4-16-2012

GEOSHARE: Geospatial Open Source Hosting of Agriculture, Resource and Environmental Data for Discovery and Decision Making

Thomas Hertel
Purdue University, hertel@purdue.edu

Nelson Villoria
Purdue University, nvillori@purdue.edu

April 16, 2012

GEOSHARE: Geospatial Open Source Hosting of Agriculture, Resource & Environmental Data for Discovery and Decision Making[1]

Authored by:
Thomas Hertel, Purdue University
Nelson Villoria, Purdue University

With input from:
Stanley Wood, IFPRI
Wolfgang Britz, University of Bonn
Glenn Hyman, CIAT
Noah Diffenbaugh, Stanford University
Andrew Nelson, IRRI
Navin Ramankutty McGill University
Stefan Siebert, University of Bonn

[1] The authors acknowledge the support of the Foresight Programme from the UK Secretary of State for Business, Innovation and Skills. A longer report is available at http://www.agecon.purdue.edu/foresight/. The authors also acknowledge support from Purdue University's Global Policy Research Institute as well as the Purdue Climate Change Research Center in developing this proposal. Funding for the GEOSHARE pilot project is being provided by the UK Department for International Development, the Economic Research Service of the USDA and the UK Department for Environment, Food and Rural Affairs.

Motivation for GEOSHARE

Feeding 9 billion people in the face of a changing climate, while preserving the environment and eliminating extreme poverty, is one of the grand challenges facing the world as we look forward over the coming decades. Agriculture and land use change account for roughly a quarter of global Greenhouse Gas emissions[2]; land-based activities are arguably the most sensitive to climate change, and farming remains the predominant source of income for the world's poorest households. Yet, the data currently available to understand how global and local phenomena affect the agriculture-environment-poverty nexus are insufficient to advance discovery and promote effective decision making. This has led the Commission on Sustainable Agriculture and Climate Change, chaired by Sir John Beddington, Chief Science Advisor to the UK government, to argue for the creation of "comprehensive, shared, integrated information systems that encompass human and ecological dimensions" as one of the key recommendations in its recently launched report.

The geospatial data needed to address these issues includes information on land cover (e.g., cropland, pastures, forests), which are typically obtained via satellite imaging, as well as land use (e.g., crop-specific harvested area and yields) which require census-type data from administrative units (e.g., counties or districts). In addition, other data are important to inform research and decision making on these issues, including soils, crop calendars, irrigation, input use, as well as poverty rates. Since these data come from many sources and are processed in many different ways, a common set of standards are necessary to ensure inter-operability. In a review of state-of-the-art datasets, in response to a request by the UK Office of the Chief Scientist in the context of the FORESIGHT study on long run food security and agriculture[3] we identified a wealth of individual data sets on agriculture, but noted that these tend to be regional or national in scope, not compatible with one another, and often not publicly available. Many of those that are available were found to be technically challenging to researchers with limited resources. Where global data sets do exist for specific attributes, for example, harvested area, yields, and irrigated area, they are not developed in a way that facilitates inter-operability.

This lack of time series, mutually compatible, geospatial data at global scale has greatly inhibited the ability of scientists, practitioners and policy makers to address the socio-economic and environmental impacts of contemporary policy issues related to poverty reduction and the long run sustainability of the world food system. An accurate assessment of these policies, as well as effective resource allocation to solve development problems, requires knowledge of local conditions; however, at the same time, these local decisions are being made within an international context, for which global analysis is required to capture the drivers of change as well as to avoid misleading conclusions. GEOSHARE aims to fill this gap.

---

[2] http://www.wri.org/chart/world-greenhouse-gas-emissions-2005

[3] Hertel, T., W. Britz, N. Diffenbaugh, N. Ramankutty, N. Villoria, (With additional contributions from S. Wood, S. Siebert, G. Hyman, and A. Nelson), 2010. "A Global, Spatially Explicit, Open-Source, Data Base for Analysis of Agriculture, Forestry, and the Environment: Proposal and Institutional Considerations." Available at http://www.agecon.purdue.edu/foresight/.

Mission and Vision:

GEOSHARE's *mission* is to develop and maintain a freely available, global, spatially explicit database on agriculture, land use, and the environment accompanied by analysis tools and training programs for new scientists, decision makers, and development practitioners.

