AIR WAR COLLEGE

AIR UNIVERSITY

DATA INTEGRATION: CHARTING A PATH FORWARD TO 2035

by

Jill E. Singleton, Lt Col, USAF

A Research Report Submitted to the Faculty
In Partial Fulfillment of the Graduation Requirements

14 February 2011

Disclaimer

The views expressed in this academic research paper are those of the author and do not reflect the official policy or position of the US government or the Department of Defense. In accordance with Air Force Instruction 51-303, it is not copyrighted, but is the property of the United States government.

Contents

Illustrations

Page

About the Author

Lieutenant Colonel Jill E Singleton is an intelligence officer and was commissioned as a Distinguished Graduate from the United States Air Force Academy in 1990. Her education includes a Master's in Public Policy from the Kennedy School of Government, Harvard University; Master's of Strategic Intelligence from the Joint Military Intelligence College; Air Command and Staff College; Joint Professional Military Education, Phase II, Joint Forces Staff College; and Air War College.

Lt Col Jill Singleton most recently served as Commander, 8th Intelligence Squadron, Hickam Air Force Base, Hawaii, providing 24x7 exploitation and dissemination of near-real-time intelligence from the Predator, Global Hawk, and U-2 platforms. Lt Col Singleton has served in squadron, group, major command, Headquarters Air Staff and Joint positions. She has experience in weapon system acquisitions and training. She is a fully qualified Joint Specialty Officer and previously served as Deputy Director, C2 Demonstrations and Assessments, Joint Systems Integration Command.

INTRODUCTION

Knowledge is Power.

--Sir Francis Bacon
Religious Meditations, Of Heresies[1]

In 2035, intelligence collection systems will include autonomous systems from the size of bugs to blimps, each equipped with multiple sensors. These systems will be trustable, flexible, survivable, composible, and agile.[2] The future collection suite will build on many platforms in use today, to include the Global Hawk, Reaper, and Predator Remotely Piloted Aircraft (RPAs). These diverse sensors platforms will be complemented by exponentially expanding storage and computing capabilities capable of emulating human intelligence.[3] Artificial intelligence will enable the integrated platform and sensor family to analyze, form opinions, make recommendations, and task collection. In this brave new world, data will be the coin of the realm. But, how do we best make use of this heterogeneous and ever expanding data?

The Air Force and Department of Defense (DoD) as a whole would benefit greatly from an increased focus on data integration as a strategic enabler. Well-executed data integration saves limited personnel resources and contributes to knowledge creation. Data integration solutions that are designed to evolve from the outset offer the best potential for quick response and acquisition savings. These benefits extend to legacy and future capabilities alike. Optimally, data integration aims at maintaining valuable data complexity while overcoming accidental complexity caused by stovepiped data silos. This accidental complexity takes the form of "physical, representational, structural, and semantic barriers between data sources, types and domains."[4] At its core, successful data integration enables improved service and agency operational integration. This paper discusses the potential for data integration solutions through

2035 with a focus on where to invest now to begin tapping the potential of the growing data stockpiles.

Data access, integration, and security are linked to America's military effectiveness and ultimately, national security. While this paper will focus on intelligence data integration, the findings are of use to any field that suffers from the data integration challenge. Several likely candidates include the logistics and medical data stores. Coherent data integration offers the opportunity to make best use of military capabilities and resources. An added consideration is the growing civilian capacity to access and integrate diverse data, which puts increasing power in the hands of superempowered individuals.[5]

The term data integration has a wide range of uses and interpretations. This paper uses the Gartner definition of data integration. Gartner provides a quarterly assessment of the status of data integration solutions and an assessment of the industry leaders based on vision, leadership and ability to execute.[6] Gartner defines the discipline of data integration as "practices, architectural techniques and tools for achieving consistent access to, and delivery of, data across the spectrum of data subject areas and data structure types in the enterprise to meet the data consumption requirements of all applications and business processes."[7] Intelligence business processes are the focus of this research paper.

---

[1] Sir Francis Bacon, *Religious Meditations, Of Heresies*, 1597. http://www.quotationspage.com/quote/2060.html.
[2] Caroline King, "Cooperative Control. Overview," (lecture, Air War College Blue Horizons Team, Air Force Research Labs, Wright Patterson Air Force Base, OH, 22 September 2010).
[3] Ray Kurzweil, *The Singularity is Near: When Humans Transcend Biology* (New York, NY: Viking Publishing, Sep 2005), Kindle location 716-29.
[4] M. Andrew Eick and Suzanne Yoakum-Stover, "Fixing Intel and Operationalizing Data – The Data & Processing Syndicate," www.imintel.org.
[5] Daniel Goure, "Wikileaks Dilemma: How Does a Nation Fight a Superempowered Person?" Lexington Institute Early Warning Blog, 6 Dec 2010, http://www.lexingtoninstitute.org/. This article references Thomas Friedman's *The Lexus and the Olive Tree* definition of a superempowered individual and asserts that Julian Assange may be the "first truly superempowered individual."
[6] Ted Friedman, Mark Beyer, and Eric Thoo, "Magic Quadrant for Data Integration Tools," *Gartner*, 19 Nov 2010, http://www.gartner.com/technology/mediaproducts/reprints/sas/vol7/ article4/article4.html,
[7] Ibid, 2.

DATA INTEGRATION BACKGROUND

No matter what anybody says, it's pathetic.
-- Maj. Gen. John M. Custer, commanding general of the Army intelligence center, said of the information sharing environment.[8]

Demand for data integration capabilities is growing with the rapid increase in available data and diversity of users. "Contemporary pressures are leading to an increased investment in data integration in all industries and geographic regions."[9] The first portion of this section will look at state-of-the art data integration and application solutions that are already in use today. This will provide a window into the power of integrated data. The second section will focus on the data integration enablers that are of most interest today.

Turning to examples of the power of data integration, Los Alamos National Laboratory is using a powerful data mining tool to focus analysis. The Data Knowledge Management Tool revealed key words across open source articles related to emerging chemical biological threats.[10] Data mining capabilities rely on the data to be available though before it can be searched. In the case of foreign language sources, the non-trivial challenge of at least basic machine-level translation is also critical before such a tool can work its magic. These may seem like obvious statements, but much of the data integration challenge lies in the fact that data is not in complementary formats, neatly metadata tagged, or readily accessible.

Also at Los Alamos, Dr Vestrand and the "Thinking Telescopes" team are taking portions of the database to the sensors and enabling precise and quick focus on celestial events of interest. Their challenge was how to break out anomalies in the universe, a pretty massive data source if there ever was one, quickly enough to drive more focused collection before the transient event was over. Humans lack the attention span, response time and memory (database) required to

monitor the data to recognize important variations and respond. The "Thinking Telescopes" team seeks to meld human knowledge with machine abilities. The critical data, called the "hot database," is with the sensor and ready to respond based on rules previously identified by a scientist and encoded into the system by a software developer. When the previously identified high-priority anomaly is picked up and recognized by the telescope system, it drives additional or more focused collection against the anomaly. In addition to the hot database which drives action within seconds, there is a warm database that is close to the collection, broader in nature and designed for decisions that take minutes. All of the data coalesces in the cold database, a central data storage facility that is ideal for finding patterns enabled with more powerful computers.[11] This autonomous, real-time, robotic interrogation and surveillance can be simply depicted as:

Surveillance➔ Anomaly Detection➔ Interrogation➔ Data Fusion
Context from DB Interleaving for max impact

AFRL's sensors directorate is working to integrate varied sensor data in the Layered Sensor Operations Center. This effort is early in development, but has the potential to spin off valuable concepts in the next 5-10 years.[12] Addressing the full joint and national sensor integration challenge is likely to take significantly longer given the many Service and Agency elements who "own" the data in its native format. The integration of the military Services operationally on a daily basis has increased the necessity to break down information barriers, but the challenge still is how to integrate the varied data in a meaningful way once it is out of its database and system-prescribed containers.

