CIFOR Annual Meeting 2017
Data discussion

1. How do we manage integration between datasets, and of different dimensions within a dataset? Can we build a framework that would favor integration?
2. What methods to ensure extrapolation domains exists, or that insights from the data can be relevant to different contexts in other locations and/or to globally insights/International public goods?
3. How to ensure quality of data? How to link quantitative and qualitative type of information?
4. Was data effective to trigger change?
5. What are the future needs? Which linkages and partnerships could be envisaged? What are our expectations versus the big data initiatives and platforms?
6. What could be the roles of FTA, as a place to valorize data and related work, to build frameworks for data collection, organization and valorization, and to build related linkages cross datasets and partners?

D R A F T

Creating value from data requires a new mindset. Silos are hard to escape, whether they are technical or conceptual. The answer is to focus on the outcomes and how they can be achieved.

This approach can be cultivated through looking at the data value chain, both (i) along the project cycle and (ii) integrated at the level of CIFOR.

A learning approach is also to be followed.

Draft Recommendations

1. Building a protocol to improve the quality of data along the research/project cycle

a. Research design and project startup

Data Management Plan (DMP) needs to be mandatory on the project start up. It is not required to provide detailed answers to all the questions in the project start up. DMP should be intended to be a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur.

DMP should consist of the following information:

- Data information
    - What is the purpose of the data collection/generation and its relation to the objectives of the project?
    - What types and formats of data will the project generate/collect?
    - Will the scientist re-use any existing data and how?
    - What is the origin of the data?
    - What is the expected size of the data?
    - To whom might it be useful ('data utility')?

- FAIR component description
- Allocation of resources
- Ethical clearance – which also link to the CIFOR RER

DMP is also to meet the requirements of the donor. The format/template of DMP on each donor usually different but has the same information. To minimize reporting effort a mechanism on how to export the information on PM system to the document needs to be considered.

Currently available table on the PM System needs to be improved with the DMP template. Information on: (1) name of the dataset, (2) description of the data and (3) type of the data should be generate into a data pipeline that is accessible to everyone through myCIFOR and published on the monthly publication update.

b. Data collection and acquisition

To ensure the consistency and enable the integration, CIFOR RDM Team need to develop a series of templates related to the data collection and data quality checklist.

Data collection plan stated in the DMP will be used as a starting point to help the team create the questionnaire beside the data collection check list. Data templates will standardize procedures and processes of a dataset that is handled by more than one person. A controlled vocabulary will be used as a guidance on data entry to reduce the number of errors made during data collection.

A series of standardization will apply for (at minimum, by using a globalized standard):
- format for geographic coordinate values.
- format on a date value
- codes and acronyms
- measurements
- unique ID of research object

During data collection, projects need to implement a quality assurance check of their data. Data should be reviewed as they are accumulated during a field research to minimize error, inconsistency, and data loss. Whenever errors or deficiencies are identified, steps should be taken to identify and mitigate their sources.

Changes in the equipment or procedures needs to be documented. The documentation will be archived in the data repositories as part of the lesson learned.

We encouraged to develop expert pool to review the collected data during the data collection process. Expert pool is a domain specific group of people that is developed by the project as a quality assurance process for the data reviewer.

c. Data curation

The main idea behind data curation is to enable more complete and high-quality data-driven models for CIFOR. Curation activities needs to be integrated into project workflow throughout the research lifecycle.

To add value to the data and to make data available for use by other researchers beyond the lifespan and purpose of the project for which the data were collected, information should be documented and published as a Data Curation Profile[1].

Data curation Profile can be published either as a working paper or as a data publication in a data journal such as Data In Brief Journal or Scientific Data Journal.

Each project to develop a "data quality case studies", to feed a compendium story of successful projects on managing their data. A documented lesson-learned book will help others to avoid the same mistake.

d. Storage

The key challenge of data capture is the variety and volume of data that can be captured. This poses a question about how to store this data, but more importantly how to process it for storage in such a way so that it is ready for analysis.

Qualitative data needs to be quantized so the result is presented into one single comprehensive dataset.

2) Building a framework to promote data integration

Interoperability issues need to be taken into account at the stage of the production of the data. A standardization and the creation of data in machine-process able format can help to overcome this barrier.

a. Laying the ground for integration: understanding the multiplicity of objects at hand.

Understanding the spectrum of dimensions/variables researched: including typologies, glossaries, units, research and measurement/acquisition protocols.
Integration of different research/data dimensions in a project

b. Integration of different projects' results: learning, mapping and interoperability from spatial dimension / geographic mapping and topic.

The resulting understanding enables us:
To understand what worked well and feed-back into a lessons-learned book: Importance to figure out how do we feed-back from lessons learned: which mechanisms to feed-back into projects?
To analyze for suitable areas to make investments that can create new value from data

Piloting integration in one country, for instance Indonesia.

[1] Adopted from http://datacurationprofiles.org/. The sample of profile can be check on: https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1027&context=dcp