EarthCube Design Approach

Hannes E. Leetaru, Illinois State Geological Survey, University of Illinois
Michael Welge, National Center For Supercomputing Applications, University of Illinois
Bernie Ács, National Center For Supercomputing Applications, University of Illinois
Shaowen Wang, Department of Geography, University of Illinois
Yu-Feng Forrest Lin, Illinois State Water Survey, University of Illinois
Kenton McHenry, National Center For Supercomputing Applications, University of Illinois

VISION FOR EarthCube

A description of the envisioned scope of EarthCube and how will it transform geosciences research, including the functionality EarthCube can provide to the whole Geosciences community.

The vision of EarthCube presented here is an open access archive of all publicly available geosciences data in the United States and eventually the world, capable of supporting the technical needs to store and offer easy access to petabyte-size datasets in the next decade, offering long-term format translation, rich provenance and metadata, advanced cross-disciplinary portals that can be tailored for use from the citizen scientist to graduate students to expert users, offers community features to facilitate and promote collaboration, and has strong governance and sustainability models to ensure long-term success. A hybrid architecture of a centralized data system with multiple distributed nodes for legacy data and specialized portals combines ease-of-integration of distributed data with the centralized needs of governance. There will be a common template for domain portals that will enable transparent access for the new user and still enable experienced users to use advanced features unique to their disciplines. EarthCube development will use an iterative approach with rapid development of prototypes used to test technical feasibility and suitability to science needs.

Important, but easily forgotten, is the need for easy access and analysis capabilities for the "Citizen Scientist" and K-12 students to leverage the same resources as domain scientists, but through interfaces and curriculums tailored for their use. These non-scientists have recently made significant contributions to science that would otherwise not happened.

EarthCube will offer unified search for across attributes, geographic region, specific methodologies, provenance, or earth characteristics. Ultimately, we envision all geoscience projects funded by the US government and nongovernmental organizations including USDOE, USEPA, corporations, and USGS covering non-classified data would integrate their data into EarthCube. Finally, to encourage initial adoption and to maximize the coverage of the system, connectors will support the integration of numerous online data environments, as well as integration with NSF XSEDE to allow for seamless application of HPC and service oriented approaches to synthesizing data and associated analytics.

The Illinois State Geological Survey (ISGS), the largest state research survey for the geosciences in the United States, will partner with the National Center for Supercomputing Applications (NCSA), which has a long history of housing national scientific data and computational infrastructures, including the first engineering grid, NEESGrid, along with two of the largest forthcoming astronomical data projects: 1) Dark Energy Survey (DES)[1] and Large Synoptic Survey Telescope (LSST)[2] The NSF-funded CyberGIS software integration framework[3], as well as the University of Illinois' #1-ranked Library and Information Science program in the country, with its long history of work on provenance, data curation, and other related topics will partner together to leverage their combined expertise for EarthCube.

COMMUNITY-BASED GOVERNANCE MODEL

The community structures necessary to acquire current and future user input/requirements, to respond to changing data and science needs, to adapt and adopt new technologies, to coordinate components and facilities, to foster partnerships and community participation.

There are three parallel relationships that must be managed in a large cross-disciplinary endeavor such as EarthCube: science/science and science/technology. At the highest level, EarthCube consists of two interrelated systems: a collection of science goals and an assortment of hardware and software technology infrastructure to support those goals. One of the lessons learned from the NSF NEESGrid project[4] was the reinforcement that a successful effort must involve strong collaboration between both domain scientists and technologists, but that the driving vision must be led by the scientists, with the technologists helping to shape that vision to be computationally tractable. The dangers of a technology-driven approach are seen in the caBIG© program[5], where costly technologies were developed without the ongoing input of the scientific community (in this case cancer researchers), resulting in a collection of tools that ultimately did not find widespread adoption.

Multidisciplinary projects face the prospect of competing scientific goals or demands, requiring strong scientific leadership and governance processes. The proposed EarthCube governance model would be patterned off the model of several large scientific projects, such as LSST, and would consist of an executive committee with a geoscientist as full-time chairperson to set overall coordinative vision, a technologist as full-time associate director, and a full-time project manager to oversee day-to-day operations. Given the wide range of disciplines represented and the lack of any one single "end user," a scientific steering committee will provide the primary scientific input, consisting of representatives from each of the geoscience domains. A technical steering committee will provide similar input on cyberinfrastructure. The research communities served by EarthCube are extremely diverse in their needs and these needs will change over time, requiring ongoing regular meetings of the committees to evaluate the impact of those changes on the scientific goals and cyberinfrastructure needs of the project. Capturing the combined insight of each discipline in a scientific steering committee ensures all needs are heard, while the use of a full-time executive committee ensures continuity and accountability for project schedules. All proposed changes or additions to EarthCube will be reviewed by the steering committees for approval.