There is broad awareness of data and system integration challenges, but solutions often try to balance the necessity of pulling all of the data into centralized repositories or dictating a specialized structure that doesn't meet the needs of all of users. In reality people need to use

data at multiple levels in multiple ways, much as the Thinking Telescope team does. A single hardwired solution is rarely sufficient to meet user needs. Dr. Jim Gray described this challenge as the "Fourth Paradigm." Dr Gray's first three paradigms were experimental, theoretical and computational science. The Fourth Paradigm involves an "exaflood of observational data" that is threatening to overwhelm scientists. A new generation of computing tools to "manage, visualize and analyze the data flood" is required and will lead to a new computing landscape. Dr Gray crusaded "It's the data stupid" and pushed for integration of scientific discovery and computation. The goal isn't building the biggest computer but getting all of the science literature and data online and interoperable.[13]

There are already early examples of powerfully-integrated data and information available for use by the average citizen. Dr Gray helped launch the integration of astronomical data which led to the Worldwide Telescope (WWT). WWT provides a view of the incredible potential of data aggregation in a system and data-agnostic, user-friendly and accessible format.[14] WWT is not alone, other data aggregators include Google Sky; similar capabilities are emerging for neurobiology, geography, hydrology, and the social sciences.[15]

Cloud computing is used frequently in descriptions of the Web and touted as a data integration pathway, but just what is it?

> Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.[16]

The government frequently employs private or hybrid cloud computing because it provides the benefits of elasticity and network services while lessening security issues, bandwidth concerns, and control over user access and network processes. [17] A community cloud is preferred by organizations with shared interests, missions, or security requirements.[18]

Given the promise of cloud computing it is worthwhile to understand what is inside a cloud (Figure 1). A cloud includes applications, infrastructure and a software environment or platform. All clouds do not include all of those levels, some are simply the infrastructure. This figure gives the reader an understanding of what is inside the amorphous "cloud."

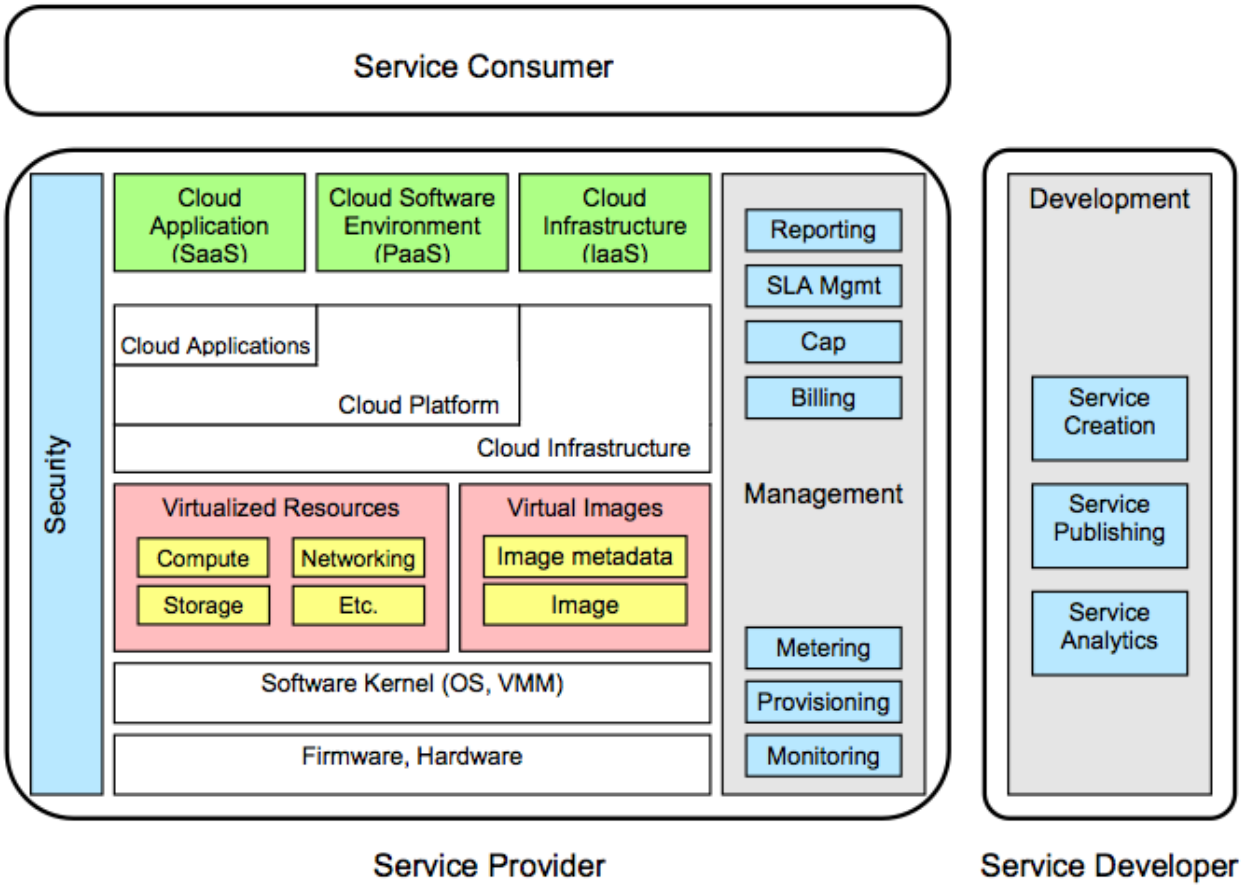

Figure 1: Overview Taxonomy of Cloud Computing[19]

While the civilian world is moving rapidly forward with cloud-based solutions, the government is hampered by lack of continued focus, unclear definition of needs, and competing organizational interests and priorities. Experts that work with the government "fear that government progress will be far slower than on the Web, or even business. We may learn the wrong lessons from Wikileaks, and hobble ourselves."[20] Even if government implements a

perfect cloud-computing environment, it is not a sufficient solution for data integration in and of itself. Cloud computing provides the elastic platform for data storage, networks, and even processing services. Investing in the cloud as the solution without further refining what is feeding the cloud will not resolve the data integration challenge.

Large amounts of data offer unique challenges and opportunities. The civilian world is responding to the same pressure as the government to deal with large amounts of heterogenous data. Handling this "big data" requires "a row-based data store powered by massively parallel processing (MPP) engines, or -- even better, according to some -- an MPP-based columnar data stores."[21] Machine-based processing may become more human. The human brain's unique power comes from its ability to perform massive parallel processing of its existing data stores.[22] In short, more diverse data in large data warehouses provides the opportunity for powerful processing to reveal more information. Layer advanced analytics onto the system and knowledge creation becomes possible.

Finally, data models are the backbone of data architecture and are necessary, but also a key challenge to data integration. Much current and past effort at integration has focused on ontology mapping or designing universal ontologies.[23] These efforts had some success but came up against the very real need for data to be bound in specific ways to enable certain processes, varied needs of different users, and the tendency of people to employ unique semantics. More recently automated metadata tagging, modularized and reusable processes, and data analytics have been moving to the fore. Master Data Management (MDM) products, which are promising and underutilized in government, learn to match "entities" across the data sources to the same identity. [24] Of note, advanced data capabilities can offer increased security while exposing appropriate data by making data about the user part of every transaction.