GEOSHARE's *vision* is that of a vibrant global network contributing to this shared infrastructure, enhancing capacity for analysis in developing countries, and applying these geospatial tools to guide decision making related to food security, land use, environmental sustainability and poverty reduction.

GEOSHARE will achieve its mission and vision by fulfilling three key objectives:
Objective 1: To provide a globally consistent, temporally opportune, and locally relevant database for better decision making. GEOSHARE's partners have a strong record creating both global and local spatial datasets widely used for discovery and decision making. Funding for these partners, obtained through GEOSHARE, will allow the development of the long-term relationships required to produce and update data under common standards for interoperability. These regional nodes are: the *International Food Policy Research Institute* (IFPRI), particularly through its activities in sub-Saharan Africa led by Node Director Dr. Stanley Wood; the *International Rice Research Institute* (IRRI, focusing on Asia), led by Node Director Dr. Andrew Nelson; and the *International Center for Tropical Agriculture* (CIAT), based in Cali, Colombia and focusing its efforts for this project on Latin America through Node Director Dr. Glenn Hyman. The data captured by the regional nodes will enrich the global data provided by the global research nodes. In the initial phase, the global nodes include three institutions: *Stanford University* (past and future climate) with Node Director Dr. Noah Diffenbaugh, *McGill University* (agricultural productivity, land cover and use) with Node Director Dr. Navin Ramankutty, and *University of Bonn* (irrigation, agricultural production and water use) with Node Director Dr. Stefan Siebert. The contents of GEOSHARE will be delivered through a HUBzero-based platform (more details below), making remote processing of data, simulation-based analysis and capacity building tools available to any individual with an Internet connection, anywhere in the world.

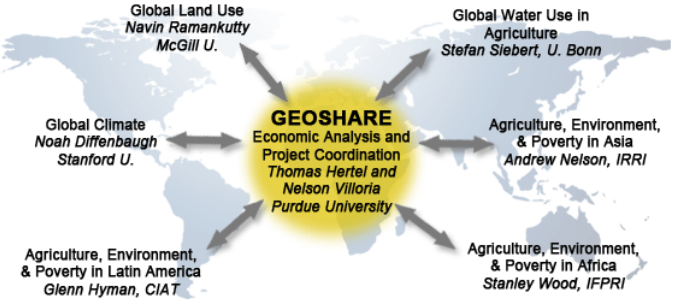

Figure 1: GEOSHARE will generate a spatially explicit database on agriculture, resources, and the environment utilizing a scalable nodal structure that matches data supply capacities with data needs.

The nodal structure of GEOSHARE is readily scalable. By involving top scientific and analysis teams from around the world, the burden of funding this effort will be globally shared, thereby leveraging sponsors' investments and increasing GEOSHARE's sustainability. To ensure that developments in GEOSHARE address local problems and policy issues, GEOSHARE's regional research nodes will interact with national and local stakeholders in the government and academic communities in their regional areas of influence.

Objective 2: To assist decision makers, policy analysts and researchers seeking to use geospatial data and analysis tools to inform activities relating to agriculture, poverty, land use and the environment. One of the key trends shaping the development landscape is that of global interconnectedness[4]. GEOSHARE capitalizes on this trend by increasing stakeholders' access to data sharing and analysis tools through remote computing via HUBzero technology, which provides interactive simulation and analysis tools via a Web browser, requiring no downloading of data or software (see Box 1). Through HUBzero, users can share their expertise by exchanging computer scripts and software ranging from simple spreadsheets to sophisticated simulation models such as Pegasus[5]. This activity takes place 'in the cloud', powered by Purdue University's high-capacity computing grids, without the need for specialized software or hardware at the user's end. HUBzero opens endless collaboration possibilities. For example, a researcher without Geographic Information Systems (GIS) expertise, sophisticated hardware, or expensive software, interested in productivity-weighted average precipitation during the growing season for a given zone in Ethiopia, could reuse computer scripts written by a skilled user that created the same variables for Iowa, or the Bolivian highlands.

