Core Science Analytics, Synthesis, and Libraries Five Year Science Strategy

07-07-2015

Introduction

Core Science Analytics, Synthesis, and Libraries (CSAS&L), a program within the Core Science Systems (CSS) Mission Area, offers a comprehensive spectrum of skills, products, and services that enhance scientific research capabilities within the United States Geological Survey (USGS).

CSAS&L Vision:

Synthesize and deliver Earth system data and information.

CSAS&L Mission:

Accelerate research and decisionmaking through data science, information delivery, advanced computing and biodiversity analytics.

This Five Year Science Strategy outlines CSAS&L goals, research questions, and strategic actions for four key science themes - Biodiversity Science, Applied Research Computing, Library and Information Science, and Human Dimensions - and one holistic, comprehensive theme - Data Science - that arises from the confluence and fusion of these fully realized science themes. The scope of these activities within CSAS&L and their relationship to the broader vision of CSAS&L, are depicted below in Figure 1.

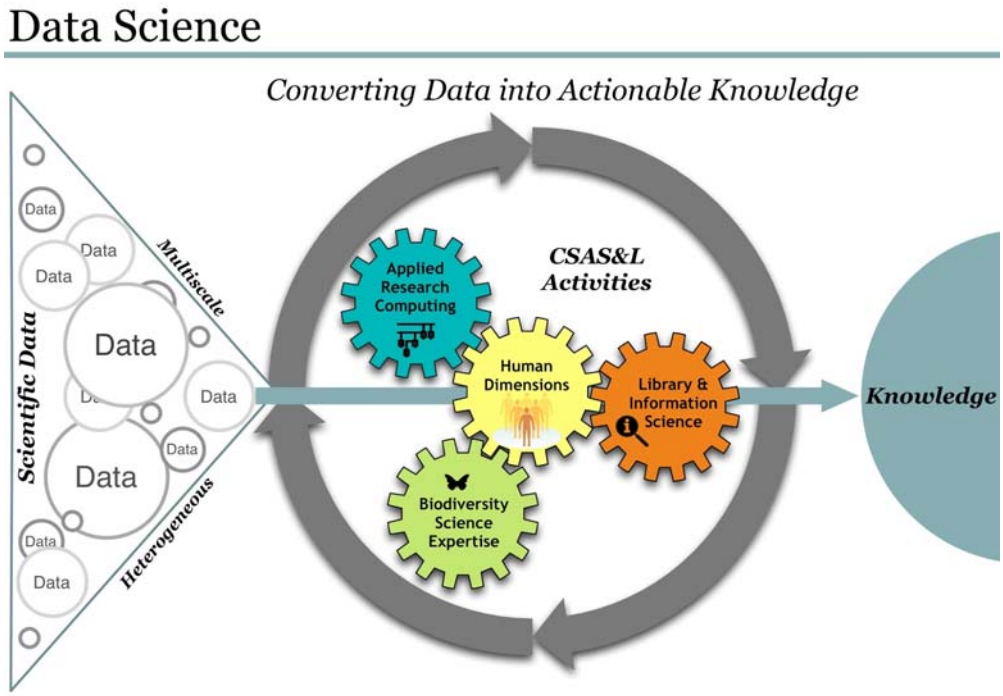


Figure 1: CSAS&L Data Science - Converting data into actionable knowledge.

CSAS&L develops, produces, and maintains a broad suite of terrestrial and aquatic data systems, established under the former USGS Biological Informatics Program, and used for biodiversity assessment and decision-making. With the incorporation of the enterprise scientific data mission from the former Geospatial Information Office and the information science and services mission of the USGS Library System, CSAS&L is well positioned with the diversity of skills and expertise needed to provide Data Science infrastructure, leadership, and expertise within the USGS.

Biodiversity Science, Applied Research Computing, and Library and Information Science are core focus areas for CSAS&L. While each of these research areas provides a suite of targeted products and disciplinary expertise, each also draws upon connections to the others. Biodiversity Science relies upon foundational Library and Information Science principles of data organization and description to make those data continuously actionable. Additionally, Biodiversity Science depends upon Applied Research Computing to provide substantial computation power for processing large volumes of data to facilitate timely understanding and decision-making. Applied Research Computing, in turn, relies upon Library and Information Science capabilities to ensure that large, heterogeneous datasets used are accessible, organized, well-documented, and persistent. Applied Research Computing's true value cannot be realized without the Biodiversity Science expertise required to glean insights and understanding from large, processed datasets. Relationships among these CSAS&L core capabilities are thus multi-dimensional.

Human Dimensions is the nexus of our core activities, as users are the focus for everything we do. Identifying our stakeholders, understanding their priorities, and addressing and anticipating their

needs drive CSAS&L development of tools, capabilities, and resources in Biodiversity Science, Applied Research Computing, and Library and Information Science areas. Human Dimensions provides the 'feedback loop' that helps us define, refine, and establish priority products and new opportunities to serve communities of practice, and is a growing area of action and research potential for CSAS&L. As depicted in Figure 1, Human Dimensions is both the cog that drives our research activities, and the bridge that connects them.

Research data are the raw materials USGS converts into the actionable knowledge used by our stakeholders to address societal needs. While each focal area in CSAS&L applies its expertise to discrete activities such as species occurrence data systems, best practices for research data management, or providing USGS scientists with access to high performance computing capabilities, our four research areas also work synergistically to provide true data science capacity for USGS. This broader application of our capabilities and expertise mirrors the emergence of 'big data,' as organizations transition from relatively homogeneous data created and managed for a narrowly focused need, to an era where data are complex, heterogeneous, multi-disciplinary, disparately structured, and applicable to a wide range of uses, both known and unanticipated. It is this 'big data' universe for which CSAS&L's experience and resources are ideally suited.

Biodiversity Science Expertise

CSAS&L's biodiversity activities focus on a landscape-level understanding of terrestrial, aquatic and marine species and ecosystems, through development, maintenance, and integration of a suite of synthesized data systems used to conduct biodiversity assessments at regional to national scales.

These data come from hundreds of source datasets, contain millions of discrete observations and measurements, and are built using robust research-based algorithms. These systems include species taxonomy (Integrated Taxonomic Information System); species observations across terrestrial, wetland, estuarine, and marine ecosystems (U.S. Ocean Biogeographic Information System and the Biodiversity Information Serving Our Nation system); detailed species distribution models and range information (terrestrial Gap Analysis Program and Aquatic Gap Analysis Program); ecological system classification and mapping to describe and analyze fish and wildlife habitat (terrestrial Gap Analysis Program and Aquatic Gap Analysis Program); and explicit mapping of areas protected with conservation measures (Protected Areas Database of the United States).

CSAS&L will continue to expand the domain science expertise to include the full range of earth sciences, in keeping with the Core Science Systems Strategy of characterizing the "Critical Zone," the physical region of the Earth that extends from the tree tops down to the deepest groundwater (Bristol et al., 2013).

Goal 1) Leadership in biodiversity data science

Continue to be an authoritative source for data (e.g., occurrence, abundance, distributions, movement, taxonomy) on species and plant and animal communities to guide conservation and resource management decisions. CSAS&L is committed to continue leading data integration and dissemination efforts based on high quality data standards, domain-specific standards, and protocols (e.g., quality assurance, metadata creation, applying standard classifications and terms) to facilitate understanding, interoperability, and synthesis of national data assets in support of Department of the Interior (DOI) mission and priorities.