[8] Maj Gen John Custer, as quoted by Stew Magnuson in "Military 'Swimming in Sensors and Drowning in Data,'" *National Defense Magazine*, Jan 2010, http://www.nationaldefensemagazine.org/archive/2010/January/Pages/Military'SwimmingInSensorsandDrowningin Data'.aspx.
[9] Friedman, Beyer, and Thoo, "Magic Quadrant," 2.
[10] Beth Perry, "Emerging Threats in CB Weapons Space," (lecture, Air War College Blue Horizons team, Los Alamos National Laboratory, NM, 25 August 2010).
[11] Thomas W. Vestrand. "Thinking Telescopes." (lecture, Air War College Blue Horizons team, Los Alamos National Laboratory, NM, 25 August 2010).
[12] "COMPASE Center, Layered Sensing Operations" flyer from Air Force Research Labs, Wright Patterson Air Force Base, OH.
[13] Dr Jim Gray as quoted by John Markoff, "A Deluge of Data Shapes a New Era in Computing," *New York Times* (December 15, 2009): D2.
[14] World-Wide Telescope, http://www.worldwidetelescope.org, accessed 26 Sep 2010.
[15] Tony Hey; Stewart Tansley, and Kristin Tolle, eds, *The Fourth Paradigm: Data Intensive Scientific Discovery*. Microsoft Research: 2009. E-book, 1-35, http://creativecommons.org/licenses/by-sa/3.0.
[16] Peter Mell and Tim Grace, "The NIST Definition of Cloud Computing," Version 15, 10-7-09, 1, http://csrc.nist.gov/groups/SNS/cloud-computing/, accessed 1 Dec 2010.
[17] Cloud definitions from Mell and Grace, "The NIST Definition of Cloud Computing," 2 and Cloud Computing Use Case Discussion Group, "Cloud Computing Use Cases," white paper, 31 July 2009, 6, http://groups.google.com/group/cloud-computing-use-cases.
[18] Mell and Grace, "The NIST Definition of Cloud Computing," 2.
[19] Cloud Computing Use Case Discussion Group, "Cloud Computing Use Cases," 9. Several key terms included in the diagram are Service Level Agreement (SLA) which is a contract between the provider and consumer that includes such things as privacy, security and backup procedures; and provisioning, which is the act of assigning scalable resources to meet user need. For a complete description of cloud taxonomy and examples see the latest version of the "Cloud Computing Use Cases White Paper," with definitions and additional information on pages 3-11.
[20] Arnon Rosenthal, MITRE Corporation, 6 Dec 10, e-mail interview. His concerns were echoed by many others, but were the clearest depiction of the potential that we will throw data integration efforts out with the Wikileaks response. Data integration does involve exposing data, but this author posits that such exposure if combined with security solutions, to include data about the user and what the user is doing with the data, can provide a better solution overall for knowledge creation. A detailed discussion of security implications and data integration would be a worthwhile paper in its own right. Dr Rosenthal identified development of "rational ways to justify and manage risk/reward as a basis for access decisions" as an area worthy of further focus.
[21] Stephen Swoyer, "Crunching the Numbers on Big Data," The Data Warehousing Institute (TDWI), 1 Dec 2010, 1, http://tdwi.org/articles/2010/12/01/crunching-big-data-numbers.aspx?admgarea=news.
[22] P.W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century (*New York, NY: The Penguin Group, 2009), 102.
[23] Leo Orbst, Mitre Corporation, e-mail interview, 6 Dec 2010. An ontology is an organization of some knowledge domain that contains all relevant entities. Ontology mapping links the individual entities to each other. A universal ontology seeks to identify all possible entities of interest across knowledge domains. The challenge is pre-identifying all possible ways in which data entities interrelate or even creating fully exhaustive ontology.
[24] Arnon Rosenthal, 6 Dec 2010.

AIR FORCE AND DOD CHALLENGES

We're going to find ourselves in the not too distant future swimming in sensors and drowning in data.

- Lt Gen David Deptula, former Air Force Deputy Chief of Staff for Intelligence, Surveillance and Reconnaissance.[25]

The Air Force and DoD, as a whole, face substantial acquisition and organizational challenges in crafting an agile and evolutionary response to the data explosion. Future sensors and sensor data will greatly exceed the capability of information operators to process much less act quickly using current data-computation and integration methods. Col Weinburg, ISR Task Force chief of operations, stated, "…military analysts often spend 75 percent of their time poring through intel data, and only 25 percent analyzing it."[26]

Inexpensive sensors aren't just an intelligence issue, these sensors will be available in multiple disciplines and the need for flexible implementation will only continue to grow as new users create new ways to work with the data. As relatively inexpensive and capable collections systems proliferate, more efficient and elegant solutions will be required to capture, analyze, share and visualize this data. As much as this is an area of intense interest to the intelligence analyst of the future, it is also an opportunity for Air Force operators, logisticians, and medical personnel of the future.

The heterogeneous data explosion is here and will only gain in force. As an example, the already fielded Global Hawks Block 10 will begin to be replaced in April 2011 with the more capable and sensor-rich Global Hawk Block 30 (Figure 4). These new Global Hawk platforms will increase in number and are slated to replace the workhorse U-2 platform; this is known as the High-Altitude Transition (HAT) plan (Figure 2 and 3).

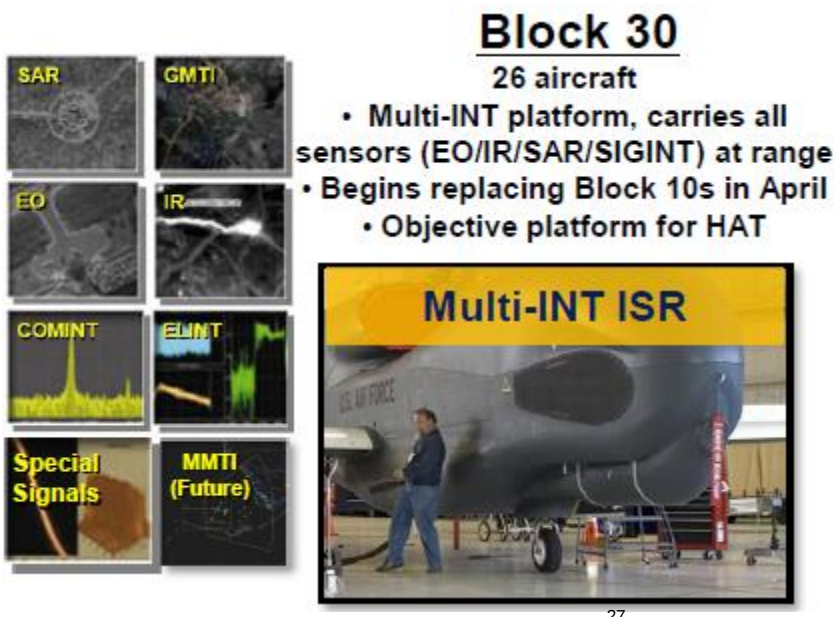

Figure 2: Global Hawk Block 30 capabilities.[27]

