Summary of breakout groups as well as general discussion on HUBZero as Cyber-infrastructure for GEOSHARE (see below for detailed comments from each breakout group)

Q1: What specific needs should cyber-infrastructure for the geospatial community address? Do these needs vary across data/model users and data/model providers?

It is very important to have standard data formats for different classes of information stored in GEOSHARE (including meta data). Any open source software developed on the HUB should be written in such a way that it operates on this particular data format (allowing for both spatial and temporal variation). The Potsdam Institute (PIK) group has developed such a data format, and they also have a tool for converting other standard data formats into this standard format, as well as re-converting the standard format back into these other, widely used formats. Developing these tools for use on the HUB should be a high priority as they will circumvent the need for multiple versions of the same tool.

Once a standard data format has been defined, then crafting a standard aggregation tool will be relatively straightforward. Offering such a tool, along with the global, gridded datasets, will have immediate value, as the data developers at the meeting underscored the strong demand for aggregated datasets and much of this burden currently falls on the developers themselves. This facility, combined with *doi*-based referencing of the datasets, should create a strong incentive for uploading these data.

In addition to the aggregation tool, it would be very useful to have tools to perform other widely desired operations, including: visualization of the data (including mapping), extracting data for a single country or region from the global dataset, computing statistics, crowd-sourcing data contributions, and uploading comments on the data base and related data bases and tables of information. [See also the related comments in data endorsement section and the recommendations for data parsing, visualization and comparison tools]. When combined, these tools will also allow for improved peer-review and endorsement of global geospatial data for particular purposes ("horses for courses") – something which presents a significant challenge at present. In the long run it will be important to also include metrics indicating the level of uncertainty in various datasets.

The incentive for users to upload data into the standard format could be further enhanced if GEOSHARE became an accepted repository for supplemental information cited in publications. Scientists are familiar with using "required formats" to meet publication guidelines. The importance of facilitating peer review (and endorsement where appropriate) of new critical spatial datasets which then remain readily available and easily cited is attractive and currently unavailable in this field.

Q2: What are the workflows implied by these needs? How much of the workflow should be online? Is the hub a place to ensure reproducible research?

Participants acknowledged that there are some steps in the data preparation stage which cannot be put online. However, once a new data layer has been prepared, the idea of being able to substitute it into a workflow, such as those discussed at the workshop, is very appealing. And being able to carry this all the way through data reconciliation to a new set of model results which might, for example, permit assessment of the impact of climate change on agriculture, or the impacts of future water scarcity, would be tremendously valuable. When combined with a *doi*-based reference which allows readers of a journal article to access, reproduce, and potentially revise the workflow, this could transform the way science and the associated citation credits is currently undertaken in our field.

Q3: What incentives can be put in place to ensure these workflows are contributed by authors and utilized by fellow researchers?

Participants recognized that GEOSHARE will not succeed without an appropriate set of incentives for the contribution of datasets and tools as well as for their use and subsequent citation. Data base developers at the workshop noted that allowing others to easily comment on their data would be a significant incentive. And having multiple data bases on the same subject, accompanied by easy-to-use tools, would make comparison easy and appealing. Citations will provide a strong incentive for contributing data on the HUB, and, to the extent that HUBZero tools facilitate peer-review, posting the data on the HUB may strengthen the case for publication in prestigious journals.

The incentives for users to access and utilize workflows on the HUB are perhaps more obvious. If this is where they can get the data and tools to do their work, they will come. And if the HUB offers them a way of doing their job more easily, then will perform the work on the HUB. This was nicely illustrated by an AgMIP tool, https://mygeohub.org/resources/agmip, developed by Nelson Villoria in collaboration with Joshua Elliot. Users can now access aggregated climate impacts for a variety of socio-economic scenarios, global climate and crop models within a matter of minutes, whereas this required economic research teams to dedicate a full year of a professional's time to the task!

The opportunity for communities of practice working within GEOSHARE to "reach-out" to encourage other groups to add "missing data" to complete/improve datasets/tools is a further incentive for contribution (i.e. "pull" form the community rather than "push" from individual groups). Being part of a global community-of-practice may be its own incentive for many individuals and smaller groups.

Global recognition that project data will be made available via GEOSHARE as a condition of grants is yet another form of incentive which could be further strengthened by the Advisory Board (see breakout session on institutional design).

Q4: In what ways will the use of HUBZero contribute to transparency, credibility, and acceptance of the underlying data and workflows available within GEOSHARE?

HUBZero automatically solves the requirement for peer-reviewed publications that the data and methods be made publicly available. The solid *doi* citations will greatly facilitate this process and aid in the 'versioning' of data bases and workflows. It will be important to use these workflows to better document, and transparently report, the overall uncertainties surrounding this area of science. For example, what are the "real" error bars around current global estimates of future climate impacts on agriculture? This uncertainty extends far beyond the array of possible socio-economic scenarios and the multiple climate, crop, and economic model forecasting methods to also encompass uncertainty about the source data.