In any large community-based initiative, especially where there are multiple domains potentially making competing demands, there must be a central master framework to allow for a quantitative uniform decision-making process. Following the LSST project's framework, the master consensus science goals for EarthCube as set forth by the steering committee will be used to construct a set of "sizing guides" that capture, within the infrastructure design, the actual dollar cost of any individual hardware or software component. A range of guides will cover everything from computing, storage, and networking, to data security. If a given science community requests the addition of a new feature, these sizing guides will be used to determine the actual dollar cost of adding that new capability, which can be evaluated against the core science goals and the availability of funding sources to cover the immediate and ongoing costs. A robust "sizing model" allows the true cost of every component of the system to be determined and what it would take to scale the system up or down as dictated by changing funding climates over time. Finally, the governance plans provide concrete criteria through which to evaluate every component of the system holistically to determine the order of development and deployment efforts, manage risk, and ensure that integration efforts meet total project objectives and schedule. Such formal project management documents, while seemingly burdensome compared with the more

fluid process typically used in single-investigator projects, is absolutely critical when coordinating a large distributed group consisting of diverse expertise and geographically remote teams and individuals. A cost analysis needs to be completed on any development of open-source tools when commercial tools are more mature architecturally and easier to use.

Another key finding of the NEESGrid project was that all software development in a scientific enterprise must be iterative. This is in contrast to traditional large software development models where all user requirements are defined at the beginning of the project, with delivery at the end of a finished product. Successful scientific projects must instead incorporate a far more iterative design process revolving around the use of rapidly-evolving prototypes. The system should also be developed in modular phases rather than a single monolithic development cycle, to allow for continual community feedback to shape the outcome.

A significant finding during the evaluation of the caBIG© project was the need for strong consensus policy on data sharing and an improved workflow for the creation of metadata. Researchers would often withhold portions of their data they were still publishing on, suggesting the need for an embargo capability on EarthCube to encourage more complete data submissions. Metadata was also found to be problematic in that most researchers wait until the end of a project and generate metadata only to meet contractual obligations[5]. The result of the temporal gap between data creation and preservation is a loss of provenance and metadata accuracy. Finally, each dataset must include metadata that references all resulting publications and presentations (including posters and PowerPoints) based on the data or method and any other key documentation such as laboratory manuals or field notes that would enable the user to understand the data and/or replicate findings based on it.

Metadata and provenance are especially critical when working across disciplines. For example, seismic reflection data is collected in time, not depth, and is later converted to depth through "time to depth" algorithms that take into effect the lithology, density, velocity, and consistency of the formations. These assumptions must be included in the provenance history.

CONCEPTUAL CI ARCHITECTURE

The architecture necessary to provide the services of EarthCube, to integrate advanced information technologies that facilitate access to distributed resources such as computational tools and services, instruments, data, and people.

There are two significant models for data storage: distributed data (also referred to as the federated model) and centralized data. Both of these data models have significant benefits. The data distributed model is initially the least costly option and can be implemented quickly. In a distributed model the data is physically distributed across multiple sites (nodes) that each specialize on one of the geoscience domains, while a smaller number of centralized nodes contain the metadata for searching. One of the benefits of distributed system model is that sites can each specialize in a single science domain and new resources can be added quickly. However, this ease of addition also makes it more vulnerable to data loss: the world wide web succeeded because of the ease of adding new web servers to the network, but the majority of the content posted to the web since its creation has subsequently been lost. To ensure long-term data availability, stronger governance and data replication facilities are needed. In addition, the movement of large datasets across the network can be problematic.

In the short term the distributed data model is probably the least expensive and most rapidly implemented because one can use the current servers and distribution centers that are already in place.

In the longer term it is a more expensive option because of the need to keep multiple nodes active. The data distributed model would need a significantly stronger monitoring system because each data node over time is likely to drift from the planned governance model for data archiving. In addition the distributed system could lead to software duplication (especially middleware applications). Many data are similar enough that modified middleware could be used for diverse types of data.