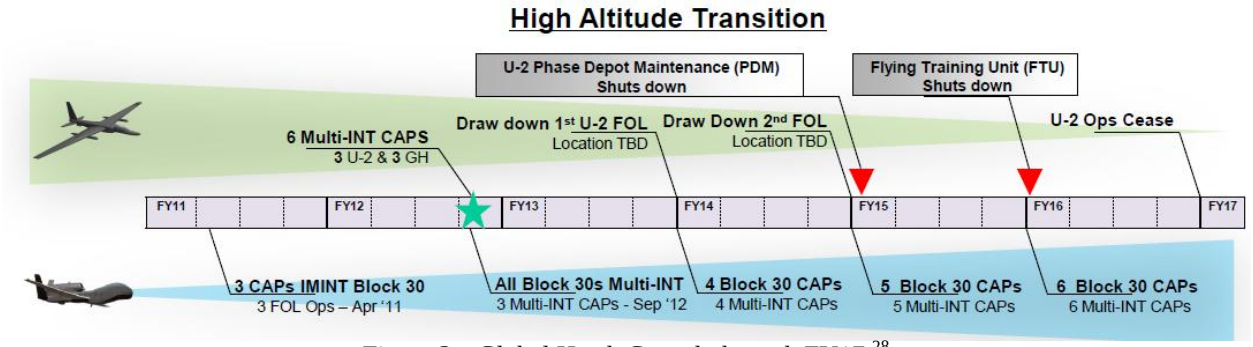

Figure 3: Global Hawk Growth through FY17.[28]

These figures represent the growth in only one platform, but similar expansion in capabilities is projected for Army and Navy systems that will frequently be operating in the same area. As an example, the Army's Long Endurance Multi-intelligence Vehicle (LEMV) is a 250-foot hybrid airship that is planned to stay aloft for three weeks, travel at speeds of 30-80 knots, and carry 2,500 pounds of sensors and data links.[29] The Army expects to have LEMVs over Afghanistan by 2011, only 18 months after the Feb 2010 solicitation.[30] The shared need for improved data integration provides a unique opportunity for the services to work together.

The rapid acquisitions of the Army's LEMV and Air Force's Project Liberty provide a template for quick fielding of vital capabilities. Project Liberty's thirty-seven platforms MC-12W were acquired, specially equipped, and fielded over just two years to provide tactical intelligence collection for operations in Iraq and Afghanistan.[31] In both cases, many of the cumbersome federal acquisition processes were waived.[32] While such a path has challenges for enterprise-wide solutions, it does offer the ability to do rapid development and evolve a promising capability. This paper posits that such an acquisition strategy would enable quick testing and further evolution of promising data integration and analysis solutions.

Joint solutions are necessary to enable service mission success. Data integration critics argue not all of the tactical and operational data will be of long-term or strategic use; this may well be true for much of the data. However, in a fiscally constrained environment, it is irresponsible to ignore planning for integration of the most valuable data in a more elegant and powerful manner than e-mailing or posting briefings for retrieval across organizational lines. The DoD ISR Task Force has stated they are focusing on data solutions over procuring more ISR platforms.[33] Such a joint effort should be supported fully by the Air Force; this engagement will enable us to leverage joint resources toward mutually beneficial solutions.

Lest this challenge be seen as a specialized intelligence issue, the reader only needs to look at the description of many new systems in the acquisition pipeline to see that a flexible system that is responsive to diverse users is vital to operations. Future-combat platforms will combine the ISR role with the bomber or fighter role by design. This multi-role mission is already a reality with Hellfire equipped Predators and Reapers. Current command and control software systems aren't designed to fully enable the flexibility of multi-role aircraft. This is an issue from air tasking order production through data dissemination to varied users. Future data

and system integration will be no different unless the Air Force begins now to build in operational flexibility. This flexibility is needed to allow use of new and old systems and data without being forever stuck in the stovepipes of the past. Bottomline: the Air Force can spend billions on meta materials, hypersonics, and nanotechnology, but the use of these advanced capabilities will be hamstrung at best and useless at the worst without data capabilities that enable decision making within the Observe Orient Decide and Attack (OODA) Loop. As that cycle gets faster, takes less time and becomes more automated and decentralized, it will move toward an OODA Point, thus requiring almost instantaneous data fusion.

[25] Lt Gen David Deptula, as quoted by Stew Magnuson in "Military 'Swimming in Sensors and Drowning in Data,'" *National Defense Magazine*, Jan 2010, http://www.nationaldefensemagazine.org/archive/2010/January/Pages/Military'SwimmingInSensorsandDrowninginData'.aspx.
[26] Col Weinberg, ISR Task Force, as quoted by John Bennett in "Gates' ISR Task Force To Join Top DoD Intel Office," *Defense News*, 7 Oct 2010, http://www.defensenews.com/story.php?i=4863676&c=POL&s=TOP, accessed 7 Oct 2010.
[27] Col Gear, AF/A2CU, "Global Hawk Update and AF RPA Way Ahead Briefing," 12 Oct 2010.
[28] Ibid.
[29] "Rise of the "Blimps": The US Army's LEMV", *Defense Industry Daily*, 2 Sep 2010, http://www.defenseindustrydaily.com/Rise-of-the-Blimps-The-US-Armys-LEMV-06438/.
[30] "Long Endurance Multi-Intelligence Vehicle Solicitation," Department of the Army, 11 Feb 2010, https://www.fbo.gov/index?s=opportunity&mode=form&id=63af8191a894981adf3879047c9800ba&tab=core&_cview=1, accessed 4 Dec 2010. The LEMV project was exempt from many US federal acquisition regulations and processes, allowed non-traditional partners, and included a rapid solicitation to operational testing timeline.
[31] Caitlin Harrington, "USAF confirms Project Liberty plans to deploy ISR aircraft," *Jane's,* 28 Jan 2009, http://www.janes.com/news/defence/air/jdw/jdw090128_1_n.shtml.
[32] Bennett, "Gates' ISR Task Force," and "Long Endurance Multi-Intelligence Vehicle Solicitation."
[33] Bennett, "Gates' ISR Task Force."

ENVISIONING DATA INTEGRATION IN 2035

> *The intellectual power and prowess of our human resources go undeveloped and are eclipsed by those of other nations while the Intelligence Community strains mightily under tectonic forces of shifting technologies, powerful organizations, and rapidly accreting mountains of data.*
>
> --Institute for Modern Intelligence[34]

Looking to the future, computing power and data storage capacity are expected to grow exponentially. Still, the data to fuel this brave new world needs to be accessible and integratable to take full advantage of the technology revolution. The fastest computer in the world today performs at 2.57 petaflops (or 2.57 thousand million million calculations) a second.[35] By 2029, $1000 computers can be expected to do "twenty million billion calculations a second, equivalent to what a thousand brains can do."[36] In the 2030's a disk-sized device could store "a trillion trillion bits of information" or even more.[37] As incredible as this all is, faster computers and bigger storage still need the data as fuel; data that is locked away with no plan for integration has little benefit for knowledge generation.

Cloud computing offers much promise for enabling data integration and is already available to the average person via smartphones and online applications. The government is inhibited from full data integration, cloud based or otherwise, due in large part to process and governance challenges. Recognizing that process and government are necessary, computer science skills "are the best hope for models that (incrementally) clarify, improve and support incremental automation of process and governance."[38]

Need for an evolutionary solution, vice a final solution, is echoed in the increasing interest in data virtualization. Data virtualization provides a flexible layer of abstraction that insulates "DI (data integration) targets from changes in DI sources as new data sources are retired or added."[39] This decoupled data also offers the ability to reconfigure business processes

and data exchanges to reflect changing needs without modifying the underlying data stores. It also allows users to reuse objects and services from data silos in a multitude of changing "consumer channels and applications."[40]

The earlier reviewed World-Wide Telescope solution points the way to the power of getting data to where it can be shared, debated, and used. Elements of the government and military have taken notice of this potential and are pursuing an integration approach called Ultra-Large-Scale (ULS) Systems. The ULS System concept has gained the attention of national-level agencies and is built on the concepts of Dr. Jim Gray's Fourth Paradigm. [41]

ULS Systems "will be interdependent webs of software-intensive systems, people, policies and economics." They are designed to operate at large scale, be decentralized, be developed and operated by various entities with different or even conflicting needs, and be built to evolve. "People will not just be users of a ULS system; they will be elements of the system. Software and hardware failures will be the norm….The acquisition of a ULS System will be simultaneous with its operation and require new methods for its control."[42] In summary, ULS Systems, whether known by this name or another, are the operating environment of the future.