Research Question(s)

1. *How can biodiversity data be synthesized effectively at relevant (e.g. national, regional) scales to create policy-relevant knowledge products?*
2. *What advances in data management practices are needed to improve access to, and the quality and application of, biodiversity data developed or aggregated by CSAS&L? How can "data" understanding be improved by modeling, field data collection, metadata, use of consistent standards, or web-based tools?*

Sub-questions

1. *How can biological and abiotic data be combined to allow for increased interoperability?*
2. *Is the terminology in current biological domain data standards sufficient to be used in semantic applications? If not, what glossaries need to be strengthened/built?*

Strategic Actions

1. *Maintain and strengthen partnerships with bureau, departmental, national and international science committees (e.g., Committee on Environment and Natural Resources Research, Subcommittee on Ocean Science and Technology, Intergovernmental Oceanographic Commission and the Federal Geographic Data Committee, National Wildland Fire Cohesive Strategy).*
2. *Represent USGS in appropriate DOI initiatives to ensure that USGS capabilities serve Departmental needs and that the Department's policies and decisions consider USGS science needs.*
3. *Continue to formalize current relationships within departmental, national and international organizations (Memorandums of Understanding, Interagency Agreements and plans).*
4. *Determine data flow processes that can better serve bureau data management needs in the areas supporting biodiversity science (e.g., species observations, monitoring).*
5. *Develop and strengthen the biodiversity informatics research and development capacity of CSAS&L.*

EcoINFORMA

CSAS&L is contributing to the establishment of EcoINFORMA, a collaborative initiative for biodiversity and ecosystems open data. CSAS&L is leading and coordinating among Federal agencies the implementation of key EcoINFORMA components. Biodiversity Information Serving Our Nation (BISON) is the designated Biodiversity Resource Hub of EcoINFORMA, currently one of only three designated EcoINFORMA Resource Hubs. BISON facilitates interoperability of species occurrence data by collaborating with data providers to integrate their

data into a national scalable system. The President's Council of Advisors on Science and Technology recommended in the Report to the President on Sustaining Environmental Capital: Protecting Society and the Economy that the "EcoInformatics-based Open Resources and Machine Accessibility (EcoINFORMA) is needed to ensure that Federal agency data relevant to biodiversity and ecosystems, as well as the socio-economic and geophysical data required in support of ecosystem valuation and decision-support, are published in machine-readable, interoperable format to facilitate research engagement by public, private, academic, and other stakeholders, and to support policy- and decision-making at Federal, state, and local levels."

Integrated Ocean Observing System

Advances in data science applied through CSAS&L leadership provided the basic technical framework (biodiversity data schema, marine biogeography) enabling the interoperability of marine biogeographic data collected by the Federal Integrated Ocean Observing System (IOOS). The innovations allow regional associations to share data, discover inter-regional relationships between biota, as well as contribute to a broader view of our nation's marine biota. Building on this success, future plans for a national marine biodiversity observation network (BON) will incorporate these same standards.

Goal 2) Research and analysis supporting biodiversity assessments

Characterize species and community distributions and status using state of the art modeling and analysis techniques to support the understanding of ecosystems at relevant scales. CSAS&L maintains data assets related to terrestrial, aquatic, and marine species and ecosystems. Those data assets should be used to create a comprehensive portfolio of analyses that provide a baseline understanding of the conservation status of those elements of biodiversity. In addition, those data and analyses should be used in conjunction with a range of threat assessments to model the potential impact of landscape changes (e.g., climate, land use, invasive species) on those same elements.

Research Question(s)

1. *How are species, and the ecosystems they depend on, changing, and what are the key drivers of change?*

Sub-questions

1. *Which species are most vulnerable to climate change?*
2. *What are the primary threats to natural systems and to the native species in each region of the U.S.?*
3. *What is the range of policy/management alternatives available to address those threats?*

Strategic Actions

1. *Expand existing CSAS&L assessments of biodiversity to inform landscape scale management.*

2. *Engage other USGS and Department of Interior programs to address important national science objectives such as reporting on the status of biodiversity at regional and national scales.*
3. *Work with partner programs to identify key data resources on primary threats to biological diversity.*
4. *Develop advanced applications to support use of heterogeneous data from diverse sources, including: graph analysis and structured data mining for combining disparate incompatible taxonomies; develop innovative algorithms for remote sensing of species and communities; and assess the completeness of models and datasets in CSAS&L products.*
5. *Collaborate with policy analysts on identifying the range of management alternatives to be explored in biodiversity assessments.*

National Gap Analysis Program

Wildlife species models developed by the National Gap Analysis Program are an integral component of the Crucial Habitat Assessment Tool developed by the Western Governors' Association. The tool provides consistent information across the Western U.S. to help address wildlife conservation questions (e.g., wildlife corridors, priority conservation areas).

Western Governors' Association Policy Resolution 13-04: Conserving Wildlife and Crucial Habitat in the West:

> *"Federal and state agencies need data-driven science, mapping and analysis to manage species and habitat. That same information is needed to address other challenges, including the potential listing of species (such as the Greater Sage-Grouse), the spread of invasive species, the impacts of drought, water transfers, and energy development."*

Goal 3) Partnerships and community collaboration

Engage communities of practice and partnerships to advance methods, tools, and access to data based on stakeholders' needs (e.g., specific resource questions, mandates, or initiatives). Identify key communities of practice related to CSAS&L products. These partnerships would address data lifecycle and data management practices as well as data of key communities [Community for Data Integration (CDI), Committee on Environment, Natural Resources, and Sustainability (CENRS), National Ocean Council (NOC), Global Biodiversity Information Facility (GBIF), International Oceanographic Data and Information Exchange IODE (iOBIS), Taxonomic Databases Working Group (TDWG), Federal Geographic Data Committee, Association of Fish and Wildlife Agencies, Integrated Digitized Biocollections (iDigBio), DataONE, System for Earth Sample Registration (SESAR), PaleoBioDB (PBDB)].

Research Question(s)

1. *What are the pressing biological diversity-related questions that can be more effectively addressed through broader participation, and what partnerships should CSAS&L prioritize to facilitate effective analyses related to those questions?*
2. *What advances are needed in biodiversity data management to improve discoverability, and the quality and visualization of, critical zone data developed or managed by USGS?*

Sub-questions

1. *What are the current best practices with respect to species distribution modeling in terrestrial, aquatic and marine systems?*
2. *What are the current business practices that encourage integration of data collected from diverse science disciplines to better inform our understanding of ecosystem resiliency and function, and species health and survival?*
3. *What are the best practices or methods for delivery of scientific data and information for decision-making?*

The following questions related to partnerships and communities are dealt with more directly in the Human Dimensions Section of this document; they are listed here to provide context.

1. *Based upon the experiences of different "data sharing communities" (from local to global) what are the factors that inhibit data sharing that we can address, and what are the factors that encourage data sharing?*
2. *How can we use communities of practice to help frame, establish, and implement policies and business practices that will improve data management within Federal agencies, and in support of CSAS&L products and goals?*
3. *How can we characterize key communities of practice and understand the constraints/barriers to data aggregation and meta-analysis across these communities?*
4. *How can we address these constraints/barriers to more effectively engage these communities and other partners to work toward aggregating data at the local and regional levels to foster meta-analyses that meets stakeholder needs?*

Strategic Actions

1. *Initiate a community of practice within the Community for Data Integration specific to species habitat modeling.*
2. *Formalize and communicate the process for prioritizing CSAS&L data assets that will most efficiently enrich the holdings in support of relevant policy initiatives.*

USGS National Paleontology Database

CSAS&L is working to establish a USGS National Paleontology Database along with the Integrated Earth Data Applications (IEDA) facility, and the Paleobiology Database to expose information on paleontological specimens for discovery by scientists and citizens. Sample records and taxa identification from Examine and Report files are being processed from the Reston Nanofossil Lab and the Denver Core Research Center.