Objective 3: To build capacity throughout the world in individuals who can effectively bridge disciplines to make decisions and to identify solutions to complex resource use and development problems using geo-spatial data and analysis tools. The GEOSHARE framework is designed to engage a broad, diverse learning community ranging from decision makers and seasoned practitioners to undergraduate students and budding scientists. Our overall outreach strategy is to advance knowledge and catalyze new approaches for development and resource use through an innovative cyber-infrastructure coupled with personal connections and networking. GEOSHARE aims to support graduate students undertaking training and research at partner institutions. It will also provide content for interdisciplinary courses at the undergraduate and graduate levels. As the GEOSHARE user community grows, we anticipate developing webinars, podcasts, online training activities, and other learning tools with direct input from the user/stakeholder community. Beginning in Year 3, GEOSHARE plans to initiate an annual global development challenge program for undergraduate students. Through the program, student teams will be given the opportunity to work with GEOSHARE mentors to apply their creativity to address a problem facing decision makers in developing countries by competing for funding that will allow them to cultivate their proposed solution, field-test their methodology in conjunction with one of the regional nodes, and measure the impact of their work. GEOSHARE's open-source data platform lends itself well to innovation and will actively encourage new ideas to improve development objectives by establishing a second challenge program topically aligned with stakeholder needs and sponsors' strategic goals, focusing on accelerating innovation through public- private sector partnerships.

[4] USAID Policy Framework - 2011-2015. Retrieved from www.usaid.gov/policy/policyframework_sep11.html.

[5] Deryng, D., W.J. Sacks, C.C.Barford, and N. Ramankutty, 2011. "Simulating the effects of climate and agricultural management practices on global crop yield." *Global Biogeochemical Cycles*, 25(GB2006), 18 pp.

BOX 1. HUBZERO TECHNOLOGY: NETWORKING IN THE CLOUDS

The cyberinfrastructure for this project will be built on the HUBzero® Platform for Scientific Collaboration.[6] A hub combines unique middleware with Web 2.0 functionality, providing a platform that is much more powerful than an ordinary website. Users are not only able to network and share information, but they can also create, publish and access interactive visualization tools powered by a computer cluster built to render computer-generated imagery, as well as facilitate online collaboration supporting research, education and outreach.

HUBzero was created by the NSF-funded Network for Computational Nanotechnology starting in 2002 with the development of the first hub, named nanoHUB.org.[7] Since then, usage of nanoHUB.org has grown exponentially. In 2011, it served 400,000 visitors from 172 countries worldwide. Of these, a core audience of more than 190,000 users watched seminars, downloaded podcasts and other educational materials, and accessed more than 230 nanotechnology simulation tools. While accessing the tools, users launched a total of 390,000 simulation runs via their web browser and spent nearly 10,000 person-days collectively interacting with tools and plotting results. To date, there are more than 719 citations to nanoHUB.org, its simulation tools, and other resources, in the academic literature.

In 2007, HUBzero was spun off from nanoHUB.org as a separate project and software package to power newly created hubs. It was released as open source software in April 2010, and today, HUBzero supports more than 40 hubs with a combined audience of more than 650,000 visitors each year. One of these new hubs, the NSF Network for Earthquake Engineering Simulation (NEES), a $105M award connecting 14 institutions, is a particularly large project using the HUBzero platform not only for simulation, but also for experimental data collection. NEES Co-PI Rudi Eigenmann will oversee the IT architecture for GEOSHARE, which will have its own, dedicated HubZero platform, with specialized personnel developing the database, geospatial processing capabilities that will facilitate building and sharing analysis and decision support tools that respond to the needs of specific users. We envision having simulation tools such as crop models that can be used by the GEOSHARE community and also repositories of scripts using standard, open-source tools such *R* and *GRASS* that can execute geospatial processing of any complexity in order to generate maps, summary statistics and analysis results for GEOSHARE network members around the world.

[6] M. McLennan, R. Kennell, "HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering," Computing in Science and Engineering, 12(2), pp. 48-52, March/April, 2010.

[7] G. Klimeck, M. McLennan, S.P. Brophy, G.B. Adams III, M.S. Lundstrom, "nanoHUB.org: Advancing Education and Research in Nanotechnology," Computing in Science and Engineering, 10(5), pp. 17-23, September/October, 2008.

***GEOSHARE's role in the current geospatial institutional landscape*:**

GEOSHARE is different from many existing activities and programs in that it has both global and local components. GEOSHARE emphasizes data production, consistency, validation, and interoperability using state-of-the art methods. GEOSHARE will also offer time-series, geo-referenced data to support analysis of long run development and sustainability challenges as well as validation of analysis tools. Moreover, the HUBzero technology sets GEOSHARE apart from other data portals that constrain the user to built-in capabilities such as spatial zooming or automated summary statistics. GEOSHARE's cyberinfrastructure has been chosen to facilitate technology transfer, both by building capacity for stakeholders in developing country institutions and engaging students and scientists.