The Army drove the ULS study because their leaders understand there is a fundamental system challenge to overcome if they are "to see first, act first, and act decisively."[43] While this challenge is shared by all of the services, the Army has already made ULS a key focus area for the Distributed Common Ground System-Army (DCGS-A) of the future.[44] DCGS-A uses a database aptly called the "Brain" which is becoming the backbone of intelligence databases in many theaters. The value of this solution is not limited to DCGS-A; it is useful for analysts and operators across services and agencies.

To enable and take advantage of the ULS System future, where very little of the evolving environment will be under one organizations control, data fusion must:

- Present minimal barriers to incorporating new data and semantics,
- Embrace all data "sources, types, models, and modalities,"
- Support diverse processing by which "structural and semantic barriers are overcome to yield information and knowledge,"
- Allow reuse of data, information, and knowledge from diverse perspectives,[45]

To achieve this operational data integration flexibility, data models must be considered from a higher level of abstraction.[46] The growth in data virtualization, discussed earlier, offers a window into the need to abstract data from its original data model and data storage containers.

Successful data-integration solutions fit the business processes of users. Intelligence business processes "include data collection, semantic enhancement, fusion from data to information to knowledge, and communication/collaboration to create understanding."[47] Figure 4 demonstrates cognitive hierarchy on the right. On the left, a simplified version of a data integration framework identifies the key layers necessary to enhance data into understanding. This specific "Data Architecture and Semantic Integration Framework" mirrors both the structure of cognition and the operations of intelligence business processes.[48]

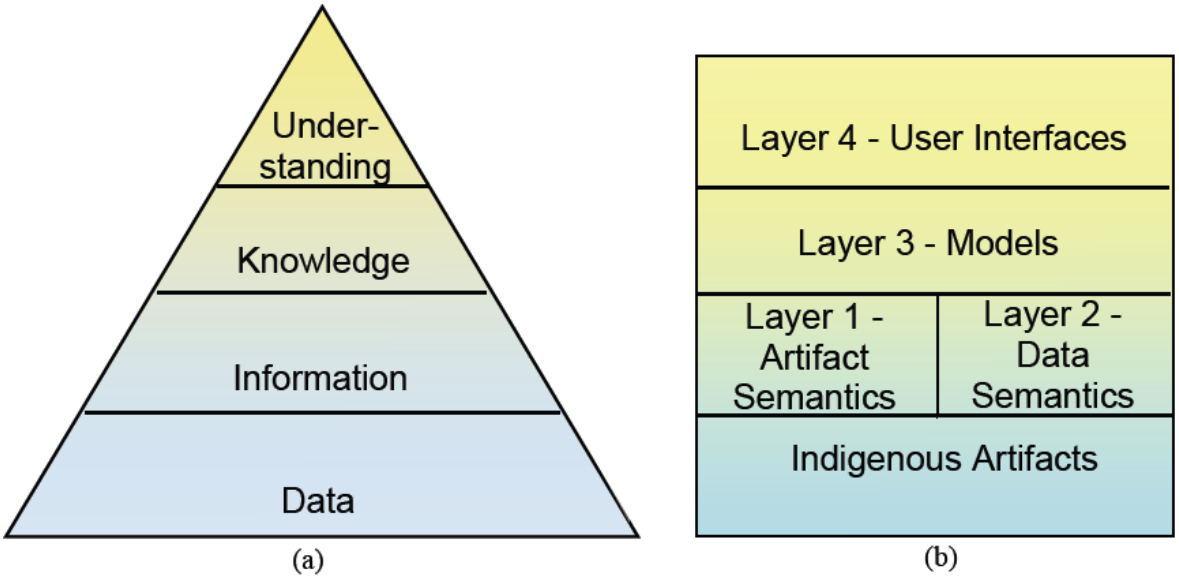

Figure 4: (a) The cognitive hierarchy. Intelligence business processes move intelligence artifacts upward through the hierarchy. (b) Organization of the Data Architecture and Semantic Integration Framework in support of the cognitive process.[49]

Of critical note, the first and second layers demonstrate the process of abstracting the data from its original source into a "Unified Data Space." Such a space goes well beyond data integration to enable data to exist unmodified by the shape of the data storage container while retaining its key identifying information (the data about the data or the Metadata). In this construct data is not just integrated it is unified. Figure 5 provides a view of how Layers 1-3 of a unified data space work.

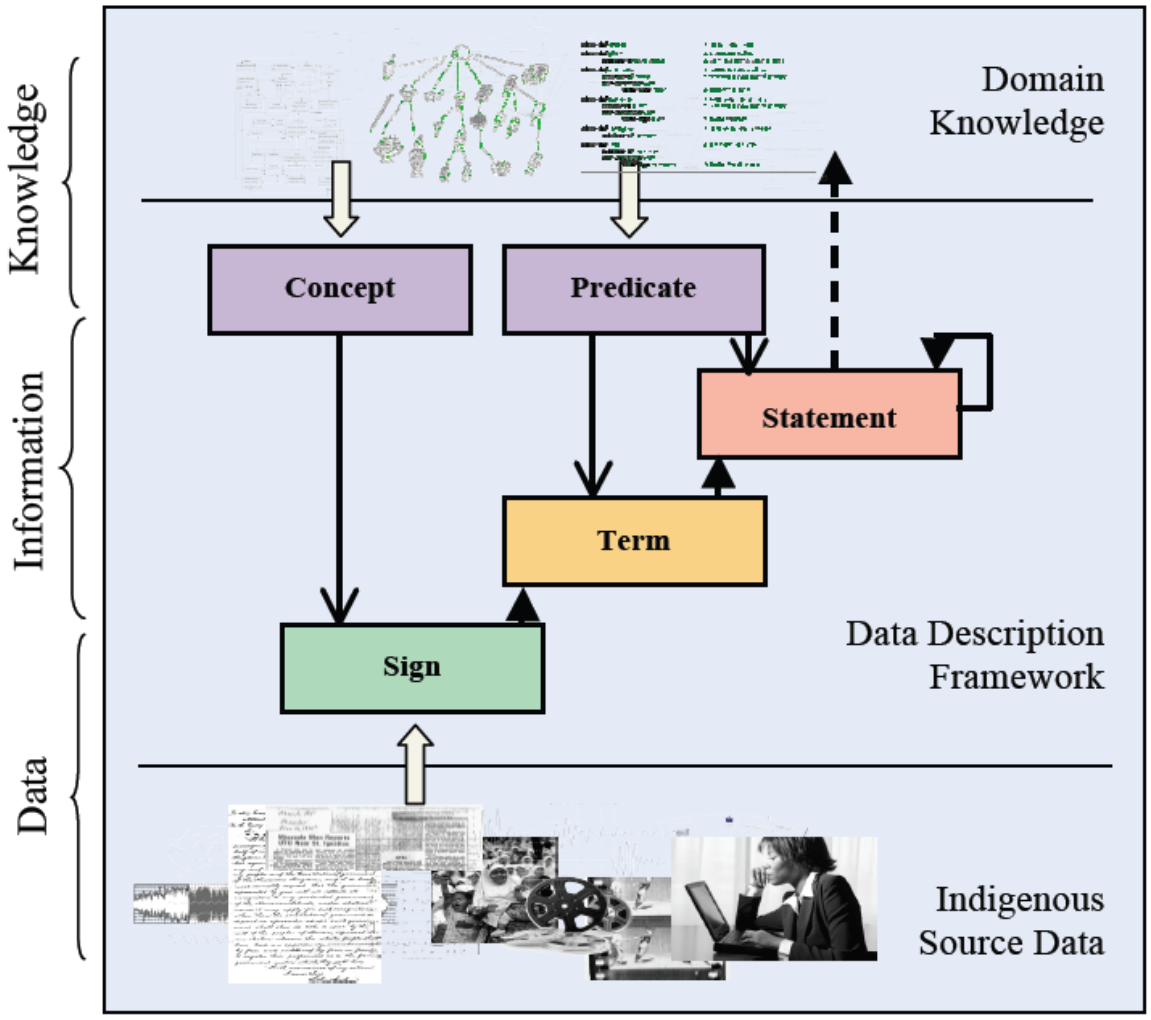

Figure 5: High-level diagram illustrating the first three layers of the Data Architecture and Semantic Integration Framework. Broad arrow: "feeds," open arrow: "binds with," closed arrow: "forms," dashed arrow:" informs."[50]

