the handling and conflation of large, scalable datasets from varied sources and configurations within the HDF5 solution space container.

- GIS Community –Until the HDF5 geospatial data management approach matures to the point that other users can adopt the technologies into their framework with relative ease, these conceptual structures will remain somewhat of a niche in the geospatial domain. We must identify any deficiencies in the areas of functional compatibility and systems integration and begin the process of working within the community for common solutions. Whether it is from an architectural point of view (Enterprise Services) or at data analysis level, the HDF5 approach can be of great benefit to solving some of the 'hard' geospatial problems we face today.
- Web 2.0 – The next generation of web development will facilitate information sharing, interoperability, and collaboration. An evolution in traditional geospatial data management operations may be required to handle these upcoming services and applications that will rely upon extremely complex data operations and interactions as described in Torrens (2009). The HDF5 technologies will be a viable mechanism to enable these operations in the upcoming World Wide Web paradigm.

Long-term goals

- Duality – The focus of HDF5 thus far has been primarily on the computational and analysis side, but there is potential to also use these structures on the visualization side of problem areas such as in modeling and simulation. These conceptual structures may serve as input to simulators for fly-thrus and various other virtual war gaming exercises requiring detailed (high-fidelity, multivariate) Battlespace representations. Imagine for instance being able to visualize realistic, real-time scene portrayals of all of the terrain parameters in a compiled database (e.g. HDF5) that could provide optimized feeds to the visualization engines.
- Virtualization – The advent of abstracting computer resources across multiple platforms (operations systems, applications) in this prevailing area of networking technology will pervade all aspects of current GIS technologies. Geospatial data management strategies that adapt, persist, and more importantly move forward within this computing environment will provide the best alternatives for supporting military operations reliant upon these key (geo)services. The HDF5 technologies are well positioned to take advantage of these resources due to the open systems nature and physical implementation of HDF5. For example, HDF5 offers

parallel computing support, resident memory addressing, high performance I/O-storage models and open API access.

- Army Future Force - The opportunity to meet Warfighter geointelligence needs will occur through exhaustive research and development in the strategic areas of Joint Operating Environments, Preparation of the Battlefield, Geo-Informatics, and Enterprise Command Services. Each of these areas presents a unique set of challenges and issues for researchers to resolve. There does exist one common denominator within these underlying technical thrust areas: the requirement for a unifying data/information structure to drive the analysis and decision-making spectrum of operations. This structure (a.k.a. geocomputational framework) must address a wide range of concerns to include fidelity, dimensionality, scalability, interoperability, computational optimization and more rigorous 'physics' based terrain representations (Nedza 2006), a very ambitious set of needs that will rely upon many innovative solutions. The introduction of HDF5 technologies into this solution space provides an important step in the appropriate direction.

References

Benger, Werner, et al., 2001. "Data Description Via a Generalized Fiber Bundle Data Model," Proceedings / *Tenth International Symposium on High Performance Distributed Computing*. San Francisco, CA: 10th IEEE International Symposium on High Performance Distributed Computing.

Bennett, K, Wen, S, Robertson, J. 2007. *An Overview of ARL's Multimodal Signatures Database and Web Interface*. ARL-TR-4337. *http://www.stormingmedia.us/-authors/Bennett__Kelly.html*

Berry, Joseph K. 2007. "Innovation Drives GIS Evolution," *GeoWorld*.

Buleu, Ana Elena. 2007. *An Ontological Scheme for Specifying Time in HDF5*, Proceedings of The National Conference On Undergraduate Research (NCUR) 2007. San Rafael, CA: Dominican University of California.

Defense Science Board. 2007. *Defense Science Board 2006 Summer Study on 21st Century Strategic Technology Vectors*, Volume 1, Main Report.

Engineer Research and Development Center (ERDC). 2008 Broad Area Announcement. *http://www.mvk.usace.army.mil/contract/docs/FY2008.BAA.pdf*.

GeoPDF. 2007. *Geo-referenced Portable Document Format (GeoPDF) Fact Sheet*. U.S. Army Topographic Engineering Center, Engineer Research and Development Center.

ISO/IEC. 2008. *Information technology — Semantic metadata mapping procedure (SMMP)*. ISO/IEC WD 20943-5, Committee ISO/IEC JTC 1/SC 32/WG 2. *http://metadata-standards.org/metadata-stds/Document-library/Documents-by-number/WG2-N1201-N1250/WG2-N1217-WD_SMMP_20081119.pdf*.

Miller, M.C., et al. 2001. *Enabling Interoperation of High Performance, Scientific Computing Applications: Modeling Scientific Data with the Sets & Fields (SAF) Modeling System*. ICCS-2001 v2.

Nativi, S., J. Caron, B. Domenico. 2008. "Unidata's Common Data Model mapping to the ISO 19123 Data Model," *Earth Science Informatics*, 1(2): 59-78.

Nedza, John A. 2006. *Terrain Reasoning Array Paradigm (TRAP) Concept Paper*. U.S. Army Topographic Engineering Center, Engineer Research and Development Center.

National Imagery Transmission Format (NITF). 2006. *Department of Defense Interface Standard, National Imagery Transmission Format Version 2.1*. MIL-STD-2500C.

NRC. 2005. Committee on Network Science for Future Army Applications. *Network Science*. National Research Council, National Academies Press.

Schmitt, John F. 2006. *A systemic concept for operational design*. Quantico, VA: Marine Corps Warfighting Laboratory.

Stein, Michael C. 2009. "Battlefield Geometry Final Report."

Torrens, Paul M. 2009. *Process Models and Next-Generation Geographic Information Technology*, ArcNews Online. *http://www.esri.com/news/arcnews/summer09-articles/process-models.html*.

Venkataraman, S. et al. 2006. *Visualizing Katrina - Merging Computer Simulations with Observations*, Umeå, Sweden: Workshop on State-of-the-Art in Scientific and Parallel Computing.

Versatile Information Systems Integrated ON-line *(VISION)*. 2008. Aberdeen Test Center. *http://vision.atc.army.mil/*.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
August 2010

2. REPORT TYPE
Final report

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
Transforming the Geocomputational Battlespace Framework with HDF5

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Mike Folk, Peter Cao, and John A. Nedza

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Topographic Engineering Center
U.S. Army Engineer Research and Development Center
7701 Telegraph Road
Alexandria, VA 22315-3864

8. PERFORMING ORGANIZATION REPORT NUMBER
ERDC/TEC SR-10-1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Success in the modern Battlefield depends on effective management of data. Every operation is unique, making it impossible to build systems that address all of the needs of a specific mission in advance. Information needs arise under intense time pressure, and the available information is often incomplete or uncertain. Data entities vary enormously in scale and resolution, and can exhibit a great deal of heterogeneity. These challenges will only intensify in the future, as networks of sensors increasingly collect and transmit huge amounts and varieties of valuable data, scaling from synoptic images to the vital signs of individuals.
The complexities inherent in mission operations make the information management task an immense challenge, one which must be addressed in part by focusing on how information is organized, integrated, accessed, and analyzed. Toward that end the goals of this investigation are to identify the role that HDF5 can play as a data management platform for Battlefield military operations, to demonstrate the use of HDF5 visualization tools to present operational data, and to identify a research and development plan to develop a prototype geoinformatic data management system based on HDF5.

15. SUBJECT TERMS
Battlefield data management
Concept map
Geoinformatic data management system
HDF5
Heterogeneous data structure
Socio-cultural analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED		74	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18