Summary of breakout groups as well as general discussion on Data Fusion for GEOSHARE (see below for detailed comments from each breakout group)

Q1. Transparency (simplicity) and sophistication (complexity) in data fusion. Where should GEOSHARE aim on this spectrum?

This question itself generated considerable discussion, eliciting questions like: "Can you have transparency in complex models?" "Can the complex models can be broken down into step by step to ensure transparency still?" One thing that all participants agreed upon is that all the steps in any data construction process should be documented. More generally, the principle of Occam's razor should be applied – namely use the simplest possible approach to get the job done. However, many of these tasks are inherently quite complex.

It was also recognized that there are multiple levels of sophistication in the prospective user communities: those who wish to simply access/view and possibly download the data, those who wish to perform some simple operations, and those who want to get inside and 'tweak' the algorithm, thereby potentially producing a new version of the dataset.

Consequently, there may be value in making data/workflows available in ready-to-use and more sophisticated formats

Q2. What is the role for prior information in this process?

This question was formulated with a particular workflow in mind – namely that involving the SPAM software which incorporates 'priors' on the allocation of land by pixel. The specification of prior information is integral to the Bayesian approach to data fusion. Others see this as establishing a baseline hypothesis. In any case, these priors must be fully documented as per the discussion in Q1 above.

Q3. How can GEOSHARE harness private- and public-sector knowledge and expertise in the process of data fusion?

Participants agreed that harnessing such external knowledge represents a key opportunity for GEOSHARE. The challenge is how to go about doing this. The point was made that the private sector has a great deal of information on hand that it considers 'out of date' but which could prove very useful for scientists and other decision makers. This includes information from yield trials and cropping calendars which could be used to improve crop models, as well as other socio-economic information gathered at geospatial scale (including food consumption statistics). There is lots of information out there. The question is one of priorities? What data are needed? This comes back to the models being used to support decision making and the underlying workflows. If GEOSHARE can make available a list of needs and the rationale for making these data available for public-use, participants felt that there were good prospects for getting industry to donate such data – especially if there are mechanisms to translate it, thereby removing proprietary content without requiring expensive private industry time. CIMSANS could serve as an effective intermediary between the private sector and GEOSHARE in this process. Finally, the point was reiterated that a few really good use cases are still very much needed in order to convince more private sector players about the value of GEOSHARE. The recent work by Navin Ramankutty and Stefan Siebert appear promising in this regard.

Q4. What is the role of ground-truthing and crowdsourcing in complementing data fusion based on more census and remote-sensing sources?

Participants were very supportive of the idea of ground-truthing the GEOSHARE datasets. This is another area where the private sector data could play an important role. Crowd-sourcing of data is more challenging. However, the work of Stefan Fritz at IIASA appears promising. He incentivizes participation through various types of 'games' which encourage additional contributions. The suggestion was made that Stefan's group could become a 'node' of GEOSHARE.

Detailed Comments follow

GEOSHARE workshop Data fusion Breakout – Thursday – Rapporteur: Paul Hendley

- Q1: There is a trade-off between transparency (simplicity) and sophistication (complexity) in data fusion. Where should GEOSHARE aim on this spectrum?
- Q2: What is the role for prior information in this process?
- Q3: How can GEOSHARE harness private- and public-sector knowledge and expertise in the process of data fusion?
- Q4: What is the role of ground-truthing and crowdsourcing in complementing data fusion based on more census and remote-sensing sources?

Q1

- Disagreed with question – we felt that Lack of simplicity does not mean lack of transparency
    - However, Occam's razor ought to be applied to ensure the simplest approach necessary to address all the needs simplicity
- Range of products – two types of customers "tweakers" and users – need to address both types
    - Because users vary in exactly what they need from certain datasets, our group recommends having products at three stages of complexity for some work products
- Self–selective – workflows in addition to standardized products
- Community endorsed "gold standard" is a key aspect of GEOSHARE
- Goes without saying that in GEOSHARE data banks we need full descriptions of all steps AND underlying data preparation steps – metadata needs to include the heritage of the data
    - Navin's three versions as an example of the 3 levels of complexity approach
- Danger is that documentation of all the underlying steps becomes exponential and metadata can get out of control
    - However we ought to consider adding ways of tracking citations and "credit" for all contributors – even to partial datasets

    - In addition to classic metadata, as a minimum, the working assumptions selected for5 a model run that makes that run unique need to be listed – especially for less expert model users
- Need to build in a "reality check process" with time that can lead to updating

Q2

Group felt that this was merely a subset of Q1 – we defined "Prior" as broad term general indicating a "starting baseline hypothesis"

- If using a Bayesian approach (or a model that uses priors) they are needed. However all the key requirements listed above for Q1 need to be addressed

Q3.