This solution preserves the sources' original data and semantics, uses diverse data of any type, can modify sources readily for evolutionary flexibility, and supports powerful processing "without limitations." Current solutions require intense "pre-integration processing (schema harmonization and data normalization) and usually entail loss/distortion of original data and semantics."[51] This heavy processing limits data fusion due to forcing the data back into a new data schema.

Since most of us aren't database managers, a picture of the type of visualization possible with unified and enhanced data is worth a thousand additional words (Figure 6).

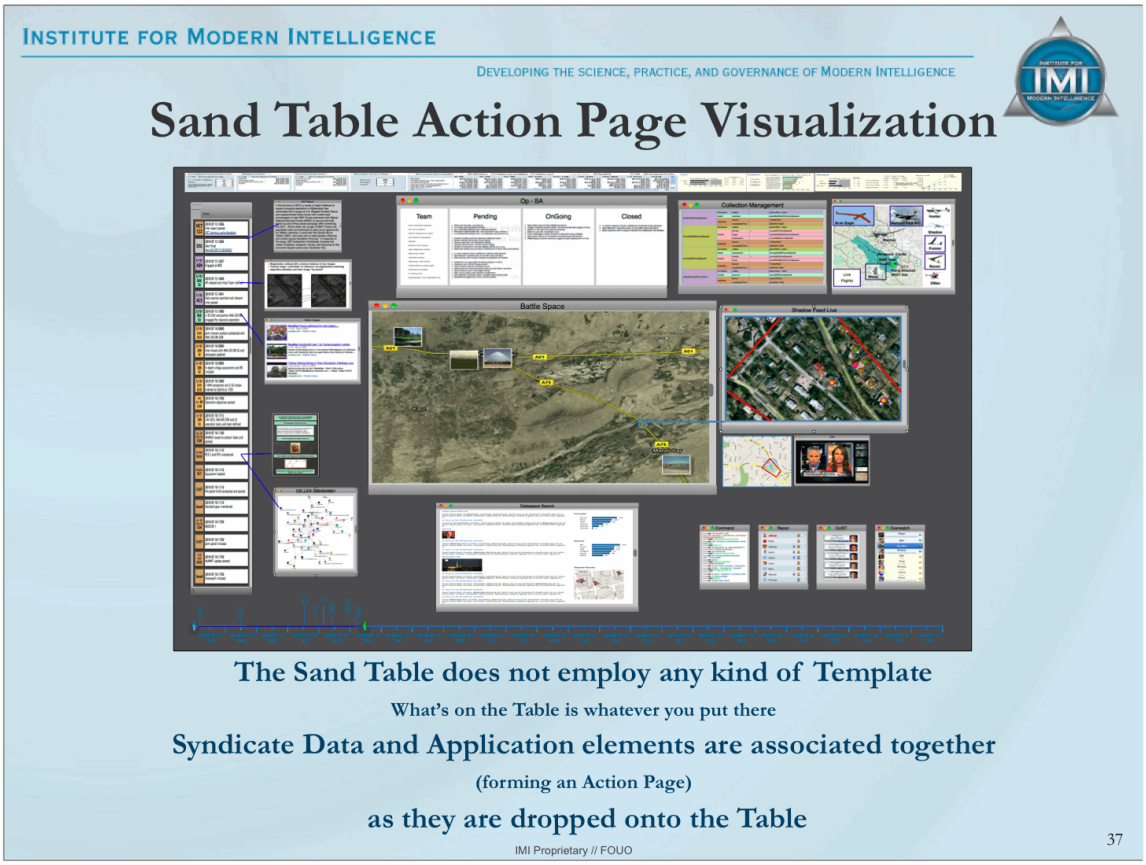

Figure 6: Depiction of a flexible visualization tool that could overlay unified data.[52]

Finally, reaching this future sooner requires incentives for sharing, semantics solutions, and access permission changes.[53] Incentives would encourage sharing and could be provided by including tools that give something back to the data providers. Semantics would focus more toward enabling data comparison versus providing an absolute description. Finally, flexible data access processes and tools will be needed to provide access based on specific missions and roles.

[34] Institute for Modern Intelligence executive summary, http://www.imintel.org, accessed 4 Oct 2010.

[35] "China Grabs Supercomputing Leadership Spot in Latest Ranking of World's Top 500 Supercomputers," 11 Nov 2010, http: www.top500.org/lists/2010/11/press-release, assessed 4 Jan 2011. The fastest supercomputer in the world is China's Tianhe-1A. It supplanted the US Blue Gene supercomputers as the leader with the release of the November 2010 Top 500 list. The list is created by a team of computer science professors from the University of Mannheim Germany; Lawrence Berkeley National Laboratory; and the University of Tennessee, Knoxville.
[36] Singer, *Wired for War*, 102.
[37] Ibid., 102.
[38] Arnon Rosenthal, 6 Dec 2010.
[39] Stephen Swoyer, "Why Data Virtualization Trumps Data Federation Alone," The Data Warehousing Institute (TDWI), 1 Dec 2010, 1, http://tdwi.org/articles/list/news-articles.aspx, accessed 4 Dec 2010.
[40] Swoyer, "Why Data Virtualization?" 1.
[41] *Ultra-Large-Scale Systems: The Software Challenge of the Future*. Study lead Linda Northrup. Pittsburgh, PA: Carnegie Mellon Software Engineering Institute, June 2006, ix-3, http://www.sei.cmu.edu//library/abstracts/books/0978695607.cfm.
[42] "Ultra-Large-Scale Systems Overview," Software Engineering Institute, Carnegie Mellon. http://www.sei.cmu.edu/uls.
[43] "Ultra-Large-Scale Systems Overview."
[44] Suzanne Yoakum-Stover, "Trends in Infrastructure: Commercial vs Military." (lecture. National Association of Broadcasters Military & Government Summit, Las Vegas, NV, 13 April 2010).
[45] All characteristics of the ideal data integration future are from Norbert Antunes, Tatiana Malyuta, and Suzanne Yoakum Stover, "A Data Integration Framework with Full Spectrum Fusion Capabilities," August 2009, 2-3.
[46] Ibid, 3.
[47] Yoakum-Stover, "DDF 2009," 2.
[48] Ibid, 2.
[49] Ibid, 2
[50] Ibid, 10.
[51] Suzanne Yoakum-Stover, Tatiana Malyuta, and Norbert Antunes, "A Data Integration Framework with Full Spectrum Fusion Capabilities," *Challenge 2009* (MSS Information Fusion Symposium, Las Vegas, NV August 2009).
[52] Suzanne Yoakum-Stover and M. Andrew Eick, "Data & Processing Syndicate," (Briefing, 15 September 2010). Obtained during 10 November 10 interview with Dr Yoakum-Stover.
[53] Arnon Rosenthal, 6 December 2010.

