

**EarthCube TIVO:
Test, Integration, Verification Organization—Engineering EarthCube**

*Chaitan Baru, David Nadeau
Center for Large-scale Data Systems (CLDS)
San Diego Supercomputer Center*

Summary

Data resources in the geosciences are made available by data archives, e.g. NCAR, IRIS, UNAVCO, OpenTopography, USGS, NASA, NOAA; via community efforts, e.g. CUAHSI HIS, CZO, EarthChem, etc; or, via individual PIs. Each has its established and growing user community. The interoperability and interworkability objectives of EarthCube require not only easy access to these extant resources, but also easy access to all data, across these resources. A number of industry and community standards have been developed and have matured over the past one to two decades, providing a good starting point for system implementations. While each individual group/organization strives to produce data, metadata, and data access interfaces according to defined specifications, using many of the well-known standards, we recognize the need for a “third-party” entity whose task is to simply ensure that the mechanical aspects of data interoperability are in working order, and adhere to the published/stated specifications. We refer to such an activity as the *EarthCube Test, Integration, and Verification Organization*, or *EarthCube TIVO*.

Context

EarthCube TIVO is proposed based on the recognition that oftentimes interoperability and interworkability “does not come for free”. Specific effort is needed to ensure that these properties are satisfied. Individual projects, groups, PIs generally focus more on producing a specific dataset and associated tools which are of use to a particular domain or sub-domain in the geosciences. As mentioned, such efforts are increasingly using best practices in adopting well-known metadata and data standards, as well as services-based interfaces. However, these projects are (a) focused much more on their immediate community of users, as it should be and (b) the nature of funding is such that, the projects generally have barely sufficient resources to achieve their primary objectives, and do not have the luxury of considering system designs that allow broad interoperability, nor the resources for the programming efforts. Even if they build systems using standardized specifications, they certainly do not have the resources for test and verification of the interoperability aspects. The concept of an EarthCube TIVO is proposed specifically as a software engineering activity whose specific and only task is to test interoperability—according to the specifications set forth by the individual data providers, and by an EC TIVO Steering Committee (as mentioned below).

Interoperability in EarthCube

EarthCube TIVO is *not* a standards organization. It is a software engineering organization and a software competency center, steered by the community.

EarthCube presents an exciting vision for the future, but also represents a massive enterprise that needs to tackle ease of data integration at an unprecedented scale. EarthCube is not a single software system, but rather a virtual organization with data and software arising from many corners of the community. The key to making EarthCube work is to ensure the interoperability and interworkability among data and

tools produced by many different groups. For the purposes of this discussion, we frame “*interoperability*” as a technological concept: can the data cross between software boundaries? Can software reach across a distributed system and get to data? “*Interworkability*” is framed as a usability concept which brings in domain semantics. Can one scientist work with the data provided by another? Can these data be used meaningfully in conjunction with other data, to produce meaningful results?

The key to EarthCube is *not* the construction of a single software behemoth, nor a system that is based on a common software substrate, but rather the use of standardized interfaces and development of a variety of EarthCube “Connectors” (ECCs) that will allow for interoperability “in the small” thereby supporting interworkability in the large. Diversity will abound in the system, in terms of how individual archives/groups implement their systems internally, and should be encouraged. There will be many providers, consumers, and ECC creators. The trick to making all this work as an EarthCube system is to “trust but verify”, i.e. set simple, easy to understand interoperability requirements and request creators to adhere to these requirements, while verifying that data produced in part of the system can, indeed, be consumed at another point in the system. Trust and verification is achieved via actual software testing. If there are incompatibilities or other problems they could be (a) reported to the providers with suggested changes, (b) where possible, fixed by EC TIVO, or (c) simply noted and left for future action.

The figure at the end of this paper shows the EarthCube ecosystem, which includes data providers and data consumers, and the role of EC TIVO. As mentioned, data providers in the geosciences already strive to employ standards across the board with the objective of ensuring greater use of their data. Each provider already has their immediate user community. EarthCube brings the larger vision of interoperability across such resources, thereby improving the situation for current users and perhaps expanding the community of users. The EC TIVO would assist with testing the interoperability. EC TIVO is not an EarthCube standards organization, but simply the “keeper” of community generated best practices, and a testing organization that checks whether different components can be integrated. It may be able to produce a “state of interoperability” report.

Engineering EarthCube

The past 1 to 1.5 decades has been about the development of a broad range of data-oriented standards—for metadata, data formats, data access services, and even some semantic information. The next decade or so of EarthCube and other initiative should be about adoption and effective use of standards. This requires an engineering approach, as described here.

As mentioned earlier, ensuring interoperability is, at one level, an engineering task. At another level, it is a domain-oriented task where scientists and experts need to ultimately comment about whether certain “mashups” of data are meaningful or not. Here, we focus on the engineering aspect first, since that is a prerequisite to the higher order integration. Testing in software engineering is characterized by concepts such as *unit and function tests* and *integration tests*. *Verification* is a concept that is captured by the notion of *regression tests*, where one ensures that previous tests are still valid and that they still produce valid results (i.e. the system has not *regressed* to an earlier state).

Unit/Function Tests

In the context of the EarthCube TIVO, Unit or Function Tests would be the tests that are performed using individual data resources, to ensure that data are accessible and parsable from individual sources, using the specification provided by that source. The

EarthCube TIVO Steering Committee, consisting of scientists representing and reflecting the diversity of the broader community, would decide on priorities for which sources need to be tested, and in what ways. In other words, they would assist in the development of the unit tests.

Integration Tests

Integration tests would be based on “integration scenarios”, also developed by the EC TIVO Steering Committee, which would be based on plausible interoperability and interworkability scenarios, where a user may wish to access data from more than one source and perform a certain set of operations on the retrieved data.

Verification Tests

Both Unit/Function tests and Integration tests would accumulate over time to form a suite of verification tests. These are tests that could be invoked at any time—or periodically—to ensure that the entire system is still working as expected, and that no regression has occurred. It is entirely possible and likely that all tests will evolve with time, as the individual systems and their capabilities change.

Conclusion

We are unable to imagine how the vision of EarthCube could be implemented in actuality without the presence of an activity like the proposed EC TIVO. Interoperability is a difficult proposition—both in terms of the software systems and in terms of the domain science. Software systems need to be engineered to support interoperability, and engineered systems need to be tested. EC TIVO provides that mechanism.

