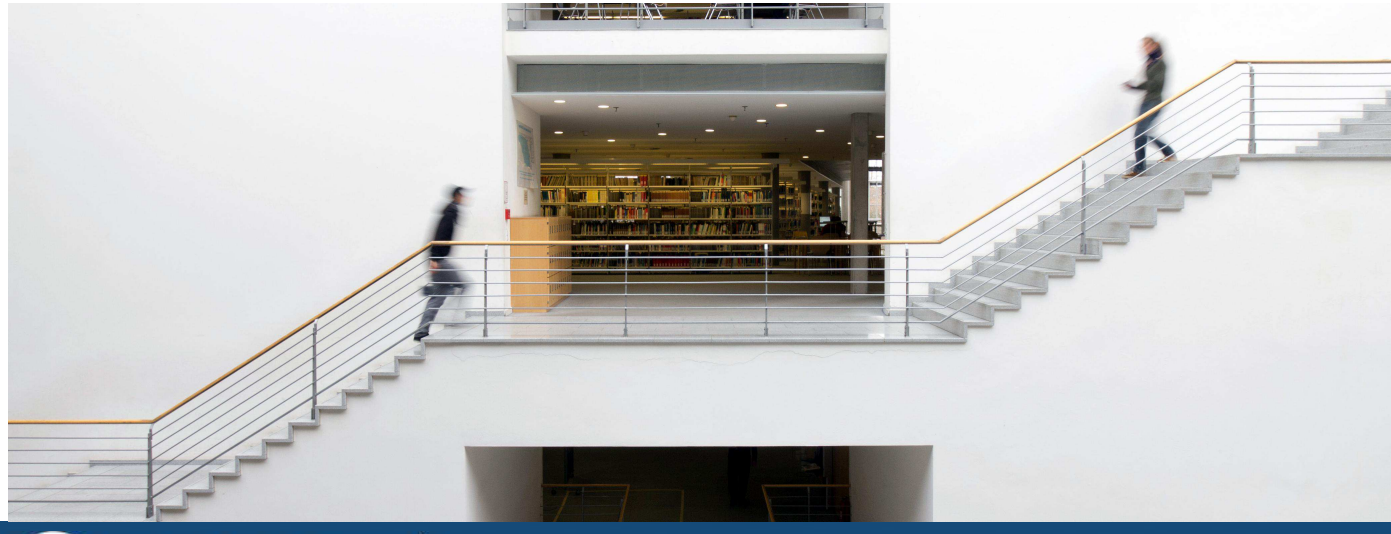
SUB | NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

Using BExIS in a complex Collaborative Research Centre

Thomas Fischer

BExIS 2 User and Developer Conference 2015

July 9, 2015



Contents

1. Short description of CRC 990
2. Some challenges not to be considered here
3. Data challenges
    - Large unstructured data sets
    - Unwieldy structured data sets
4. Metadata challenges
    - Structure
    - Download and Upload
    - Interface
    - Contents
5. General considerations

Collaborative Research Centre 990

Ecological and Socioeconomic Functions of
Tropical Lowland Rainforest Transformation Systems
(Sumatra, Indonesia)
Collaborative Research Centre 990
(= CRC 990 = SFB 990 – Sonderforschungsbereich)

6 Faculties plus the Library:

- Faculty of Agricultural Sciences
- Faculty of Biology
- Faculty of Economic Sciences
- Faculty of Forest Sciences and Forest Ecology
- Faculty of Geosciences and Geography
- Faculty of Social Sciences
- Göttingen State and University Library (SUB)

11 departments and institutes

Research in Indonesia

Partners in Indonesia, the country of research :

- Indonesian Institute of Science (LIPI)
- Bogor Agricultural University (Java)
- University of Jambi (Sumatra)
- Tadulako University in Palu (Sulawesi)