A centralized EarthCube model would be similar to the GenBank model where all of the data are archived in three large centralized data nodes across the world. One of the GenBank data centers is in the United States, another in Japan, and the third one in Europe. LSST also has two centralized data nodes (one at the telescope array and one at NCSA). Similarly, under this model, EarthCube would need at least two or more data centers in different portions of the country that would mirror each other.

EarthCube would have to start as a hybrid model because of the need to integrate all of the legacy data centers. It is anticipated that it would take years to include all of the data from the current distributed nodes in a central database. A set of centralized servers would provide data storage and computing resources, while remote nodes would allow for specialized portals and legacy data. While ultimately all of this data would be encouraged to be housed in the main system, a connector will be built to allow the central node to access these remote datasets.

The largest projects, such as LSST, have reached a point where data has become too large to be moved, and so computation on the largest datasets will have to move to where the data is. Based on the LSST model, ultimately a local compute cluster colocated with the data will allow allocated computation directly across the largest datasets.

CI design will include a Common Data Interface (CDI) against which discipline-specific portals can access and publish data. The primary focus will be on offering a range of data ingest and delivery mechanisms, normalization, long-term format translation and preservation, data protection, provenance, and enhanced metadata. These capabilities will be leveraged for global search and discovery, ranging from selecting a polygonal shape on a map and requesting all datasets covering that region, to looking up a common dataset identifier from a publication and seeing all other datasets generated using that particular processing technique.

CDI will allow a wide range of tailored community portals to be built around this common backend infrastructure, including XSEDE science gateways. Other services like GISolve will be provided to allow new capabilities, including computation, visualization, analysis, annotation, and other capabilities to be housed by the system. These will be modular services, sharable across all disciplines, and allow each domain to group them together and form customized workflows and portals matching their distinct needs from the same set of core building blocks.

Initially the system will be a depository where data is compiled by researchers at their own sites and then ultimately deposited with the system, but this generate-metadata-at-the-end approach severely limits the accuracy and availability of metadata, so EarthCube requires interfaces where researchers can use it as their *primary* storage grid. Ultimately, support for realtime ingest where sensor networks can write directly into this storage architecture is required. All of this requires a careful governance plan to understand the costs involved with the highest-volume projects. For example, likely access patterns include high-compute on XSEDE infrastructure on a single large dataset, fully cloud-based analysis of a dataset with no data downloaded other than visualizations, downloading small datasets to local

desktop, and full-scale compute across massive datasets using local compute infrastructure provided by EarthCube.

To better preserve data files and ensure data will be available and accessible going forward into a future where file formats are a moving target, data will be translated into a common format where possible. In this way, risks involving accessibility can be mitigated, and data integrity and availability can be ensured. From this normalized format, translations into popular "formats de jour" will be supported so data may be delivered to clients, or client software, in the form most palatable given their needs.

Ensuring that digital files are preserved on tape or disk is only one aspect of digital preservation. The other, far more difficult aspect is in ensuring that the saved files are still viewable, editable, and in general accessible as time goes on. We have documented over 140 different file formats for just 3D content, a number we believe to be only a small fraction of the actual number of 3D formats available. One solution might be to integrate the NCSA Polyglot service, which offers transparent format conversion, including calculations of translation data loss.

Portal Design

Human interface design is probably the most critical feature of any data retrieval and analysis system, having a significant impact on the ease with which data is accessed. A review of existing portals for many geoscience database nodes shows a wide variability in user interface designs. Prototyping would be used to construct a general purpose portal that could be used by all of the distributed notes. It would provide a general portal template ensuring consistent familiarity, while allowing specific tailoring for each discipline's individual needs.

EarthCube must include facilities to reference all publications using a given dataset. This implies technology and policy functionality to encourage continual updating of the system or integration with external bibliographic services. The data will be invaluable for future groups to find out what worked and what didn't work and what other information can be gained from these data. Additionally, the linkage of diverse data sets in a single project is a critical, yet under-discussed capability. Large site evaluations such as those that are occurring on the United States Department of Energy carbon sequestration partnerships involve data from many sub-disciplines in the geosciences. For example, the same site could have micro seismic data, seismic reflection data, groundwater movement data, LiDAR data, multiple types of geochemical data, and geologic mapping data. These would be commonly archived in the different sub-disciplines but must have a linkage to the main reference project. It is critical that all publications that reference the original data must also have their providence referenced with the original archive via a reference identifier.