"The USGS National Paleontology Database is being set up to be accessed primarily through two, best in class "third party" distribution vehicles. The System for Earth Sample Registration provides the registry for physical sample/specimen information and assignment of International Geosample Numbers (IGSN) that are linked to fossil identification and taxa information in the PaleoBiology Database. The IGSN links the two systems together such that a user finding information about a USGS fossil specimen in either system can cross reference to get the full information tracking all the way back to the lab or archive that maintains the physical data." (Arko et al., 2014)

Goal 4) Expand access to, and application of, biodiversity data through innovation

Actively investigate the addition of evolving data types (e.g., DNA), new methods (e.g., citizen science, semantic linkages), and visualization and decision support technologies (e.g., web services, biogeographic analysis tools) to meet emerging stakeholder needs.

Research Question(s)

1. *What are the data collection activities (e.g., partners, technologies, citizen science) that provide the best data, in an effective and efficient manner, and what factors will encourage those sources to provide access to the data?*
2. *What are the current data gaps with respect to addressing the highest priority questions related to biological diversity?*
3. *What technologies will be needed to support collection of biological observations and to provide access to data in real or near-real time?*
4. *What technologies and processes could facilitate delivery of data and information in a more biologically-robust, accessible fashion?*

Sub-questions

1. *What new connections/patterns emerge when semantic linkages between related but distinct biological datasets are discovered?*
2. *How can citizen science data be effectively integrated into a statistically robust modeling framework?*
3. *Can we quantify uncertainty in modeling species occurrence in order to inform and guide citizen science (where, when, species of interest)?*

Strategic Actions

1. *Expand ongoing semantic technology research to explore linkages between related but distinct biological datasets.*
2. *Expand on partnerships with the citizen science community and statisticians to explore statistically robust modeling frameworks to better inform policy and resource management.*
3. *Increase access to biodiversity data and CSAS&L products by providing and registering coordinated and well documented web services.*

CSAS&L Biodiversity Science Products

CSAS&L is a national leader among federal agencies in collaborations on biological diversity and open data. The efforts span terrestrial, freshwater, and marine biodiversity and include a diverse array of activities including:

***Biodiversity Information Serving Our Nation (BISON)** is leading among Federal agencies the integration and interoperability of species occurrence data in order to discover the patterns of distribution and correlations among all species in the United States and the environments at landscape scales.*

Integrated Taxonomic Information System (ITIS) *represents an international partnership working together to provide the authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world.*

National Fish Habitat Partnership (NFHP) *implements the National Fish Habitat Action Plan with the primary goal of protecting, restoring, and enhancing the nation's fish and aquatic communities.*

National Gap Analysis Program *produces core datasets (land cover, protected areas database, and wildlife distributions) and tools to guide proactive management key ecosystems and the wildlife they support.*

National Aquatic Gap Analysis Program *evaluates aquatic biological diversity and aquatic habitats using spatial analysis and habitat suitability models to identify gaps in species distribution and work toward more effective conservation prioritization.*

National Vegetation Classification Standard (USNVC) *represents a central organizing framework for how all vegetation in the United States is inventoried and studied, from broad scale formations (biomes) to fine-scale plant communities.*

Ocean Biodiversity Information System – USA (OBIS-USA) *is part of an international data sharing network that brings together marine biological occurrence data throughout the U.S. and where U.S. funding supports the research and monitoring.*

Applied Research Computing

Today's societal challenges compel scientists to approach problems at increasingly larger scales, which require intensive computational power. Traditional computers are often not powerful enough to handle large computational problems in reasonable amounts of time, pushing scientists to navigate advanced computing paradigms without any traditional computer science training. High Performance Computing (HPC) clusters, often called supercomputers, offer a network of computing nodes, each containing its own processors and memory, which work collectively to solve these large scientific computational problems at more reasonable speeds than can traditional computers. Applied Research Computing is the investigation of high performance computing methodologies, software, and hardware that can be used for the advancement of scientific discovery within any discipline. As the technology of HPC progresses and research data grows in scale, CSAS&L provides high performance computing capabilities and expertise to scientists for the acceleration and expansion of scientific discovery. The following goals have been outlined to further extend and enhance this mission:

Goal 1) Investigate advances in computational hardware, technologies, and methodologies for the advancement of scientific computing and large-scale data management within USGS.

Research Questions

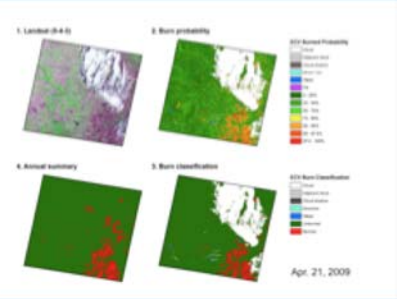
1. *Are innovations in coprocessor technologies viable and effective solutions for USGS scientific computing needs?*
2. *How can traditional HPC technologies be intertwined with newly emerging large-scale data-intensive computing frameworks to better serve the computational and data management needs of USGS scientists?*
3. *How can data visualization techniques and tools be best utilized on large scale computational clusters to analyze data and better communicate USGS science?*
4. *Which types of high performance storage systems are best suited to address issues with bottlenecks in Input/Output that occur within many scientific codes?*
5. *How can cloud computing services augment current HPC capabilities to create efficient, cost effective services, such as on-demand cloud provisioning, for researchers in need of short term computational resources beyond the scope of currently available systems.*

Strategic Actions

1. *Continually evolve and adapt our technical footprint to meet the changing computational needs of the USGS science community.*
2. *Develop a CSAS&L research agenda specific to computing methods for processing, modeling, and analysis of large-scale data for understanding the Earth's critical zone.*
3. *Architect and communicate high performance scientific computing solutions to technical barriers impeding the USGS science mission.*

National Science Foundation

"Dramatic increases in computing power, storage capacity, and networking capabilities are widening the scope of investigation and allowing scientists to address increasingly complex questions" (National Science Foundation, 2007).



Burned Area Essential Climate Variable Algorithm
The objective of this study was to develop an algorithm to automatically identify burned areas in Landsat imagery.

Principal Investigator
Todd Hawbaker, Geosciences and Environmental Change Science Center

Computational Challenge

"On my workstation (3.2 GHz Xeon W3670 w/ 24 GB RAM, writing across the network), it takes about 55 minutes for R to process 1 landsat scene. There are 410 scenes in our test site, so with 1 thread, ***it will take about 15.66 days to process all 410 images****!"* Todd Hawbaker

Applied Research Computing Solutions

ARC staff worked with Hawbaker to parallelize code to run efficiently within an HPC environment and automated the processing pipeline to reduce manual data manipulation prior to model runs. Using these advanced methods, analysis of all 410 scenes in the test site can now be completed in **less than 1 hour** on CSAS&L's HPC cluster.

Goal 2) Enable Expertise in High Performance Computing

Applied Research Computing outreach activities focus on providing educational opportunities to USGS scientists so that they become empowered with the skills and knowledge needed to effectively utilize high performance computing resources.