Detailed comments from breakout groups follow:

Hub Zero Break out – Room 121. **Group 1:** Carol Song*, Paul Hendley**,

- Q1: What specific needs should cyber-infrastructure for the geospatial community address? Do these needs vary across data/model users and data/model providers?

- Q2: What are the workflows implied by these needs? How much of the workflow should be on-line? Is the hub a place to ensure reproducible research?
- Q3: What incentives can be put in place to ensure these workflows are utilized?
- Q4: In what ways will the use of HUBZero contribute to transparency, credibility, and acceptance of the underlying data and workflows available within GEOSHARE?

Question from floor: - would HubZero run client side or server side

Discussion notes

The discussion was free flowing and did not focus in too tightly on the sequence of questions – I have added in brackets the Question(s) to which each discussion point relates

- Need to handle time series – e.g. MODIS RS data. Does GEOSHARE plan to be able to make this type of time series available so users could plot out data. The answer would be "NO" unless the MODIS time series was part of a desired workflow or part of a supporting dataset to a publication involving this time series (**Q1**)
- Data heritage issue is increasingly important – e.g. Data.gov initiative which is becoming more significant and will require pointers to supporting documentation and other forms of metadata. The innate HUBZero capability to create log of all input parameters and all sources for EACH run of the tool was seen to be a very strong plus to help users address these increasing requirements– We will need to make sure this is covered thoroughly in endorsement session.(**Q1, Q3 and Q4**)
    - A related discussion covered the need to make sure validation is in place and rationale for model selection as part of the workflow (**Q1**)
- It would be a considerable benefit to provide a really solid DoI for each key dataset – the question to answer is what is the right thing to be versioning" (**Q1 and Q4**)
    - does each revision get a new DoI or each run?? The recommendation was for each run to be uniquely identified.
- A key incentive and contribution to transparency is vetting of new datasets and workflow approaches. Peer review is not seen to be working well. Approval of new datasets needs to be via credibility by consensus and this requires tolls to allow potential users to rapidly assess what is included/excludes, the resolution issues and data quality/gaps. HUBZero offers the opportunity to provide reviewers and the community with ways of viewing and visualizing the data. (**Q4/Q3**)
    - Need a systematic way to generate diagnostic code for analyzing datasets for anomalies
    - Journal paper – reference to Datasets in GEOSHARE
- GEOSHARE could provide a key ability to help assess datasets – visualization tools – Box and whisker. Map tools, data comparison. GEOSHARE Team needs to develop understanding of what primary metrics are needed/may be useful for QA/QC. (**Q2,Q3, Q4**)
- HUBZero convenience – collaboration is the answer – it makes collaboration easier for reproducibility across groups (**Q3/Q4**)
- Data vs models is important to answer the question on how much of the workflow that should be on line

- Metadata system – fusion products are particularly key – algorithms of all workflows need to be rigorously described. (**Q2/Q4**)
    - GEOSHARE needs to include/recongnize/adopt the FGDC and ISO standards of metadata
- Incentives – collaboration, citations, Journal requirements, available system designed for sharing code and data, (**Q3/Q4**)
    - HUBzero permanence/sustainability is a question that has come up a number of times – the group saw there were several ways of addressing this such that it should not be a problem.
        - When an entry goes into GEOSHARE then copies are retained.

HUBZero and GEOSHARE Breakout 14:30-15:15 September 11th, 2014

Group 2 Room 129

Rapporteur: Jingyu Song, Purdue University

What do you think about the HUB?

It consists of a lot of resources, requires people to contribute and upload, but is very useful to various groups and benefits a huge community. Diverse groups such as geographers and economists need examples of people actually using the HUB to see how it works and there is a need for the HUB to act as a bridge over different groups.

User cases and goal of GEOSHARE:

With HUBZero and GEOSHARE, there is no need to worry about software licensing issues, no need for training of using models. One can just go in the browser and click.

People do not need to be an expert in a particular field, the data and models will help them in their own fields by making use of other people's analysis available. Different groups can get different implications from the same model.

However, GEOSHARE is much more than just clicking of buttons. More people will be able to share and communicate ideas, models. There will be different versions of models for users to decide which to use, depending on what suits their purposes. GEOSHARE serves as a platform for groups to communicate and gain better understanding of the data and models. It lowers the entry barriers, and provides opportunities for people to understand the data, and reduces the efforts of creating models. Fundamentally, GEOSHARE is about sharing the burden of producing public goods, i.e. data, and bring together publically available data for researchers and private sectors to share.

Citation of codes & data:

The hard work needs incentives. Researchers require the HUB to have the ability to publish their datasets and to be recognized for their work. Data collectors need to be identified, scholarly products of code and data need to be cited correctly. There should be good tracking of contributions, improvements, i.e. “durable citations”.

Example: AgMIP

It is the final step towards the public, but there are barriers during the process of sharing models and data. Difficulties are there before publications becoming available to the public. Thus, GEOSHARE needs to pay attention to registration, usage tracking over time, application of tools, etc.