In addition: Harapan Rainforest Initiative, National Park Bukit Duabelas

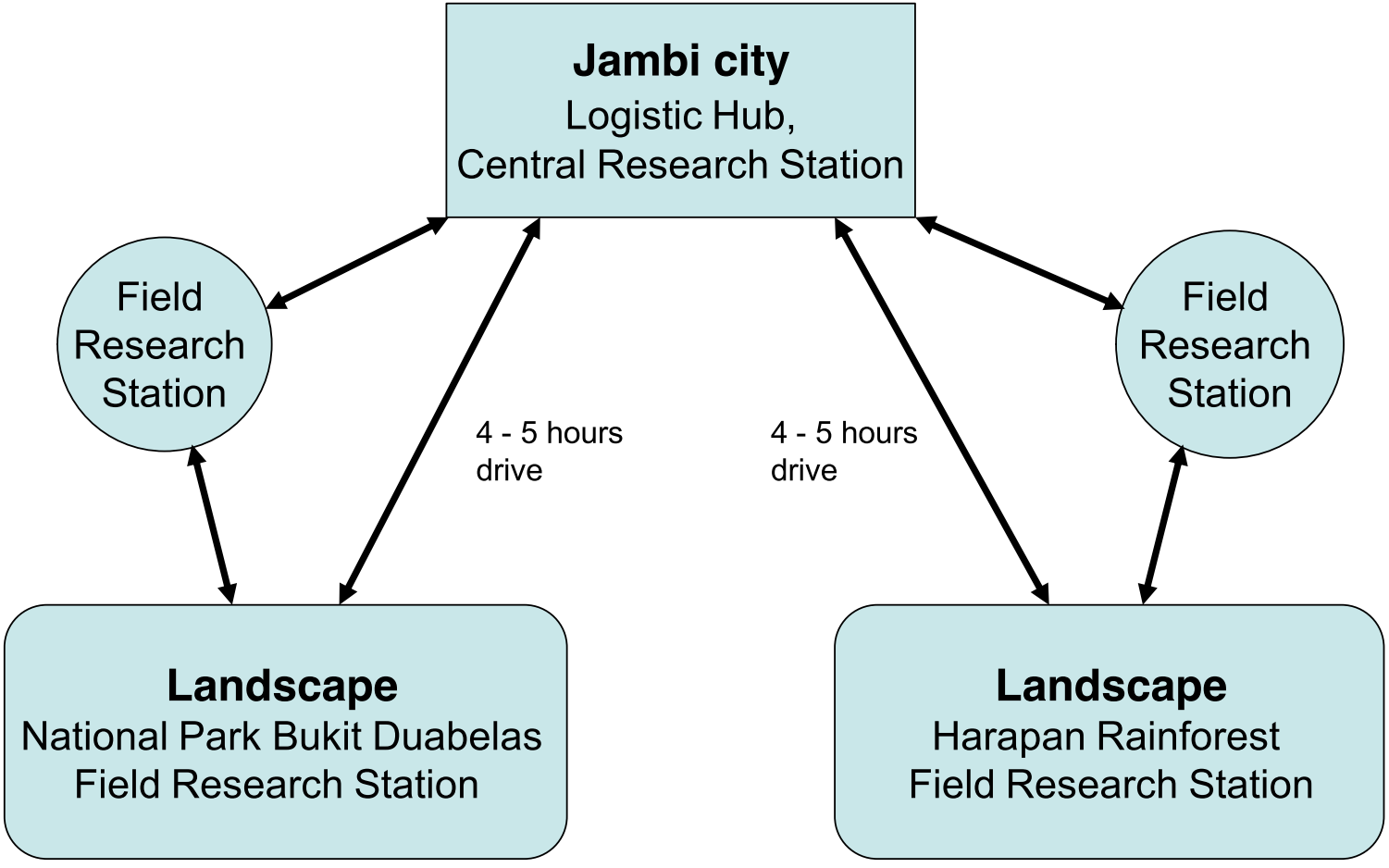
Research stations in **two landscapes** with **four transformation systems**:
oil palm, rubber, jungle rubber & tropical lowland forest

Four core plots (50m × 50m) in each transformation systems (= 32 plots) with 5 subplots (5m × 5m) in each plot