Strategic Actions

1. *Host HPC and large-scale data management training workshops to enhance the computational skills and knowledge needed for the current and future USGS scientific workforce.*
2. *Consult with USGS Science teams to advance the understanding and application of HPC technologies for the expansion and acceleration of science.*
3. *Facilitate a USGS wide knowledge base and collaborative forum for HPC users.*
4. *Establish a High Performance Computing student development and training program.*
5. *Partner with organizations and educational institutions with expertise in High Performance Computing to educate and enhance the computational skills of the current and future scientific workforce of the USGS.*
6. *Offer development-level high performance computing resources that provide safe learning environments where scientists can test new HPC methodologies and best practices.*

CSS Science Strategy

"The rapid pace of technological advancement and the importance placed on wise applications of technology to science described in the Core Science Systems vision demand investment in continual learning and workforce modernization" (Bristol et al., 2013).

Goal 3) Provide Access to High Performance Computing Resources

Facilitate scientists' access to the appropriate high performance computing resources and expertise in order to increase the scope and timeliness of research efforts within USGS.

Strategic Actions

1. *Provide access to high performance computing resources to all USGS researchers who require them.*
2. *Partner with other government agencies, universities, and regional computing consortia to expand the opportunities for USGS scientists to collaborate with scientific computing experts in the field.*
3. *Improve the access methods and networking capabilities for the processing, transfer, and management of large data sets.*

USGS Science Strategy

"The USGS will use its information resources to create a more integrated and accessible environment for its vast resources of past and future data. It will invest in cyberinfrastructure, nurture and cultivate programs in Earth-system-science informatics, and participate in efforts to build a global integrated science and computing platform" p. 56 (USGS, 2007).

Goal 4) Collaborate to create efficiencies throughout the realm of computational capabilities and communities of expertise within the USGS, DOI, partner organizations, and the broader scientific community.

Strategic Actions

1. *Collaborate with USGS advanced computing resource owners to establish a cooperative community to identify and create efficiencies throughout the USGS.*
2. *Partner with NSF funded programs, university computing groups, and other federal agencies to expand the scope of available high end computational resources and HPC experts available to help USGS researchers with their computational and large-scale data management needs.*
3. *Participate in professional communities of practice to gain expertise and model our Applied Research Computing program after successful scientific HPC center endeavors.*

USGS Helping Scientists Utilize High Performance Computing Resources through the Extreme Science and Engineering Discovery Environment (XSEDE):

XSEDE
Extreme Science and Engineering Discovery Environment

The USGS Core Science Systems mission area has signed a Memorandum of Understanding (MOU) with the National Science Foundation (NSF)-sponsored Extreme Science and Engineering Discovery Environment (XSEDE)Project. The XSEDE network is the most powerful and robust collection of integrated advanced digital resources and services in in the world. It is a single virtual system that allows scientists to interactively share computing resources, data, and expertise. The MOU designates USGS as the first government

participant in the XSEDE Champions program. USGS scientists will now have direct access to XSEDE, its expert staff, computational resource allocations, and technical assistance. The champions program designates local representatives as sources of knowledge related to high-performance and high-throughput computing and other digital services, opportunities and resources. CSAS&L ARC staff will act as the champions for USGS by providing knowledge and guidance to USGS researchers to advance scientific discovery and analysis.

Library and Information Science

USGS produces some of the most comprehensive and impactful scientific data in the world. USGS research routinely advances understanding on key societal issues, informs planning by governments and businesses, and is used prominently in disaster response activities. Integrating, managing, and ensuring access to critical data and publications in a timely manner represents an increasingly acute challenge for the bureau. CSAS&L enables access to USGS scientific data and publications by partnering with USGS programs/centers producing these data to facilitate lifecycle data management practices, and by providing comprehensive library resources and services through the USGS Libraries Program.

Goal 1) Identify and implement best practices, policies, and standards for lifecycle management of USGS data.

USGS scientists, data stewards, and managers operate at all levels of the Science Data Lifecycle from planning and acquisition, to preserving and sharing data. Managing USGS science data throughout the lifecycle ensures appropriate credit and recognition to authors and encourages access and discovery of data. CSAS&L seeks to identify the data management challenges facing USGS scientists and staff and explore new solutions to improve data management practices.

Research Questions

1. *What are the relevant practices, techniques, tools, standards, and methodologies necessary to provide seamless access to all USGS data artifacts (metadata, data, publications, citations, media, people, etc.) for discovery, use, and visualization by multiple stakeholder audiences?*
2. *How can we identify and track data and related contextual information immediately upon creation all the way through release and usage over time?*

Strategic Actions

1. *Engage with USGS centers and programs as well as the science data management community to test and identify best practices for lifecycle data management.*
2. *Implement processes and workflows for documenting new and existing data types (e.g. versioning for dynamic datasets and mashups).*
3. *Encourage the adoption of data standards and best practices to improve the exchange of data as well as data reuse.*

Community for Data Integration (CDI)

CSAS&L employees have played a critical role in the Community for Data Integration (CDI) Data Management Working Group's efforts to develop and strengthen policies addressing metadata and the release of USGS data. These policies form the basis for important modifications to the USGS Survey Manual that will detail and reinforce lifecycle data management practices for the Bureau's scientific data, and facilitate compliance with federal Open Data access requirements.

Goal 2) Provide and sustain ubiquitous access to library collections, resources, and services using collaborative technologies that support research, knowledge transfer, and scientific discovery.

The USGS Libraries Program is a multi-site, centralized library system with over 1,500,000 volumes, 500,000 maps, 30,000 field records, and 500,000 historical photos in the collection. Specialized library collections include USGS-authored publications, field records, photos, video, film, maps, and multidisciplinary, peer-reviewed publications supporting USGS science activities. In addition to the extensive physical collections, the USGS Libraries Program provides electronic access to eJournals and eBooks, resulting in an average of 1400 article downloads per day. Making these unique resources accessible and discoverable through integrative technologies and processes is a high priority.

In response to the OSTP's Open Government Directive (Holdren, J.P., 2013), the USGS Libraries Program has started to digitize USGS publications that were previously available only in print. These digitized materials, as well as born-digital publications, are stored on and served from a variety of servers, hard drives, and systems. This lack of a single point solution leads to confusion over which group "owns," controls, and is responsible for the objects. Preservation methods are not standardized and the digital

copies are not curated post-production. A digital repository solution would provide sustainable long term preservation and access to USGS digital assets.

Strategic Actions

1. *Implement an Integrated Library System that will increase functionality within the Library and allow integration with other USGS information products to improve efficiency of the public online catalog interface and the internal application used by the library staff for cataloging and circulation transactions.*
2. *Create a plan to address scope, process and specifications, and resources needed to digitize unique historical Library holdings and USGS publications.*
3. *Examine approaches and potential implementation of a digital repository solution for the access and preservation of USGS publications and research data.*

USGS Libraries Program

The USGS Libraries Program is the premier and largest earth science library in the world. Of the physical materials held by the Library, 7.9% are unique and only owned by the USGS, and more than 36% of the materials are held by fewer than 10 libraries globally. Scanning and routing of existing print resources mitigated $2.2M in deferred costs to fill scientist requests for documents in FY13. In the digital information arena the USGS Libraries Program provides access to over 1400 eJournal titles, 2,750 eBooks, and numerous other databases and online resources through a single search point. This single search point facilitates over 524,000 article downloads per year from USGS scientists. The cost to order articles individually would be over $23M, nearly 12 times the total spent on collections in FY13. Additionally, the Library provides access to USGS publications through the Publications Warehouse, resulting in over 720,000 unique visitors per year.

Goal 3) Improve the discoverability, interoperability, and successful use and reuse of USGS data and information products.