Users must be able to immediately and easily access data and understand their geographic and physical characteristics. Portals, like all software, tend to become more complicated as they incorporate greater functionality over time. A tiered interface system will be necessary to ensure that K-12 students, citizen scientists, graduate students, new researchers, and expert users with advanced needs, can all access the same common infrastructure, but do so through interfaces tailored for their respective complexity levels. Underlying all of the portals is a common unified middleware backend accessing a centralized dataset that has standard geographic coordinate system, including data projection, and must be searchable by geography, project, and subject. Specialized portals would have a common template but would be specifically designed for the needs of each geoscience discipline.

Data Transfer

The movement of data across the network may be the most significant difficulty in designing EarthCube. There are currently few solutions for moving large data sets of even 50 GB across the network outside of the high-speed networks connecting major research institutions. This implies that a central database may be the only method of computing available for large data sets. LSST is a good example of the difficulties and infrastructure needs for moving large amounts of data.

DESIGN PROCESS

User requirement-driven design methodology, identification of design team members, qualifications of development team, time-line for design demonstration and scale-up, design tools and practices that create robust, sustainable, well-documented and open source infrastructure

It is impossible to separate the governance structure from technology in developing a cyberenvironment. A Requirements Traceability Matrix (RTM) must be used to tune the original scientific goals to ensure they are tractable and to quickly illustrate the cost of any changes in the goals in terms of time, effort, and funding. The RTM can also be used to quickly assess the impact of scheduling delays in one component of the project on other sections.

Combined, the Illinois State Geological Survey and Illinois State Water Survey have ongoing projects touching on four of the six major disciplinary data classes identified in the EarthCube call and have a long history of serving as the primary geospatial data clearinghouse for the State of Illinois, including climate data. In this way, the two Surveys offer a model in miniature of the broader disciplinary needs, challenges, and opportunities of EarthCube, yet with the benefit of substantially easier interaction with disciplinary experts given the close physical proximity with the design team. Thus, the first phase of project design would make use of these organizations as small-scale testbeds, and later phases can test larger numbers of more iterative prototypes on these groups before transitioning to the national user community.

The design process will be based around increasingly-larger pilot or prototype projects conducted to understand and advance specific infrastructure needs and their capabilities. Prototyping must scale from relatively simple systems to more complex systems during the design phase. The design phase will look at the entire lifecycle of the project with requirements for gathering and analysis, design of the software, development, software documentation, systems testing, software release strategy, and deployment and support strategies.

The design team organization includes a project director who will play a critical role in making decisions regarding important issues such as priorities and distributions of resources to meet the project goals. The project director needs to have a strong background in the geosciences to better communicate with the stakeholders and must have strong support from the development team. The project director will share much of the decision-making processes with the associate director who must have a strong background in the development of cyberinfrastructure.

There must be a technical project director who is experienced in large software development and can provide leadership and guidance on IT issues. There should also preferably be a project manager who can oversee the day-to-day activities and responsibilities of the team members. There also must be lead investigators for each discipline that can discuss technical issues and meet on weekly basis to discuss project status.

Design Team

The design team will be composed of staff from the National Center for Supercomputing Applications (NCSA), Illinois State Geological Survey (ISGS), Illinois State Water Survey (ISWS), University of Illinois Geography Department, and the Graduate School of Library and Information Sciences (GSLIS). This team has extensive experience with large multi-petabyte hosted datasets such as the Large Synoptic Survey Telescope (LSST), the development of robust enterprise data management platforms, data curation, provenance and metadata, long-term file archival and format translation loss, data visualization and interaction, and outreach and training. NCSA is also the lead for the NSF-funded Extreme Science and Engineering Discovery Environment (XSEDE). XSEDE is already being used to enhance earthquake research with improved tools for mapping the movement of the Earth's mantle. The ISGS is the largest state geological survey in the United States devoted only to geoscience research, with a staff of 200 scientists and technical support staff. The University of Illinois GSLIS is the top-ranked department in the country in library and information science education, research, and practice and has considerable experience modeling user requirements, including observing usage of a system and capturing user requirements in actual usage.