GEOSHARE is highly complementary to other important global data infrastructure efforts relating to land use, food security and sustainability. The most obvious complementarities are with similar efforts currently underway at regional scale. In this context, the Gates Foundation-funded *HarvestChoice* initiative led by IFPRI and the University of Minnesota, is worthy of special note due to its scale and ambitions. This important effort has a strong regional focus, with greatest emphasis placed on Sub-Saharan Africa. GEOSHARE will capitalize on *HarvestChoice* activities through the Africa node, led by GEOSHARE Co-PI, Stanley Wood, who is also Co-PI on *HarvestChoice.* GEOSHARE will benefit the regional activities by bringing to bear the expertise of some of the world's leading scientists via the global research nodes, as well as by making available state of the art cyberinfrastructure.

Another important type of complementarity is that offered by existing global data base infrastructure projects offering related, and sometimes overlapping, data sets. For example, the data on soil fertility generated by the *GlobalSoilMap.net* project can be linked to GEOSHARE allowing users to match these soils data with other relevant information archived in GEOSHARE. Similarly the CMIP archive of climate model results will provide an important source of information on historic and future climate. The United Nations Food and Agriculture Organization (FAO) is the world's leading source of globally comparable, national scale data on agriculture, land and water use. GEOSHARE will draw on these FAO resources in evaluating its aggregated geospatial data. And, of course, GEOSHARE will follow the guidelines and protocols established by Global Earth Observations (GEO) and the associated system of geospatial data bases (GEOSS) which will provide an important link to the broader global, geospatial community. It will also collaborate closely with the sub-groups working on agriculture, including the agricultural mapping subtask being led by the International Institute for Applied System Analysis (IIASA).[8]

GEOSHARE will benefit from recent initiatives to invest in agricultural monitoring across the developing world, as discussed by Sachs *et al.*[9] and currently under implementation through on a recent grant from the Gates Foundation. Based on discussions with that project's leadership, there appear to be at least

[8] See, L., Fritz, S., Thornton, P., You, L., Becker-Reshef, I., Justice, C., Leo, O. and Herrero, M. (2012). Building a Consolidated Community Global Cropland Map. *Earthzine*. 24 Jan 2012.

[9] Sachs, J., R. Remans, S. Smukler, L. Winowiecki, S.J. Andelman, K.G. Cassma, D. Castle, R. DeFries, G. Denning, J. Fanzo, L.E. Jackson, R. Leemans, J. Lehmann, J.C. Milder, S. Naeem, G. Nziguheba, C.A. Palm, P.L. Pingali, J.P. Reganold, D.D. Richter, S.J. Scherr, J. Sircely, C. Sullivan, T.P. Tomich, and P.A. Sanchez, 2010. "Monitoring the World's Agriculture." *Nature*, 2010. 466(7306): p. 558–560.

three areas of complementarity. The first is the potential for 'ground-truthing' GEOSHARE data. Since the latter are typically constructed by combining satellite and administrative unit data, a critical method for validating these geospatial data is via surveys and other forms of 'on the ground' monitoring. "Crowd-sourcing" offers another innovative source of monitoring, whereby individuals communicate field level observations to a centralized data base via geo-referenced cell phone applications. Two examples with immediate relevance are the new partnership between the World Bank and Google[10], as well as the IIASA-based "Geo-Wiki" project aimed at crowd-sourcing of global data on land use. These efforts will offer further avenues for validating GEOSHARE data.

A second area of complementarity between GEOSHARE and these agricultural monitoring projects stems from the desire to 'scale up' findings and recommendations from the site-specific studies. In order to do so, it is important to know how representative are the biophysical and socio-economic conditions at the existing sites when compared to the policy-affected region. It will also be useful to refer to GEOSHARE data on agro-climatic conditions, land use and poverty when selecting regions for new monitoring sites. Given the high cost of these monitoring activities, it is important that such sites be selected to be as representative as possible of the country/region in question.