Over the past two decades, successive administrations, congresses, academic reports and private sector initiatives have encouraged broad sharing, reuse and openness of government data to support decision making and foster economic growth. Scientific data supplied by USGS and its partners represent one of the most critical components of this broader data sharing enterprise; these shared resources have routinely supported high profile energy development, land and ecosystem management, water resources policies, and disaster response. Recent Open Data mandates require Federal agencies to make their research data products more readily discoverable, accessible, and usable by the public. CSAS&L, in coordination with other USGS programs and external partners, is helping to organize data management, integration and delivery efforts across the bureau to continue expanding the impact and use of USGS data.

Implementing best practices, such as supporting open data formats and encouraging the use of data standards, will improve the exchange of data as well as data use and reuse.

Research Questions

1. *What are the characteristics, capabilities, and requirements of a repository solution for USGS data and data products? How would a repository be established, resourced, implemented, managed, and used within USGS?*
2. *Would an integrated data, metadata, and publication system facilitate improved discovery and access to both physical and digital USGS and USGS-licensed data and information assets?*
3. *What metrics and analytical techniques can facilitate understanding of how USGS data are used, beyond the original research purpose?*

Strategic Actions

1. *Initiate and co-lead a USGS focus group to determine needs and requirements for repository solutions, and develop a repository strategy for seamless access to USGS information assets.*
2. *Employ data citation mining techniques to document and analyze other uses of USGS data products. Facilitate tracking of USGS data citation by providing tools, training, and support for assignment of Digital Object Identifiers to these products.*
3. *Identify and evaluate USGS metadata catalogs, enterprise data systems, and publications mechanisms, and explore opportunities for optimizing stakeholder discovery of, and access to, the Bureau's data products.*
4. *Improve interoperability by providing application programming interfaces and web-services to USGS data and encourage the use of open data standard formats.*
5. *Explore how USGS information products continue to benefit society at large and support stakeholder needs for relevant and timely data.*

CSS Science Strategy

"The Core Science Systems mission includes responsibilities for scientific data and information management, curation, archiving, preservation, and long term availability." (Bristol, et al., 2013).

USGS Science Data Catalog (SDC)

The Science Data Catalog (SDC) developed by CSAS&L describes USGS-produced, authoritative scientific data that have undergone the data release process outlined in the USGS Fundamental Science Practices. The SDC includes metadata describing individual datasets, data collections and observational or remotely-sensed data contained in national systems (rather than records about individual observations). The SDC meets Open Data requirements by partnering with Centers and Programs across USGS to make their data more accessible and discoverable.

Human Dimensions

The data and tools CSAS&L produces are intended for use by a variety of groups such as natural resource managers and policy-makers, government and academic researchers, and the general public. For these stakeholders, the relevance of CSAS&L products is directly dependent upon their applicability and use towards natural resource conservation issues and societal concerns. In order to develop the most effective data and tools for these communities of practice, we need to understand both the people who use these products and the context for their use. What are the needs of these stakeholders? What specific questions and issues are they trying to address or resolve? What are their motivations to seek out and use data and tools, and what requirements for use do they have? What are the barriers that prevent them from using data and tools? How do they identify, access, integrate, and apply the data and tools to address the natural resource conservation questions that are important to them? Understanding and characterizing communities of users will help CSAS&L to improve existing and develop new data and tools that meet the specific needs of these users and ensure that our products are successfully used and applied to answer critical natural resource questions.

DataONE

"Principal data challenges lie in discovering relevant data, dealing with extreme data heterogeneity, and converting data to information and knowledge. Addressing these challenges requires new approaches for managing, preserving, analyzing, and sharing data, which are equally challenging from technical and socio-cultural perspectives." (Michener et al., 2011).

U.S. Digital Services Playbook

"The American people expect to interact with government through digital channels such as websites, email, and mobile applications. By building better digital services that meet the needs of the people that use our services, we can make the delivery of our policy and programs more effective…We created a playbook of 13 key "plays" drawn from successful best practices from the private sector and government that, if followed together, will help government build effective digital services."

> *Play #1 Understand what people need: "We must begin digital projects by exploring and pinpointing the needs of the people who will use the service, and the ways in which the service will fit into their lives...The needs of people — not constraints of government structures or silos — should drive technical and design decisions."*

For the entire list of 13 key plays see: http://playbook.cio.gov/

In assessing the human dimensions aspect of data use, CSAS&L will consider a variety of social science analysis tools such as surveys, stakeholder analysis, focus groups, interviews, and assessment of user attitudes/knowledge/values. We will build on efforts that are already underway to understand aspects of data use. For example, the findings of a 2014 survey of USGS scientists and data managers, conducted by CSAS&L in coordination with the University of Tennessee, illuminate several Bureau priorities for improving lifecycle data management policies, processes, and capabilities. CSAS&L also will coordinate with social science researchers who have expertise in this type of assessment, such as the USGS/Fort Collins Science Center, Policy Analysis and Science Assistance (PASA) Branch, who lead projects that integrate social, behavioral, economic, and biological analyses in the context of human–natural resource interactions. PASA has conducted a number of studies that involve assessing human dimensions.

Understanding and characterizing communities of practice is also a focus of the Human Dimensions Science Theme. These communities are formed by individuals engaged in a process of collective learning who share a concern for something they do and how to do it better through their regular interaction (http://wenger-trayner.com/theory/). The USGS Community for Data Integration is an example of a data-focused community of practice. CSAS&L needs to understand communities of practice and how to best involve them in our data science efforts, including the development of policies and best practices.

Understanding and characterizing the impact of emerging high performance computing technologies on USGS scientists and their research is a focus of the Human Dimensions Science Theme that supports the Applied Research Computing Science Theme. To support Biodiversity science activities of CSAS&L, the Human Dimensions science theme will also focus on characterizing the drivers that lead to development of CSAS&L science products. Understanding how these drivers affect our priorities and how these priorities help us to address the natural resource issues and societal concerns of our users is critical to developing relevant biodiversity science products.

Goal 1) To understand and characterize various communities of users of data and tools in order to ensure CSAS&L data products meet their needs.

Research Question(s)

1. *What different types of user communities need access to what types of CSAS&L data products and tools (for example, natural resource managers and policy makers, government and academic researchers)?*
    a. *What types of data do these user communities need to acquire and use?*
    b. *What types of tools do these user communities need in order to interact with and use the data?*
    c. *What are the specific natural resource questions/issues and societal concerns these user communities are trying to answer/address with the data and tools?*
    d. *What are the motivations and barriers to use of data and tools for these user communities?*
2. *Based upon the experience from different "data sharing communities" (from local to global) what are the factors that prohibit or encourage data sharing that we can address?*
    a. *USGS researchers represent one user community for CSAS&L data products and tools. What are their expectations about data reuse, credit, usage responsibility, and other dynamics of data use, and to what extent do these expectations align with current USGS policies, practices, and technologies?*

USGS Data Management Assessment

Lifecycle management of USGS data and perceptions of data sharing can be heterogeneous across the Bureau. It is generally unknown how USGS scientists and data managers manage and share their science data. CSAS&L developed a Data Management Assessment survey to better understand and address specific challenges and opportunities for data sharing and management excellence within the Bureau. Results from this investigation will guide the development of targets that will allocate resources towards maximizing its return on investments. This study will also provide a "baseline" for the overall impacts and value of better data management practices within the Bureau that can then be measured and monitored in the future.

Goal 2) To understand and characterize communities of practice and how to best involve them in CSAS&L data science and data management efforts.