RECOMMENDATIONS

Make everything as simple as possible, no simpler.
-- Einstein

Treat data integration as a strategic competency. The Air Force is investing a great deal in platforms and manpower that will be ill served by a tactical, reactive data integration approach. Those who "focus only on implementing data integration architectures as cheaply as possible and optimized for narrow needs will continue to fall farther behind."[54] This change in focus will require a corresponding investment in experience and a conceptual shift to including data integration into project and process planning at the beginning, not after delivery to the field.

Energize funding and research on data-integration solution development with a DoD focus. Such an effort requires a multi-disciplinary approach as moving toward integrating heterogeneous data relies on software and system engineers, policy developers, security experts, mission area experts, cognitive psychologists, and human factors engineers. [55] The goal is identification of areas that are ready for fielding or worthy of investment for development. Gartner's analysis of available commercial data integration offerings is a good starting point for identifying leading concepts associated with companies that have the ability to execute solutions on an enterprise scale such as Informatica, IBM, and Microsoft.[56] The USAF should leverage and support IARPA's ongoing effort to identify promising data integration solutions.[57] Work with the ISR Task Force, which this paper predicts will lead the overall DoD ISR data integration effort, is essential.

Shift the perspective from rational, top-down engineering to enabling and regulating a complex, decentralized system. ULS systems and the data that drives them are by their very nature evolutionary; portions of the system or data sources will come and go; segments

will need to be taken down for repairs while the whole remains operational. Users will have varied purposes from pure analysis to operational activities; these efforts will involve rapidly forming teams with unique demands.[58] Distributed participation and solutions are a necessary part of the process.

Actively drive integration of computer engineering, scientific research, and the policy/process team. The principle take-away of Dr Gray's Fourth Paradigm and ongoing efforts to implement his vision is that scientists and system and software engineers need to work together to overcome the data challenges of the present and future.[59] This focused integration is already yielding results in astronomy, ecology, oceanography, neuroscience, healthcare delivery and holds promise to deliver much more within even the next five to ten years. This integrated approach was echoed by the Mitre team in recommending computer science engineers work in tandem with the policy and process communities to move forward on data integration while acknowledging the real needs of the user community for security and reliability.

Synergize service and agency efforts. Budget limitations are a fact of life. Working aggressively with the other services and agencies to craft a data-integration path is vital. The Army is already delivering first-phase solutions and the Air Force would benefit from putting energy into synergizing data integration efforts. The services can put the shared data to good use using the skill sets that each of the components brings to the fight. The heart of the data challenge is that different users have different needs from the same data. The now six-year joint effort to integrate service DCGS elements is a start, but there is still much work to be done on even basic system-level integration of the services' data sources.

Invest in a Unified Data Space testbed. The testbed does not have to include all available intelligence community data. Instead, this paper recommends an optimal starting

location due to shared mission is the Army and Air Force DCGS data. This data is being created daily and already includes much of the newest sensor data. If this cannot be accomplished due to service issues, partnering with the National Geospatial Intelligence Agency (NGA) Innovision team, which is pursuing enriched data and investing in data integration, would be beneficial. A unified data space is specifically designed to evolve to include additional data sources, so starting with a focused area will not prevent future expansion. Criteria for success will be the ability to enrich already existing data.

[54] Friedman, Beyer, and Thoo, "Magic Quadrant," 3-4.
[55] *Ultra-Large-Scale Systems*, xi.
[56] Friedman, Beyer, and Thoo, "Magic Quadrant."
[57] The Intelligence Advanced Research Projects Activity (IARPA) website, http://www.iarpa.gov/solicitations_kdd.html (accessed 6 Dec 2010). IARPA "invests in high-risk/high-payoff research programs that have the potential to provide our nation with an overwhelming intelligence advantage over future adversaries." IARPA's Knowledge Discovery and Dissemination (KDD) Program specifically focuses on data integration.
[58] *Ultra-Large-Scale Systems*, 6
[59] Hey, Tansley, and Tolle, eds, *The Fourth Paradigm: Data Intensive Scientific Discovery*, location 4507-4520.

SUMMARY AND CONCLUSIONS

Perhaps the battlefield of the future will benefit from plug and play satellites that are launched within days and capable of autonomous and deconflicted cross-cueing with each other and sensors on the ground, in the air, and sea. Whether or not we reach the elusive goal of autonomous, real-time interrogation and surveillance, we are assured of being in a data-rich environment. The challenge is how to make the data useable by many not just locked in valuable but separate stockpiles available to a few. While there are very real institutional, structural, process, and security challenges inhibiting data integration, these challenges are not insurmountable and are not a reason to accept the status quo. The ultimate goal is to convert the data challenge into an incredible intelligence, decision-making, action-optimizing resource.

The business processes and policies that shape intelligence community interaction are part of the operational environment. Engagement directly between computer science, data management, policy and operational community offers the best hope for real movement forward. Such an effort would be greatly aided by a unified data space that takes advantage of data attributes to enable data fusion while freeing the data from model and storage constraints. The unified data space would become a tool for the entire community providing access to data for further testing and development of analysis and operations tools and visualization capabilities. Such a solution is designed to be evolutionary and flexible from the outset and takes advantage of growing commercial interest and capabilities in data integration tools.

If we don't attack the data integration challenge we will continue with the inefficient and data limited structures and processes of the day. At best, these current solutions slow analysis. At worse, they are key contributors to "intelligence failures," unnecessary loss of life, and poor decisions. Those that think fusion centers with access to many discreet data stockpiles are the

solution need only look at the Department of Homeland Security's stymied $426 million Homeland Security Information Network (HSIN). HSIN links 72 state and local "fusion centers" but does not offer the ability to search across multiple databases and or systems. As a result, the DHS Inspector General found that analysts logged on to HSIN fewer than five minutes a month and preferred to rely on e-mail for exchanging data. Of note, the DHS IG recommended "single sign-on and comprehensive search capabilities."[60] This echoes the findings of this author; leaving data in discreet data stockpiles that rely heavily on separate searches is too costly in time and hampers the overall quality of analysis.

In a fiscally constrained environment, it is irresponsible to ignore planning for integration of the most valuable data in a manner more elegant and powerful than e-mailing or posting briefings for happenstance retrieval across organizational lines. Analysts need much more powerful data discovery and integration capabilities to make sense of the data deluge. Decisions are only as good as the information and knowledge that underpin them. Too often, we are undermining our operations and policy decisions by "flying blind" when we could be seeing deeper with data that already exists.

---

[60] Alice Lipowicz, "Fusion Centers Hampered by Limitations of DHS nets, IG says," *Federal Computer Week*, 16 November 2010, http://fcw.com/article/2010/11/16/dhs-fusion-centers.aspx.

Glossary

Cloud computing—Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.[61]

Data Virtualization—The process of aggregating data from a variety of information sources so it can be accessed without regard to original physical storage or data structure. This data can then be used by front end solutions such as applications or portals. Early data virtualization is often referred to as data federation or "data mashups."

Massive Parallel Processing—A computer system with many independent processing units that run in parallel.

Metadata—The data about the data.[62] Metadata provides information about a certain item's content, such as the file size of a picture, date, processing information, author and even key search tags. These tags are used to enable file identification and retrieval.

Ontology—In computer science, a exhaustive organization of some knowledge domain that is frequently "hierarchical and contains all relevant entities and their relations."[63] Ontology mapping links the entities and a universal ontology seeks to identify all possible entities of interest across knowledge domains.