Hannes E. Leetaru, Senior Geologist at the Illinois State Geological Survey is the Geology and Geophysical Coordinator of the USDOE funded Illinois Decatur Project that is testing the concept of CCS (Carbon Capture and Storage).

Michael Welge, Senior Research Scientist, NCSA, leads a research team specializing in data-intensive technologies and applications including visualization and data mining.

Bernie Ács, Informatics System Designer and Database Architect, NCSA, leads research teams in developing database systems in the arena of constructing, customizing and implementing computer systems in personal, professional, and industrial environments.

Shaowen Wang, Associate Director for CyberGIS and Senior Research Scientist at NSCA; Directs the CyberInfrastructure and Geospatial Information (CIGI) Laboratory that researches and develops cutting-edge cyberinfrastructure to advance GIS and geospatial problem solving and decision making, including the NSF CyberGIS project and open source SimpleGrid Toolkit.

Carole L. Palmer, Professor and Director of the GSLIS Center for Informatics Research in Science and Scholarship (CIRSS). The center conducts research on information problems that impact scientific and scholarly inquiry. Projects and activities focus on how digital information can advance the work of scientists and scholars, the curation of research data, and the integration of information within and across disciplines and research communities.

Yu-Feng Forrest Lin, Hydrogeologist and Director of ESRI-GIS Development Center at the University of Illinois at Urbana-Champaign has ongoing research in 3D visualization applications in hydrogeology.

Kenton McHenry, Research Scientist, NCSA, leads a team focusing on Digital Curation and Computer Vision Advanced Computing and is currently working with the United States National Archives (NARA) on scalable file migration and long-term data archival.

Michelle Butler, Technical Program Manager, NCSA, leads the storage systems group at NCSA. The group manages most of the storage systems at NCSA that encompass multiple petabytes of data and help design new architectures for petabyte data motion on its HPC systems.

OPERATIONS AND SUSTAINABILITY MODEL

Operational aspects of a community-wide enterprise that address such activities as centralized functions, coordination of services, user services, including training, and identification of what it will take to sustain a viable infrastructure over a long periods of time and who will carry out these functions.

The full-time executive committee will provide the day-to-day high-level operational support of EarthCube. As the lead site for XSEDE, NCSA already has a 24/7 helpdesk and operations center that could be leveraged to provide technical infrastructure monitoring and error remediation, including afterhours support. The scientific advisory board will be used to provide ongoing input into emerging trends in each discipline, especially around the growth rates of new datasets and likely end-user load on those datasets. Load demand is especially important with respect to hardware overprovisioning: to sustain projected user demands, LSST will ultimately have 100PB of disk housing four mirrored copies of the 25PB of actual data in order to provide the necessary access bandwidth. Unlike LSST, where the data sizes and growth rates are known at the start of the project, the advisory board will help with monitoring for emerging trends and ensuring that the hardware refresh cycle adds sufficient capacity by the time the need emerges, leveraging exponential hardware improvement rates.

Funding Sustainability

The long-term cost of maintaining and operating large data infrastructures such as EarthCube can be considerable and play a significant role in sustainability. For example, it is estimated that the total cost to archive the CERN collider data will be $90M, or 1% of the instrument's entire budget.

Operational cost is the largest factor affecting the sustainability of EarthCube and it is critical that there be continuous funding to support ongoing infrastructure costs. It is unlikely that private sector funding sources would be sufficient to support its continued operation and federal support is likely to play a critical role. This would not be limited to support from NSF, but would include other federal agencies wishing to deposit research data in EarthCube. In addition, a hybrid fee structure similar to that used by the Inter-University Consortium for Political and Social Research (ICPSR) could be assessed, which charges a membership fee starting at $15,000 a year for large institutions and decreasing in fee structure for smaller Masters and Undergraduate institutes. Under the ICPSR model, institutions must be paying members to be able to submit or request data from the archive. In addition, this fixed flat fee assumes that datasets will be reasonably small (social science datasets tend to be measured in kilobytes to megabytes), while EarthCube data will likely be in the gigabyte to terabyte range. Thus, while such a fixed fee could generate some additional revenue, it would be unlikely to cover a substantial portion of operating costs.