Examples of what private industry might have and why they might be prepared to share it were given e.g.

- Cocoa, trial data, cultivar data, day of anthesis
- Timing of data release may be a useful approach for industry to release their data – often the most up to date data is not important for model development and therefore old industry datasets might still be very useful for model development or calibration
- Other examples of private data
    - Cattle production
    - Dairy production….

GEOSHARE opportunities in public and private sector

- Use spatial connection/communities of practice
    - Operate through government agencies - ministries
    - IFPRI, CIAT, Agromap, CG system, FAO – example
    - GEOSHARE could reach out to organizations to justify why data release makes sense and to show examples of how it can be a general and local benefit
        - May be politically awkward – census responses cannot always be believed
    - Start with the best information – CENSUS's
        - Why not verify via private industry data?
- GEOSHARE to identify key data gaps
    - What might be available from private industry
    - Need a matrix of data needs
    - What is the crop calendar,
    - Other types of land use, what is pasture
    - **GEOSHARE needs to use these lists of needs and the rationale for what the data can do to help improve overall modeling etc. etc. in order to make detailed requests for specifics from private industry**
        - CIMSANS to act as intermediary

- AgMIP is an example of outreach
    - But was highly localized

Q4

- How can GEOSHARE facilitate this
- Examples discussed included Phenocam – CA oak project. However group felt that anything requiring extensive investment would be unlikely to get off the ground
- The group felt that examples like the Stefan Fritz ground sourcing game program are more likely to be feasible ways of getting Ground truthing across global and regional scales
    - Geoshare process – maybe ground truthing for "crowd sourcing approaches" should be a community of practice (i.e. a node) within GEOSHARE
- GEOSHARE to access scientist owned data to assist with land cover GT
    - E.g. request Access to already classified Remote sensing data identified from publications
    - Alternatively GEOSHARE could approach GOOGLE for datasets
        - Across the longer term, the best way of ensuring such data would be useful would be to design a pseudo standard for sharing classified data that could be used for supporting data submitted with journal article s (and stored /accessed via GEOSHARE)
- Guidance on data formats

Detailed Break out Comments: Rapporteur: Jawoo Koo

Q1. Transparency (simplicity) and sophistication (complexity) in data fusion. Where should GEOSHARE aim on this spectrum?

- Can you have transparency in complex models? Complex models can be broken down into step by step to ensure transparency still? Need to document how to get there. Not a tradeoff. You always need to have transparency. Sometimes you need to have the complexity, sometimes you don't. Either way, you need to know (be able to show) how you get there.
- Any example of so complex not transparent? Explaining complex crop models to producers; wouldn't be easy, will take time, will confuse.
- Opposite of transparency is "black box". For GEOSHARE, it should be transparent no matter what!
- Different audiences; transparency to one may not be so to others. Amongst us, it should be transparent.
- Users (like farmers) will favor something ***simple***, and we should be ready to respond that too.

Q2. What is the role for prior information in this process?

- What is the prior information? Input data to the models? Probability of input data? Statistical data? Time series?
- For transparency purpose, not just the model output but also the input prior data should be made available. Need raw data, and the steps all the way leading to the prior datasets (then through output).

Q3. How can GEOSHARE harness private- and public-sector knowledge and expertise in the process of data fusion?

- Companies see GEOSHARE's offer to have better quality data to the end users (farmers) as the benefit, for free. Put right information to the farmers, with uncertainties (how accurate do we have to be?). Companies don't want to keep this type of data private; they also want to make the data open and free and benefit users/farmers. Right level of granularity to give relevant advice (out of global datasets) is more challenging.
- Adapt language/topics into ones that can interest private sectors. Not developing the better model. Risk, profitability, economics are good.
- Companies can't ignore big/small data; they are interested.
- GEOSHARE is still too science-driven, not particularly welcoming to private sector.
- Can private sector's trial data shared for the model improvement? Yes!
- Can we directly approach farmers (private sector) to collect the information? You can collect very much up-to-minute information directly from farmers.
- Would research community be committed to interact with private sector partners (including farmers) to do this? A lot of work.
- What's the use-cases of GEOSHARE? Any success stories from the Hub we can learn from – like the nanohub, that also has the industry linkage? Need to have good group of PIs behind this and really use.

Q4. What is the role of ground-truthing and crowdsourcing in complementing data fusion based on more census and remote-sensing sources?

- Great, we should do, but tricky how to incentivize the crowd. Customized, packaged data delivered through Apps.
- Travel to visit farmers fields? Use a CGIAR App, to record where you are going and who you are visiting. Geotagged photos from camera. SMS-based survey and data collection too!
- What to do with FAO data? We know this won't agree with official statistics. Need strategies on the use of groundtruthing data, too.