Semantics—The meaning/meanings of a word, element, or text.[64]

---

[61] Mell and Grace, "The NIST Definition of Cloud Computing."
[62] *Princeton's WorldNet*: *a lexical database for English*. http://worldnetweb.princeton.edu/, accessed 18 January 2011.
[63] Ibid.
[64] Ibid.

Bibliography

Bacon, Sir Francis. *Religious Meditations, Of Heresies.* 1597. http://www.quotationspage.com/quote/2060.html (accessed 4 December 2010).

Bennett, John. "Gates' ISR Task Force to Join Top DoD Intel Office." *Defense News*, 7 October 2010. http://www.defensenews.com/story.php?i=4863676&c=POL&s=TOP (accessed 7 October 2010).

"China Grabs Supercomputing Leadership Spot in Latest Ranking of World's Top 500 Supercomputers." *Top 500 Supercomputer Sites*, 11 November 2010. http: www.top500.org/lists/2010/11/press-release (assessed 4 January 2011).

Cloud Computing Use Case Discussion Group. "Cloud Computing Use Cases," white paper, 31 July 2009. http://groups.google.com/group/cloud-computing-use-cases (accessed 20 November 2010).

"COMPASE Center, Layered Sensing Operations" flyer from Air Force Research Labs, Wright Patterson Air Force Base, OH.

Eick, M. Andrew and Suzanne Yoakum-Stover. "Fixing Intel and Operationalizing Data – The Data & Processing Syndicate." www.imintel.org (accessed 6 October 2010).

Friedman, Ted, Mark Beyer, and Eric Thoo. "Magic Quadrant for Data Integration Tools." Gartner, 19 November 2010. http://www.gartner.com/technology/media-products/reprints/sas/vol7/article4/article4.html, (accessed 28 November 2010).

Friedman, Thomas. *The Lexus and the Olive Tree*. New York: Farrar, Straus, and Giroux LLC, 1999.

Gear, Col, AF/A2CU. "Global Hawk Update and AF RPA Way Ahead Briefing," 12 October 2010.

Garreau, Joel. *Radical Evolution: The Promise and Peril of Enhancing Our Minds, Our Bodies - and What it Means to be Human*. New York, NY: Broadway Books, 2005.

Goure, Daniel. "Wikileaks Dilemna: How Does a Nation Fight a Superempowered Person?" Lexington Institute Early Warning Blog, 6 December 2010. http://www.lexingtoninstitute.org/, (accessed 8 December 2010).

Harrington, Caitlin. "USAF confirms Project Liberty plans to deploy ISR aircraft." *Jane's,* 28 January 2009. http://www.janes.com/news/defence/air/jdw/jdw090128_1_n.shtml (accessed 15 November 2010).

Hey, Tony; Stewart Tansley, and Kristin Tolle, eds. *The Fourth Paradigm: Data Intensive Scientific Discovery*. Microsoft Research: 2009. E-book. http://creativecommons.org/licenses/by-sa/3.0 (accessed 23 September 2010).

Institute for Modern Intelligence fact sheet. http://www.imintel.org (accessed 4 October 2010).

The Intelligence Advanced Research Projects Activity (IARPA) website, http://www.iarpa.gov/solicitations_kdd.html (accessed 6 December 2010).

King, Caroline. "Cooperative Control Overview." Lecture. Air War College Blue Horizons Team, Air Force Research Labs, Wright Patterson Air Force Base, OH, 22 September 2010.

Kramer, Franklin D.; Stuart H. Starr, and Larry K. Wentz, eds. *Cyberpower and National Security.* Washington, D.C.: National Defense University Press, 2009.

Kurzweil, Ray. *The Singularity is Near: When Humans Transcend Biology*. New York, NY: Viking Publishing, September 2005.

Lipowicz, Alice. "Fusion Centers Hampered by Limitations of DHS nets, IG says," *Federal Computer Week*, 16 November 2010. http://fcw.com/article/2010/11/16/dhs-fusion-centers.aspx (accessed 5 January 2011).

"Long Endurance Multi-Intelligence Vehicle Solicitation." Department of the Army, 11 February 2010. https://www.fbo.gov/index?s=opportunity&mode=form&id=63af8191a894981adf3879047c9800ba&tab=core&_cview=1 (accessed 4 December 2010).

Markoff, John. "A Deluge of Data Shapes a New Era in Computing." *New York Times* (December 15, 2009): D2.

Mell, Peter and Tim Grace. "The NIST Definition of Cloud Computing," Version 15, 10-7-09. http://csrc.nist.gov/groups/SNS/cloud-computing/ (accessed 1 December 2010).

Perry, Beth. "Emerging Threats in CB Weapons Space." Lecture. Air War College Blue Horizons team, Los Alamos National Laboratory, NM, 25 August 2010.

Princeton's WorldNet: *a lexical database for English*. http://worldnetweb.princeton.edu/ (accessed 18 January 2011).

Orbst, Leo. Mitre Corporation, e-mail interview, 6 December 2010.

"Rise of the "Blimps": The US Army's LEMV." *Defense Industry Daily*, 2 September 2010. http://www.defenseindustrydaily.com/Rise-of-the-Blimps-The-US-Armys-LEMV-06438/ (accessed 2 December 2010).

Rosenthal, Arnon. MITRE Corporation. e-mail interview, 6 December 10.

Schwartz, Peter. *Inevitable Surprises*. New York, NY: Gotham Books, 2004.

Seligman, Len. Mitre Corporation, e-mail interview, 6 December 2010.

Singer, P.W. *Wired for War: The Robotics Revolution and Conflict in the 21st Century.* New York, NY: The Penguin Group, 2009.

Swoyer, Stephen. "Crunching the Numbers on Big Data," The Data Warehousing Institute (TDWI), 1 December 2010, 1, http://tdwi.org/articles/2010/12/01/crunching-big-data-numbers.aspx?admgarea=news, accessed 4 December 2010.

Swoyer, Stephen. "Why Data Virtualization Trumps Data Federation Alone," TDWI, 1 December 2010. http://tdwi.org/articles/list/news-articles.aspx (accessed 4 December 2010).

Ultra-Large-Scale Systems: The Software Challenge of the Future. Study lead Linda Northrup. Pittsburgh, PA: Carnegie Mellon Software Engineering Institute, June 2006. http://www.sei.cmu.edu//library/abstracts/books/0978695607.cfm (accessed 23 September 2010).

"Ultra-Large-Scale Systems Overview." Software Engineering Institute, Carnegie Mellon. http://www.sei.cmu.edu/uls (accessed 9 September 2010).

Vestrand, W. Thomas. "Thinking Telescopes." Lecture. Air War College Blue Horizons team at Los Alamos National Laboratory, NM, 25 August 2010.

World-Wide Telescope. http://www.worldwidetelescope.org (accessed 26 September 2010).

Yoakum-Stover, Suzanne, Tatiana Malyuta and Norbert Antunes, "A Data Integration Framework with Full Spectrum Fusion Capabilities," *Challenge*. MSS Information Fusion Symposium, Las Vegas, NV. August 2009.

Yoakum-Stover, Suzanne and M. Andrew Eick. "Data & Processing Syndicate," Briefing, 15 Sep 2010. Obtained during 10 Nov 10 Interview with Dr Yoakum-Stover.

Yoakum-Stover, Suzanne. "Data-Descriptive Framework 2009," White Paper. Obtained from author during 10 November 2010 interview.

Yoakum-Stover, Suzanne. Telephone interview, 10 November 2010.

Yoakum-Stover, Suzanne. "Trends in Infrastructure: Commercial vs Military." Lecture. National Association of Broadcasters Military & Government Summit, Las Vegas, NV, 13 April 2010.