Research Question(s)

1. *What types of communities of practice are needed to further USGS data science and data management goals and how can we better learn from our existing relationships?*
2. *What are the characteristics of successful communities of practice (e.g., established mechanisms for social interaction)?*
3. *How can we characterize key communities of practice to better understand the constraints/barriers to data aggregation and meta-analysis across these communities? How can we address these constraints/barriers to more effectively engage these communities and other partners to work toward aggregating data at the local and regional levels to foster meta-analyses that meets stakeholder needs?*
4. *What are the sociological, organizational, and disciplinary challenges to networked data repository architectures (e.g., DataONE), and how can they be mediated to develop organizational norms for implementing this framework?*
5. *How can communities of practice help to frame, establish, and implement policies, business practices, and standards that will improve data management within Federal agencies and in support of CSAS&L products and goals?*
    a. *How do we encourage the application of emerging biodiversity standards based on our understanding of the social dynamics of the development, adoption, and evolution of data management standards?*
    b. *How can we influence data management practices to ensure that biological data are captured and described sufficiently to support future applications or integration with "big data" or other initiatives?*

Data Release

As USGS Programs and Centers develop a new generation of data services, the question of how and whether or not data can be released has become a growing concern. Current policies and processes facilitate the release of publications but data release has not entirely been addressed. In collaboration with the Community for Data Integration (CDI), CSAS&L helped to form a Data Release Use Case working group with representatives across the Bureau to explore possible workflows for the release of data. The team has constructed several use cases including data release via the web for review by USGS leadership. This effort may form the basis for important modifications to the USGS Survey Manual that will detail and reinforce lifecycle data management practices for the Bureau's scientific data, and facilitate compliance with federal open data access requirements.
https://my.usgs.gov/confluence/display/cdi/Data+Release+Use+Case+Team

Goal 3) To understand the socio-cultural aspects of technology diffusion and adoption as they relate to CSAS&L products and users.

An important part of understanding our users is also to understand how the social groups they are part of affects their process for adopting new technologies. Diffusion of innovations is an area of study in social science that relates to the process of spreading and adopting innovations within a social system (Rogers 2003). Understanding this diffusion process and barriers to adopting new technologies can help CSAS&L to address these barriers in developing our products for a variety of users.

Research Question(s)

1. *How can we characterize and describe the process for diffusion and adoption of technologies for various communities of users, as identified under Goal 1, in order to identify barriers that should be addressed as we develop CSAS&L products?*
2. *How can we characterize and describe the process for diffusion and adoption of new data management policies and best practices for various communities of users in order to identify barriers that should be addressed as we develop new policies and best practices within USGS?*

Strategic Actions for Goals 1 through 3 of the Human Dimensions Science Theme

The following strategic actions apply to Goals 1 through 3 of the Human Dimensions Science Theme. These strategic actions are intended to: 1) describe the actions needed to develop a baseline understanding of relevant human dimension studies that can be used to inform future CSAS&L human dimensions research and 2) describe the actions needed to build and leverage partnerships to help develop and implement a detailed plan for a CSAS&L human dimensions study.

Strategic Actions

1. *Build on efforts already underway to understand human dimension aspects of data and technology use; produce a summary of existing human dimensions studies that will be used to guide development of a plan for CSAS&L human dimensions research.*
    a. *Review results of the USGS Scientists and Data Managers Assessment to identify information that was collected that could address some of the human dimensions research questions in this Science Strategy.*
    b. *Review past and current work of the USGS Community for Data Integration to identify pertinent information that may address some of the human dimensions research questions.*
    c. *Review studies from the USGS Fort Collins Science Center, Policy Analysis and Science Assistance Branch for relevance to CSAS&L's needs to understand human dimensions.*
    d. *Identify other human dimensions studies to review for relevant information and assessment methods.*

2. *Coordinate with groups involved in social science/human dimensions research with expertise in this type of assessment in order to plan and implement a human dimensions study.*
    a. *Develop and implement a CSAS&L human dimensions study in collaboration with key partners.*
    b. *Leverage existing user assessment partnerships with University of Tennessee (UT) and DataONE.*
    c. *Build a working relationship with the USGS Fort Collins Science Center (FORT), Policy Analysis and Science Assistance (PASA) Branch.*
    d. *In coordination with FORT/PASA and other partners such as UT and DataONE, develop a structure and plan for a human dimensions study that addresses the goals and research questions identified in this Science Strategy and implement this study.*
3. *Publish a USGS Open File Report or other publication on the results of a CSAS&L human dimensions study.*
4. *Using the lessons learned from the CSAS&L study, develop a plan for a USGS-wide human dimensions study focusing on data and products from the Modular Science Framework as a whole.*

Goal 4) To understand how the availability of high performance scientific computing solutions affects the ability of USGS scientists to address increasingly complex science questions and can enhance the collaboration of scientists across multiple disciplines. To better understand and plan for the impact these capabilities will have on the future working environment of USGS scientists.

Research Question(s)

1. *How can we describe the current effect of computational limitations on USGS scientists' research scope and the effort required to complete standard analyses?*
2. *How can we describe the potential effect of high performance computing (HPC) capabilities on USGS scientists' research scope and activities in the future? How can we characterize and describe how different types of computing processes best address different types of data and science questions?*
3. *How can we characterize and describe the process for diffusion and adoption of HPC technologies for users in different scientific disciplines in order to identify barriers that should be addressed as HPC capabilities are developed within USGS?*

Strategic Actions

1. *Using information gained from Research Question 2, assess the future impacts of HPC capabilities on the working environment and relationships of USGS scientists across all scientific disciplines.*

2. *Using information gained from Research Questions 2 and 3, develop a plan for addressing impacts on the research work environment that reduces barriers and enhances motivations to work in this new computing environment for all USGS scientists.*

Goal 5) To characterize the organizational and social drivers that lead to development of CSAS&L science products. To assess our process for identifying priorities based on these drivers and how well these priorities meet the needs of our communities of users.

Research Question(s)

1. *What are the organizational goals (e.g., DOI, USGS, other federal agency) that drive the development of each CSAS&L product? Both long-term and short-term drivers.*
2. *What are the social drivers (e.g., conservation or economic issues, societal concerns) that drive development of each CSAS&L product? Both long-term and short-term drivers.*
3. *How are these drivers used to identify CSAS&L priorities in the short and long-term?*
4. *How do the priorities identified address the user needs, resource issues, and societal concerns that are characterized by Goal 1, Research Question 1 of the Human Dimensions Science Theme.*

Goal 6) To understand how USGS Modular Science Framework (MSF) products (data, information, and knowledge) can be applicable to decision-making at three levels: a) national-level policy; b) regional conservation and resource management actions; and c) local conservation and resource management actions. To understand the results and impacts of the use of MSF products at these three decision-making levels.