Rapporteur notes on HubZero session (Nelson Villoria):

Breakout sessions to discuss prospects and potential for HUBZero to transform the landscape for analysis of food and environmental security at global scale – 45 minutes for discussion and 15 minutes to report back to the group.

SUMMARY

- HubZero should provide a place where to put raw data to be shared, comment on it, and allow people to use it. This is key for having incentives for sharing.
- From the user perspective, a tool that standardizes disparate datasets in a set of common classification ways and allows custom querying and aggregation is highly needed. (Note of the rapporteur: this is taking back us to the origin, when we noticed that all existing sites were static and therefore useless for the users.)
- HubZero can help with reproducibility, but more importantly, with traceability. I.e., how is the data used and what assumptions are made along the way.
- The prestige of being a node leader is an incentive to act.
- The idea of changing data layers to explore the consequences of the data is neat.

HIGHLIGHTS BY QUESTION

Q1: What specific needs should cyber-infrastructure for the geospatial community address? Do these needs vary across data/model users and data/model providers?

- A tool to harmonize disparate datasets would be very useful. Later conversation with Jan Phillip Dietrich and other refined thinking on this. What is needed is a tool that gets any spatial dataset, and can classify along its dimensions. This has implications on how to store the data, and will be discussing this in the next few weeks.
- Ability to extract data by country. (Navin proposed it.)
- It would be also nice to have a facility to collect data. The idea is that different people collecting tables and the like, could upload them to HubZero for all to see. Navin also mentioned that he’d like to use it for getting feedback on its data.

Q2: What are the workflows implied by these needs? How much of the workflow should be online? Is the hub a place to ensure reproducible research?

- Experts such as Navin and Andy discussed that the reconciliation stages are very case specific so they are hard to be put online. The idea of the Hub should not be to reproduce a process from scratch.
- The idea of reproducible research have a lot of traction. It would be nice to see how models use data and what the parametric assumptions are.

Q3: What incentives can be put in place to ensure these workflows can be utilized?

- Navin pointed out that the biggest incentive would be: (1) having data posted by others as explained above (i.e., people uploading tables and the like); (2) Having people commenting on their data. A sustainable Agromaps of sorts.
- The Open Data Policies and Initiatives should provide also incentives for sharing data.
- Being a node leader is also an incentive if it is associated with the prestige of belonging to GEOSHARE.

Q4: In what ways will the use of HUBZero contribute to transparency, credibility, and acceptance of the underlying data and workflows available within GEOSHARE?

- No time to discuss specifics. However, points above are relevant for this.

Some notes on the discussions about HubZero and GEOSHARE: Ulrike Wood-Sichra

Breakout, Wed Sep 10 2014, 14:30 – 15:15

Q1: Specific needs …

- Data harmonization for input models
    - Tool to make any data aggregation
    - Compute statistics on data
- Does aggregation change the data?
- Options to edit datasets
- Put own datasets on the Hub
- Create tools to design tools

Q2 + Q3: Workflows implied by these needs …

- Crowd sourcing
    - also of tabular data
    - incentivize (eg by sharing of data – conflict with open access?)
    - how to make sense of results?
    - lots of human interaction, not automatic
- Harmonization creates workflow
    - define which data can/must be harmonized
    - define who harmonizes
    - must be flexible on boundaries

- Hub to keep track of use of data, where, how often
- Need to believe in primary data, correct it through user input (problem with non-proprietory data)
- Efforts of putting a model in the pipeline – are they justified?

Q4: In what ways will the use of HUBZero contribute to transparency …

- Requirements for publishing and availability solved on the web

---

HUBZero and GEOSHARE Breakout (Group 4)

Rapporteur: Dave Gustafson

Q1: What specific needs should cyber-infrastructure for the geospatial community address? Do these needs vary across data/model users and data/model providers?

- Data versioning
- Device independence
- Great example: Crowd-sourcing corrections to land-use classification
- NASA has some tools for doing great classification, very expensive (ref. Molly Brown)
- Important to document levels of confidence in data
- Different users have different data needs. Scientists tend to want higher resolution, confidence, etc. Others, not so much
- Need data of farmer behavior, agronomic inputs

Q2: What are the workflows implied by these needs? How much of the workflow should be on-line? Is the hub a place to ensure reproducible research?

- Should make it easy for scientists to produce new workflows
- Maybe we need a "Siri" for converting free-text questions into models
- Some nervousness about this idea, unskilled people start using it
- Need to make sure that we accurately characterize and communicate uncertainty in model predictions
- Workflows are a great way to generate documentation and enable reproducibility of modeling

Q3: What incentives can be put in place to ensure these workflows are utilized?

- Put emphasis on users, what are the questions that they want answered?
- Let users help dream up the questions.

Q4: In what ways will the use of HUBZero contribute to transparency, credibility, and acceptance of the underlying data and workflows available within GEOSHARE?

- Need test datasets
- Automatic creation of documentation is key
- Data validation

- Communicating uncertainty is important
- GIOVANNI tool developed by NASA, could be relevant