# Comments to the National Research Council regarding the 2017-2027 NRC Decadal Survey in Earth Science and Applications from Space

# Delivery of high volume data for Earth System Science and Remote Sensing Applications

Nancy H.F. French, Michigan Technological University, MichiganView

Micah Beck, University of Tennessee, Knoxville

P.R. Blackwell, EODN Project Consultant, TexasView

Sam Batzli, University of Wisconsin-Madison, WisconsinView

Martin Swany, Indiana University Bloomington

An important aspect of remote sensing is delivery and access to the data sets collected by earth observation systems. High-resolution views of the Earth are collected by multiple instruments, providing unique value for all Earth Science disciplines. Many satellite remote sensing systems operate continuously, generating large streams of data that are of interest to science, application, and operation communities. The nature of communication in most remote sensing applications is both high volume and asynchronous. While some remote-sensing applications require data be delivered on a demanding and rigid timetable, many do not, but they still demand access structures that provide these valuable resources when and where needed.

In this white paper we review the need for addressing the challenge of data access within the National Research Council Decadal Survey in Earth Science and Applications from Space. We represent a group of researchers and educators affiliated with the AmericaView consortium (<http://americaview.org/>) who are working on issues of earth observation data access. The impacted community for this challenge is very large, and solutions range widely. The challenge is fundamental to data use and will become more of a challenge as data volumes increase and remote sensing becomes more ubiquitous, especially globally. Due to advances in data logistics, new solutions are possible as long as there is sufficient consideration of the challenges as remote sensing systems are developed and deployed. The role of our public service agencies in providing free and open access to the valuable data collected with publically funded satellite systems needs close consideration by the Decadal Survey team.

# What are the key challenges or questions for Earth System Science across the spectrum of basic research, applied research, applications, and/or operations in the coming decade?

One of the key challenges for Earth System Science is the delivery of high-volume data for science, applications, and operations of all kinds. In particular is access to data in places that have challenges in their connection to the part of the global Internet where such data is generated and stored. This can be due to a lack of connectivity or because the data resources are “big” with respect to the connection capacity. Fast transfer of large files on demand can serve applications that are well connected, or a fast network can be used to replicate large data libraries in remote locations and serve those applications from local copies of those files. These approaches are the basis of the commercial strategy known as “Content Delivery” that is used by applications such as Netflix to distribute large collections to a very wide audience. The strategies developed for effective content delivery, and the framework for providing this capacity, need to be considered by NASA, NOAA, and other federal agencies providing remote sensing data to its user base.

# Why are these challenge/questions timely to address now especially with respect to readiness?

As data volumes increase and remote sensing data use becomes more ubiquitous, data access will be very important to address, particularly for users with limited connectivity. Fortunately, there are new ideas to help address this challenge. The confluence of the prospect of more and larger earth observation data along with new solutions to connectivity that go beyond extensive infrastructure development makes the challenge we have described a timely topic for the Decadal Survey.

In addressing the problem of how to serve applications with connectivity challenges, several approaches are possible. One approach is to upgrade the infrastructure that serves such applications, but this approach may be expensive and difficult to implement in some parts of the world. For example, one development strategy includes deploying balloons or solar-powered drones flying permanently in the stratosphere over Africa, an idea with grand challenges and expense.

An alternative approach takes advantage of the fact that high volume asynchronous communication can be accomplished over connectivity that is in many ways challenged, as long as the cumulative volume of data delivered is large enough and fast enough to meet application requirements. The term used for this approach is Logistical Networking (http://loci.cs.utk.edu/scidac/), and the central concept is to stage data delivered from the network to storage servers that are close to the application from the point of view of network topology. When data is replicated to locations that are well connected to the application network, data access can be high performance; from the point of view of the user, these data are fully accessible. An example of this concept is the Earth Observing Depot Network (EODN; http://data-logistics.org/?q=EODN), a content delivery network being developed by AmericaView and the NSF-funded Data Logistics Toolkit (DLT) project to distribute Landsat 8 and other data from satellite-based instruments. USGS currently generates Landsat 8 data on the order of 1TB each day. The scale of data access requirements can strain the capacity of the central repository at the USGS EROS Center. Once fully on-line, the EODN will be able to serve a purpose similar to that of commercial Content Delivery Networks in spreading and localizing this data-access traffic, providing the capability to support unlimited downloads at the highest available network speeds. One key point is that the EODN infrastructure is built and operated by the scientific end-user community, rather than by the data provider (USGS) or by a private sector service provider that is in the business of delivering and charging profit-making rates for access to content.

When data is staged locally at network locations that are connectivity challenged, the result is an “island” of content that is now available at very high performance within that locality. An additional advantage of locally staged data lies in the ability to control what data are moved to local servers. Controlling the flow of data so that it is available when and where it is needed is an important consideration, particularly in connectivity challenged environments. The plethora of data available today can be overwhelming, both for network infrastructure and analysts, particularly during periods of high stress such as those caused by manmade or natural disasters. The experience of users accessing content staged in this way on local servers will depend on the nature of backbone connectivity available. For example:

* A server that is regularly, but not continuously, connected at high bandwidth (for instance, by satellite connectivity) may be capable of large-scale scheduled updates to locally staged content.
* A server that is constantly connected by a very low bandwidth connection, possibly with a high error rate, might be able to support a large content collection that does not change too often.

The idea of using a local server to stage and serve content locally is not a new one, nor is Logistical Networking the only approach currently being offered. Some examples include:

* Delay Tolerant Networking (DTN; originally Interplanetary Internet) is a project that was proposed by JPL scientists as a means to avoid building a new data networking infrastructure for every space mission. It has been adapted in a DARPA project as Disruption Tolerant Networking for battlefield networking, and has also been proposed as a means of connecting remote areas.
* Outernet is a satellite-based Content Distribution project that proposes solar-powered servers to be placed in remote locations, creating a wireless Local Area Network to serve downloaded content.

A unique feature of the Logistical Networking architecture is that it is built on an infrastructure that is designed to be as *simple, generic, and limited* as possible. One of the key aspects of networking that has been provided by a common Internet infrastructure is interoperability. The more complex or specialized services described above have their place in the stack of protocols that can be delivered to connectivity-challenged locations; Logistical Networking provides a common basis on which they and many other possible services can be implemented. The idea is that these design goals are most likely to result in an infrastructure that can survive over time, meeting the goals of many user communities and be implemented in many ways as technology evolves. This approach is modeled after the most successful Internet infrastructure models of the past 50 years, namely the Internet and the Unix/Linux kernel interface, and defines a unique strategy for achieving global ubiquity.

# Why are space-based observations fundamental to addressing these challenges/questions?

For this final question, of course, space-based observations are not fundamental for addressing this challenge, but to turn this question around, access to space-based earth observation data fundamentally requires improved and progressive means to deliver timely data and data sets to the people and decision systems that use the data. Remote sensing users within U.S. Federal agencies, universities, K-12 educational entities, and resource management decision-makers need the NRC Decadal Survey team to consider the challenge of access to earth observation data, since it is the public sector that is financing and creating these data and who have the most at stake if they are not available. There is a concern by some in our community that free and open access to data developed, collected, and stored by government groups will become so difficult to access that we will be reliant on the private sector to use these valuable resources. Data stewardship is fundamental to good and open science, and it starts with free data access and effective dissemination. We urge the NRC Decadal Survey team to consider making data access a fundamental point of discussion so that our science and applications communities can rely on these data for their needs.