Research Question(s)

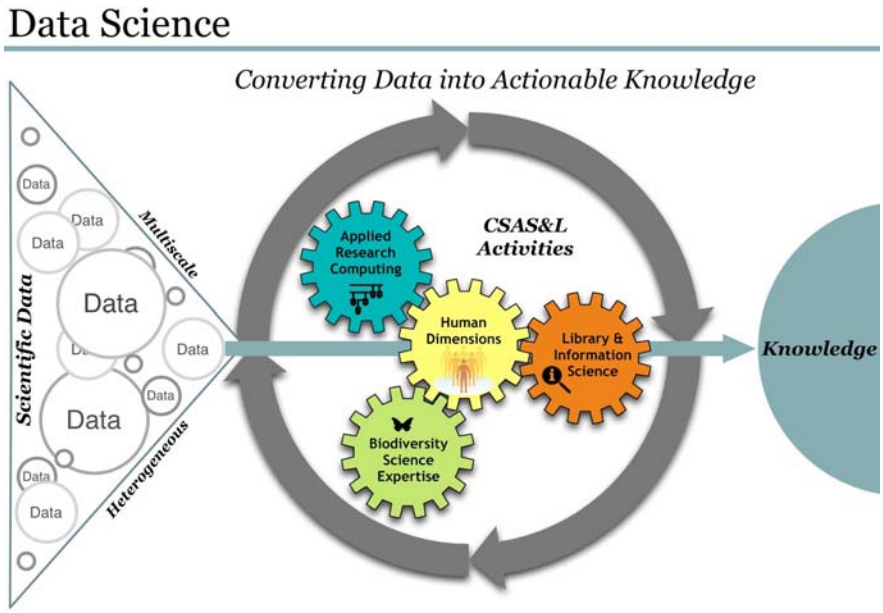
1. *How are existing USGS Modular Science Framework products used by various communities of users at the national, regional, and local levels to address natural resource policy/conservation/management issues and societal concerns?*
2. *What are the results of the use of USGS science products by these communities of users at the national, regional, and local levels? (For example, a regional resource management action identified by natural resource managers based on use of a specific science product.)*
3. *What are the impacts of the use of USGS science products at the national, regional, and local levels?*

Data Science

Data Science, an emerging, cross-disciplinary science, weaves together threads of expertise taken from computer science, information science, mathematics, statistics, and domain specific sciences (National Science Board 2005). In the simplest terms, Data Science is the practice of converting data into actionable knowledge (Nicole 2012). It focuses on 'big data,' which

emphasizes not only the ability to merge together many individual datasets, but also the "things one can do [with those data] at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationships between citizens and governments, and more" (Mayer-Schönberger and Cukier 2013).

Data, and specifically large integrated or synthesized data systems at regional or national scale, are the foundation for scientific research and discovery within and beyond CSAS&L. Mayer-Schönberger and Cukier (2013) observe that historically, the available tools and resources to collect, organize, store, and analyze data have limited the abilities of scientists to work with very large, multidisciplinary datasets, resulting "unconscious self-censorship" that restricts the breadth and quantity of information included in scientific analyses. Mayer-Schönberger and Cukier contrast these past limitations to the capabilities engendered today by 'big data' initiatives, which "rel[y] on all the information, or at least as much as possible…[and] allow[] us to look at details or explore new analyses….and test new hypotheses at many levels of granularity." Through our four research focus areas - Biodiversity Science, Applied Research Computing, Library and Information Science, and Human Dimensions - CSAS&L produces, or supports other data producers to ensure, data that are large, interdisciplinary, more accessible, better documented, more standardized, and readily integrated and processed. These capabilities facilitate informed decision-making, based upon the ability to consider a broad range of appropriate, well managed scientific data.



Goal 1) Architect a digital scientific data framework for the USGS

The digital scientific data framework is a suite of tools and capabilities that will support scientists at each stage of the data lifecycle (Faundeen et al., 2014), enabling more efficient discovery, access, and use of the wealth of scientific data from the USGS and our many global partners. The framework will include capabilities that address the mechanics of interacting with the data lifecycle at the stages of data discovery, data management planning, advanced computational analysis, data release, documentation, and preservation through digital archiving (See Figure 2 below).

The digital scientific data framework concept is a working instance of the conceptual Modular Science Framework described in the Core Science Systems Strategy (Bristol, et al., 2013). It is designed to catalyze and accelerate scientific discovery through advancements in Data Science that enable exploratory analytics on integrated and synthesized data from throughout the Critical Zone.

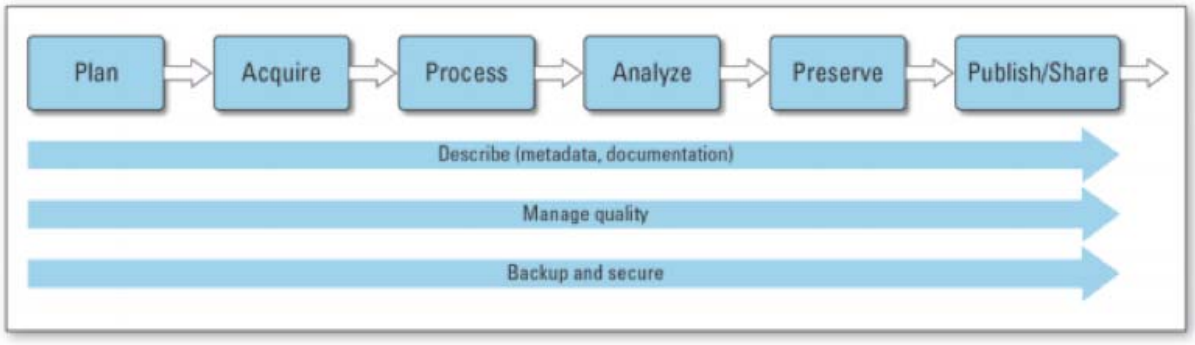


Fig 2. USGS Science Data Lifecycle Model (Faundeen et al., 2013)

Research Questions

1. *What methods and technologies can be introduced into the scientific workflow across USGS disciplines to enable more robust data documentation and efficient final data release?*
2. *How can provenance tracking methods be used to dynamically record the necessary details of the scientific data workflow to enhance reproducibility and increase efficiency of future research?*
3. *How can scientific workflow and version tracking methods be employed to build enhanced scientific publications to capture and publish results from the research process?*

Strategic Actions

1. *Create lifecycle data management tools and solutions for data management planning, documentation (metadata), persistent identifiers, and cataloging.*
2. *Provide access to High Performance Computing (HPC) resources and expertise for the processing and analysis of large-scale research data.*

3. *Produce and maintain digital holdings of USGS scientific publications, datasets, and science data products through a digital repository solution.*
4. *Design and implement the modular science framework, starting with the creation of a USGS-led integrated aquatic and terrestrial ecosystem data system that facilitates access to biodiversity information across a range of data types, data scales, and data producers.*

Goal 2) Develop the skills, knowledge and expertise in data science technologies and methodologies for the management, processing, analysis, synthesis, storage and publication of both large-scale scientific data and increasing volumes of heterogeneous, project-level datasets.

In the emerging era of Big Data, advanced methods for data collection, processing, analysis, integration, synthesis, and distribution will become increasingly critical in examining major societal issues and increasing actionable scientific knowledge. The White House Office of Science and Technology Policy (OSTP) recognized the importance of effectively utilizing "Big Data" for scientific discovery in the National Big Data Research and Development Initiative (OSTP 2012). The initiative aims to:

1. Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
2. Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
3. Expand the workforce needed to develop and use Big Data technologies.

CSAS&L takes an active role in national leadership for Big Data and related coordination activities in data science. This leadership provides USGS expertise to help inform the strategic direction of federal data policies, and develops important tactical partnerships with agencies, academic institutions, and other organizations. These partnerships help to identify opportunities, such as those presented in our collaboration with the National Science Foundation DataONE and EarthCube efforts, through which CSAS&L can leverage research and development for advanced technologies and methods. Through these partnerships we can influence research directions that will contribute to our own data platforms, participate directly in projects, and guide our own limited research budget in the most efficient direction.