The third area of complementarity between GEOSHARE and agricultural monitoring projects resides in the cyberinfrastructure. Inevitably biophysical and economic modeling will be required in order to extrapolate from the observed (i.e. monitored) to the unobserved regions. Such extrapolation will require two things which GEOSHARE's cyberinfrastructure can provide: comparable data on monitored and non-monitored areas, and modeling tools needed to make decision-support predictions based on these data. By leveraging Purdue's HubZero architecture, analysis tools developed anywhere in the world can be brought to bear on local and regional problems of this sort.

GEOSHARE will also be an important source of input data for other biophysical and economic modeling efforts. Of immediate relevance is the family of crop models developed under the Agricultural Model Inter-comparison and Improvement Project (AgMIP).[11] In particular, global data on current yields will be critical as the AgMIP team seeks to scale up its results. The Global Trade Analysis Project (GTAP)[12] offers another community of modelers who would benefit from use of GEOSHARE data. Comparable, but outdated, data currently underpin the widely used GTAP-AEZ data base and associated models[13]. Updating this data base from its current (ca. 2000) benchmark and enhancing it with information on irrigation, climate and poverty would greatly extend the applicability of GTAP-based analyses of food security, land use, climate and poverty by the 10,000 members of the GTAP user community.

---

[10] Anstey, C., "Empowering Citizen Cartographers", in New York Times, http://www.nytimes.com/2012/01/14/opinion/empowering-citizen-cartographers.html: Jan. 13, 2012.

[11] www.agmip.org/

[12] www.gtap.org

[13] Monfreda, C., N. Ramankutty and J. Foley, 2008. "Farming the Planet: 2. Geographic distribution of crop areas, yields, physiological types and net primary production in the year 2000", *Global Biogeochemical Cycles* vol. 22. Hertel, T. W., S. Rose and R. Tol (eds.) (2009). *Economic Analysis of Land Use in Global Climate Change Policy.* Abingdon: Routledge

Governance

The GEOSHARE concept was originally developed at the request of the UK Science Advisor and it has been validated by two dozen peer-reviewers (http://www.agecon.purdue.edu/foresight/). In the wake of the inaugural workshop hosted by Purdue University (www.geoshareproject.org), funding for a pilot project has been obtained from the UK Department for International Development, the USDA Economic Research Service, and the UK Department for Environment, Food and Rural Affairs. The pilot project provides modest funding for two of the global nodes, two of the regional nodes, as well as a start for the Purdue-based cyberinfrastructure within the context of several case studies aimed at demonstrating the value of GEOSHARE for facilitating decision making. In addition, there is support for stakeholder consultation as well as development of economic and institutional analysis of GEOSHARE. The pilot effort will culminate in a donors' forum aimed at mobilizing support for, and stakeholder participation in, the first decade of activities under GEOSHARE.

This pilot project notwithstanding, it is worth reviewing some of the ideas for the governance structure which have emerged thus far. The current proposal envisions establishment of a Scientific Committee and a Governing Board (Figure 2). The Scientific Committee will be responsible for recommendations on database content and interoperability, as well as analysis tools and methods of delivery. Membership will comprise the node leaders, as well as other experts as needed, and will guide technical aspects of the Project. The Governing Board will oversee management of GEOSHARE. Membership will comprise Project leadership, representatives of funding agencies, as well as key stakeholders. Professor Thomas Hertel, Distinguished Professor of Agricultural Economics at Purdue University, will serve as GEOSHARE Executive Director and will be responsible for the overall direction and leadership of GEOSHARE, including the strategic selection of research nodes and new partnerships. Dr. Nelson Villoria, Research Faculty member in Purdue's Department of Agricultural Economics will serve as Director responsible for the day-to-day operation of GEOSHARE. Professor Rudolf Eigenmann of Purdue's Department of Electrical and Computer Engineering will serve as the Center's Cyberinfrastructure Leader and be responsible for guiding the development of the overall GEOSHARE IT strategy. Each node in the network will have its own Director and the organization of individual nodes will be at the discretion of that Director.