An alternative model is that taken by the HathiTrust, which charges a storage-based fee of $4/GB per year for data archived there. This is very similar to the cloud storage model used by Amazon and other companies. However, given that HathiTrust is just launching, it is unclear how this model would cope with institutions losing funding to participate or reducing their funding commitment. In those cases, would data be removed from the archive? Many universities are deploying institutional data repositories to comply with increasing requirements for data preservation and depositing. Yet, the

majority of these are little more than file servers, storing uploaded ZIP files in binary stasis. The added benefits of the EarthCube repository together with NSF best practices for depositing NSF-funded research into the system could potentially make it possible to tap into these funding streams for additional institutional support of EarthCube from contributing institutions.

Education and Training

NCSA and ISGS have considerable experience in outreach and training to non-traditional computing disciplines. National and regional workshops and the XSEDE "campus champions" model can all be used to increase awareness and participation in the system. Yet, training and workshops can only go so far: to achieve truly institutionalized success, one must provide the tools for the communities themselves to develop their own sustainability. We need to develop new infrastructure revolving around knowledge management. For example, systems like GISolve permit customized workflows combining multiple tools and specific configurations to be saved as workflows that can be shared with other users. Thus, instead of uploading only the raw dataset used for a publication, the computational workflows, customized visualization processes, and all other steps in the production of that publication can be shared.

Bulletin boards, community profiles, recommender services, and find-an-expert services will all be incorporated into the interface components of the system, encouraging and facilitating collaboration. In particular, the assemblage of data, tools, and workflows in a single location, and the ability to replicate a new paper by applying the same tools in the same workflows to the same data, will offer tremendous capacity for encouraging graduate students and new researchers to experiment with those techniques. In particular, the best learning happens through hands-on experimentation, and so by dramatically lowering the barrier to testing a new technique, a graduate student, for example, can apply a cutting-edge new analytical technique published in a paper using one dataset, and apply that to a range of other datasets in an area of interest to that student, accelerating the uptake of new techniques considerably.

The role of the "Citizen Scientist" in EarthCube has not been heavily discussed, but there have been numerous documented examples where their role has been critical to the success of research initiatives. EarthCube must have the capability of adding applications or portals that enable the Citizen Scientist to contribute to geoscience research. Projects need to provide fun, a sense of community, and the ability to contribute to science. The Citizen Scientist in particular provides a critical resource for visual pattern recognition that is difficult to achieve through purely computational means[7]. We agree with Raddick's vision of having the citizen scientist as a collaborator that analyzes data, compares simulations to observations, and might even collect data.

The same software could be used as part of a K-12 education program. We must find ways of including our youngest and most enthusiastic members of society into the thrill of science. Elementary grades would likely be reliant on teaching materials that could be integrated into the classroom. However, middle and high school students are commonly required to take a course in either earth science or environmental science and curriculums could allow direct integration of EarthCube into classroom use, much as the NSF-funded Institute for Chemistry Literacy through Computational Science (ICLCS) has integrated computational chemistry into K-12 classrooms. For students and Citizen Scientists to maximize their success, participation in a community enabled by networking tools such as forums and blogs will be critical.

Finally, EarthCube must have an outreach program for non-scientists to review and collaborate on research. The strongest supporters for a program are those that feel a program is part of their

community. EarthCube will contain core data that describes the climate, earth processes, and anthropogenic environmental impacts. Ideally, the non-scientific community of teachers, students, and citizens can learn how to study the data and how research scientist came to their conclusions

References

[1] http://www.darkenergysurvey.org/

[2] http://www.lsst.org/lsst/scibook

[3] http://cybergis.org

[4] Spencer.B.F., R.Butler, K. Ricker, D. Marcusiu, T. Finholt, I. Foster, C. Kesselman, 2006, cyberenvironment project management: Lessons Learned: www.nsf.gov/od/oci/CPMLL.pdf

[5] Board of Scientific Advisors, 2011, An assessment of the impact of the NCI cancer biomedical informatics grid; http://deainfo.nci.nih.gov/advisory/bsa/bsa0311/caBIGfinalReport.pdf

[6] Fleischer, D., and K. Jannaschk, 2011, A path to filled archives: Nature Geoscience, V. 4, p. 575-576

[7] Raddick, M.J., G. Bracey, K. Carney, G. Gyuk, K. Borne, J. Wallin, S. Jacoby, 2011, Citizen science: status and research directions for the coming decade: http://www8.nationalacademies.org/astro2010/DetailFileDisplay.aspx?id=454