Research Questions

1. *How could USGS biodiversity programs achieve a standardized data model (data structures and formats) to provide the optimal, shared framework for their observations, measurements, and models, and to support both computational data integration workflows and analytical processes?*
2. *How can advanced data structures, ontologies and other forms of explicit semantics be used to organize software-actionable relationships between heterogeneous data systems used in biodiversity and other domain specific research and drive ongoing data synthesis and recommend analytical pathways?*

3. *How can Science DMZ network architectures (Goodin, 2012) be utilized within the USGS and partner networks to aide USGS scientists with the access, transfer, and storage of large-scale scientific data for analysis within high performance computing environments?*

Strategic Actions

1. *Engage biodiversity communities of practice within USGS to research, share, and implement streamlined approaches for the effective and efficient lifecycle management, integration and sharing of large volumes of heterogeneous, multi-disciplinary, scientific data.*
2. *Continue to provide leadership and input into the Federal Government data science agenda.*
3. *In collaboration with relevant USGS IT communities of practice, investigate technologies, methodologies, and related policy transformations that will provide solutions for the storage, transfer, and publishing of extreme-scale scientific data.*

Goal 3) Research and develop scientific software methods, including advanced analysis and visualization capabilities, to accelerate scientific synthesis across disciplinary boundaries.

Scientific software is becoming an integral part of scientific publication, dramatically increasing the ability to reproduce published results, empowering peer reviewers and proposal evaluation panels to rapidly test the veracity of scientific assertions, and enabling new science to be pursued in efficient and innovative ways.

CSAS&L and other USGS programs are using scientific programming in both data integration activities and analyses. CSAS&L provides a framework for advancement of knowledge and practice through the CDI and tools like the myUSGS code repository. Other research groups throughout the USGS are pursuing similar goals for improvement in their methods and practices; therefore, CSAS&L has partners ready to collectively advance these capabilities. As we pursue other goals described in this science plan, we need to develop and employ methods that enable full provenance tracking for data integration and analyses. For CSAS&L data products, this will facilitate fuller characterization of status and trends and assertions about potential threats and changes through time.

Research Questions

1. *What indexing methods and associated data structures can support cross-platform discovery of shared scientific analysis and visualization code from public to internal code repositories?*
2. *What scientific programming methods will promote long-term viability of code and connections to data inputs and outputs?*
3. *What scientific publishing methods and practices will most seamlessly include scientific software as an integral part of journal articles and other scientific publications?*

4. *What scientific data visualization methods provide the most accurate and usable method to communicate analytical results for conservation planners and other decision makers?*
5. *How could the sharing of negative results from experimentation increase opportunities to learn from failures as well as published successes?*

Strategic Actions

1. *Expand CSAS&L scientific software research and development capability by engaging partners' efforts through ESIP, DataONE, NSF, ORNL, RDA, and others.*
2. *Enhance the CSAS&L data workflows for projects to include the scientific programming component, facilitating shared methods and practices and testing the efficacy of methods through the peer review process.*
3. *Develop and implement software management practices across CSAS&L programs that include effective modular programming, code libraries, enhanced software documentation, and use of open source projects, and test viability through metrics such as pull requests and tracked code reuse.*

Conclusion

The 2007 USGS Science Strategy outlined the key role and strengths of the USGS as the "only integrated natural resources research bureau in the Federal Government." The Science Strategy also, however, noted the severe challenges in organizing, integrating and scaling the many types of data USGS produces such that scientists and policymakers might utilize it to resolve societal challenges.

The 2010 bureau realignment, which oriented USGS to specifically address this Science Strategy, created Core Science Systems, and ultimately, CSAS&L, as one of the key integrative actors addressing this challenge. CSAS&L's science activities, including its suite of synthesized data systems, leadership in initiatives such as Big Earth Data and EcoINFORMA, and advanced repositing and sharing of USGS data and publications, position us to address this critical need described in the bureau's Science Strategy.

> ***USGS Science Strategy***
>
> *"As the only integrated natural resources research bureau in the Federal Government, the USGS has a substantial role to play in helping the economy remain strong, the environment remain healthy, and the quality of life in the United States remain high now and into the future." (U.S. Geological Survey, 2007)*

The research themes identified in this strategy, and their associated strategic goals and science questions, outline a five-year path for CSAS&L to build on existing strengths, continue improving the utility and impact of USGS science, and fulfill the CSAS&L Vision to "*[d]rive innovation in biodiversity, computational and data science to accelerate scientific discovery to anticipate and*

address societal challenges." Addressing the research questions outlined in this strategy, and implementing the associated strategic actions, will:

- link our work directly to our vision;
- support the USGS and Core Science Systems Science Strategies;
- enhance the scientific capabilities of USGS scientists;
- leverage broader initiatives such as the Big Earth Data Initiative (BEDI) and EcoINFORMA in support of our products and expertise; and
- help us better understand and serve the data needs of our stakeholders within and outside the USGS.

References

Arko, R., Bristol, R.S., Hsu, L., Peters, S., Song, L., Uhen. D.M. 2014. Exposing USGS sample collections for broader discovery and access: Collaboration between ScienceBase, IEDA:SESAR, and Paleobiology Database. Poster presented at: American Geophysical Union, Fall Meeting; 2014 Dec 15-19; San Francisco, California.

Booz, A. 2013. The Field guide to data science, 110 p., accessed 11/5/2014 at: http://www.boozallen.com/content/dam/boozallen/media/file/The-Field-Guide-to-Data-Science.pdf.

Bristol, R.S., Euliss, N.H., Jr., Booth, N.L., Burkardt, N., Diffendorfer, J.E., Gesch, D.B., McCallum, B.E., Miller, D.M., Morman, S.A., Poore, B.S., Signell, R.P., and Viger, R.J., 2013, U.S. Geological Survey core science systems strategy—Characterizing, synthesizing, and understanding the critical zone through a modular science framework: U.S. Geological Survey Circular 1383–B, 33 p., (Also available at http://pubs.er.usgs.gov/publication/cir1383B.)

Coursera, 2014, Introduction to Data Scientists Toolbox, accessed 11/5/2014 at: https://www.coursera.org/course/datascitoolbox.

Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martín, E., Montgomery, E.T., Ladino, C.C., Tessler, S., and Zolly, L.S. 2013. The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265, 4 p., (Also available at http://dx.doi.org/10.3133/ofr20131265).

Goodin, D. 2012. Securing Supercomputer Networks (Without Disrupting 60Gbps Data Flows). Ars Technica, June 26, 2012. Accessed 4/7/2015 at http://arstechnica.com/security/2012/06/science-dmz/.

Holdren, J.P. 2013. Increasing Access to the Results of Federally Funded Scientific Research. [Memorandum]. Washington, DC: Executive Office of the President, Office of Science and Technology Policy.

Mayer-Schönberger, V. and Cukier, K. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. New York: Houghton Mifflin Harcourt, 242 p.

National Science Board. 2005. Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century, pp. 27.

Nicole, K. 2012. Turning Data Into Actionable Knowledge is Key to Big Data Management [CEO Series]. Accessed 10/10/2014 at: http://siliconangle.com/blog/2012/04/16/turning-data-into-actionable-knowledge-is-key-to-big-data-management-ceo-series/.

Office of Science and Technology Policy. 2012. Big Data is a Big Deal. Accessed 11/4/2014 at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

Rogers, E.M. 2003. Diffusion of Innovations, 5th Edition, New York: Free Press, 576 p.

U.S. Geological Survey. 2007. Facing tomorrow's challenges—U.S. Geological Survey science in the decade 2007–2017: U.S. Geological Survey Circular 1309, x + 70 p., (Also available at http://pubs.usgs.gov/circ/2007/1309/.)