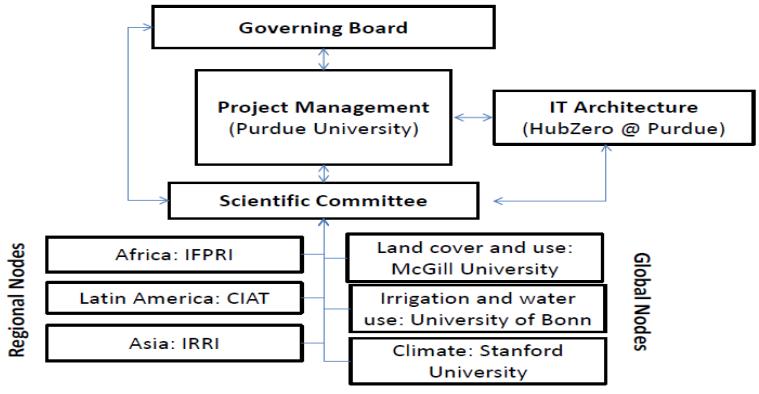

Figure 2. Organizational Structure of GEOSHARE

Long Run Sustainability: A Key Role for Stakeholders

Funding for GEOSHARE will come from a consortium of interested stakeholders. Here, we draw inspiration from the Purdue-housed Global Trade Analysis Project (GTAP), a successful database infrastructure effort, initiated by Professor Hertel 20 years ago, which now supports a network of 10,000 individuals in more than 150 countries undertaking global trade and environmental analysis. Core support for GTAP comes from 30 sponsor institutions, each of which is represented on the Advisory Board. In addition to its analytical data base, short courses (of which there have been more than two dozen), and the Annual Conference on Global Economic Analysis have been important elements of success for that Project. GEOSHARE will also utilize short courses and conferences as tools to build a global network of users, contributors and stakeholders. Membership on the Governing Board will require a multi-year commitment to project sponsorship. As with GTAP, the Board will play a key role determining the Project's strategic direction. And, by involving a large and diverse set of sponsors, the fate of the Project will not rest in the hands of a single institution.

The GTAP experience also shows the importance of flexibility and responsiveness to the evolving needs of stakeholders, if such an effort is to be sustainable over the long run. GTAP started out in 1992 primarily as a vehicle for analyzing GATT/WTO issues. Over time, the primary areas of application evolved to bilateral and regional trade agreements and, most recently, to global environmental issues – particularly climate change policies and impacts. The scalability of the GEOSHARE platform will readily facilitate the addition of new topical research nodes, making it readily responsive to new stakeholder needs. For example, the recent GEOSHARE workshop at Purdue University gave rise to discussions for a node on land tenure data led by The World Bank. By combining data on land tenure with that on agricultural productivity and environmental conditions, GEOSHARE can facilitate sound decision making regarding the leasing of agricultural lands to foreign entities – an area which has been problematic, particularly in Africa.

GEOSHARE seeks to engage stakeholders in developing countries and work with them to identify key areas of need for geospatial data. This will be facilitated by the regional nodes which comprise three members of the Consultative Group for International Agricultural Research (CGIAR): IFPRI (Africa), IRRI (Asia) and CIAT (Latin America). For example, the IFPRI-led Africa node will provide links to the Regional Strategic Analysis Knowledge Support Systems (ReSAKSS) whose aim it is to support evidence-based policy and investment analysis. GEOSHARE project leadership will also participate in these regional meetings to engage in a dialogue about potential applications of GEOSHARE's geospatial data and tools. By way of illustration, Box 2 offers a specific example of how GEOSHARE can benefit regional stakeholders in Africa.

BOX 2. SAMPLE GEOSHARE PROJECT:

Identifying Technological Solutions for Increasing Food Security in Sub-Saharan Africa

As part of the pilot phase of the project, funded by DFID, DEFRA, and USDA, the IFPRI node has proposed to use spatial and household survey data to map out the geographies of national agricultural research innovation needs (e.g. priority high-poverty regions, commodity production zones, irrigated areas, etc.) as well as the geographies of regional and sub-regional research activities (e.g., farming systems, agro-ecological zones, highlands, semi-arid areas etc.) with the objectives of: (i) helping individual countries to identify the location of relevant research that responds to the needs identified, for example, in the national plans derived from the Comprehensive Africa Agriculture Development Plan; (ii) helping international and sub-regional research organizations to determine the locations across Sub-Saharan Africa in which there is a demand for innovations being currently developed or tested in different research sites and (iii) informing research organizations and donors on how closely research priorities and investments (the supply) match the sum of demands expressed across all countries in need, thus pointing out to themes and zones in which there could be over-investing relative to current and expected demands.