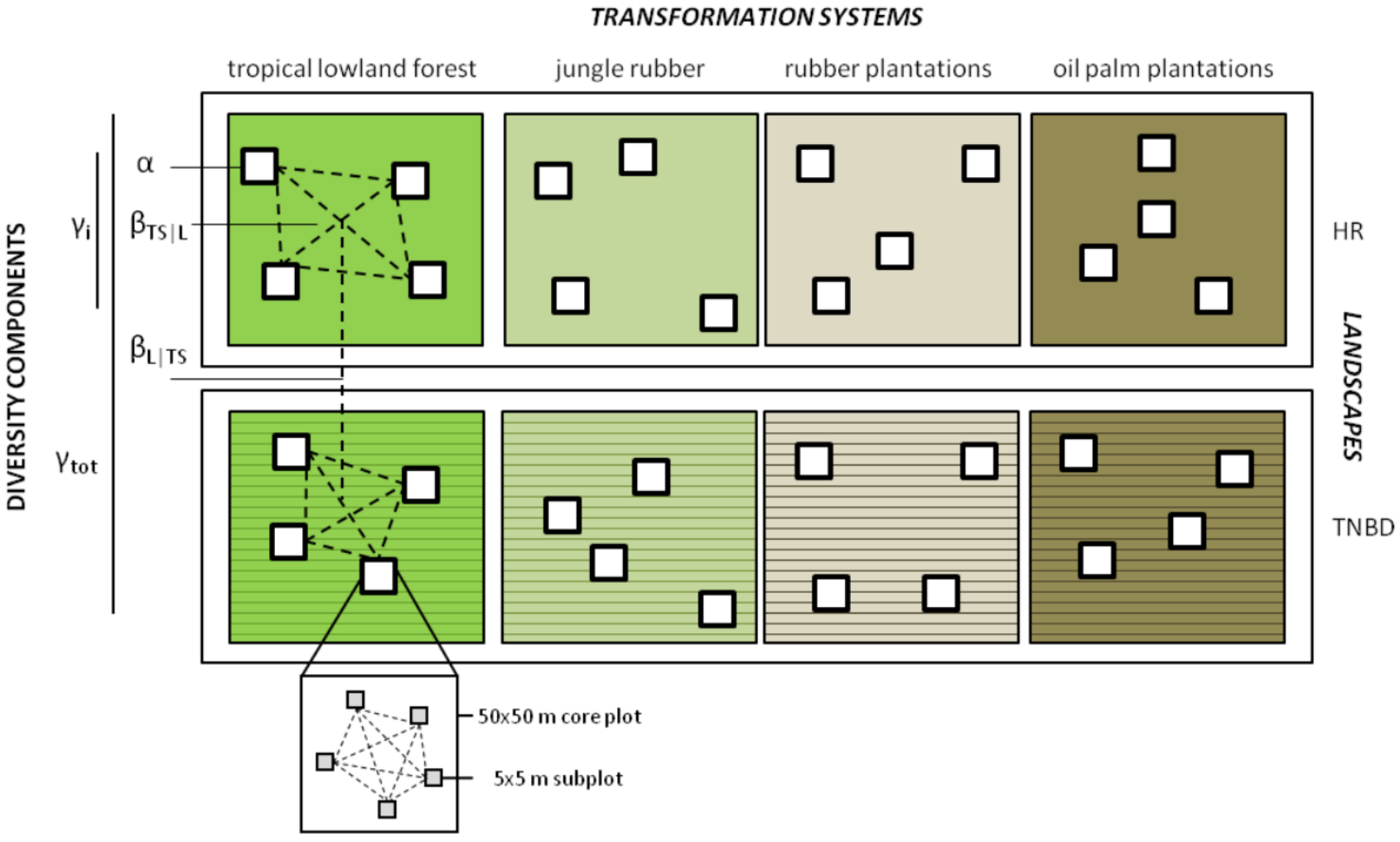
Essentially these plots and subplots are supposed to be fixed, but due to unforeseen events three core plots had to be abandoned and three new ones added.

(We found adapting BExIS to this core plot and subplot setting complicated and changes were not easily applied.)

Logistics


Jambi city
Logistic Hub,
Central Research Station
Field Research Station
Field Research Station
4 - 5 hours drive
4 - 5 hours drive
Landscape
National Park Bukit Duabelas
Field Research Station
Landscape
Harapan Rainforest
Field Research Station

Plot and Subplot Layout



Project Structure

25 projects in three Areas

- Project Group A: Environmental Processes
- Project Group B: Biota and Ecosystem Services
- Project Group C: Human Dimensions

37 Principal Investigators

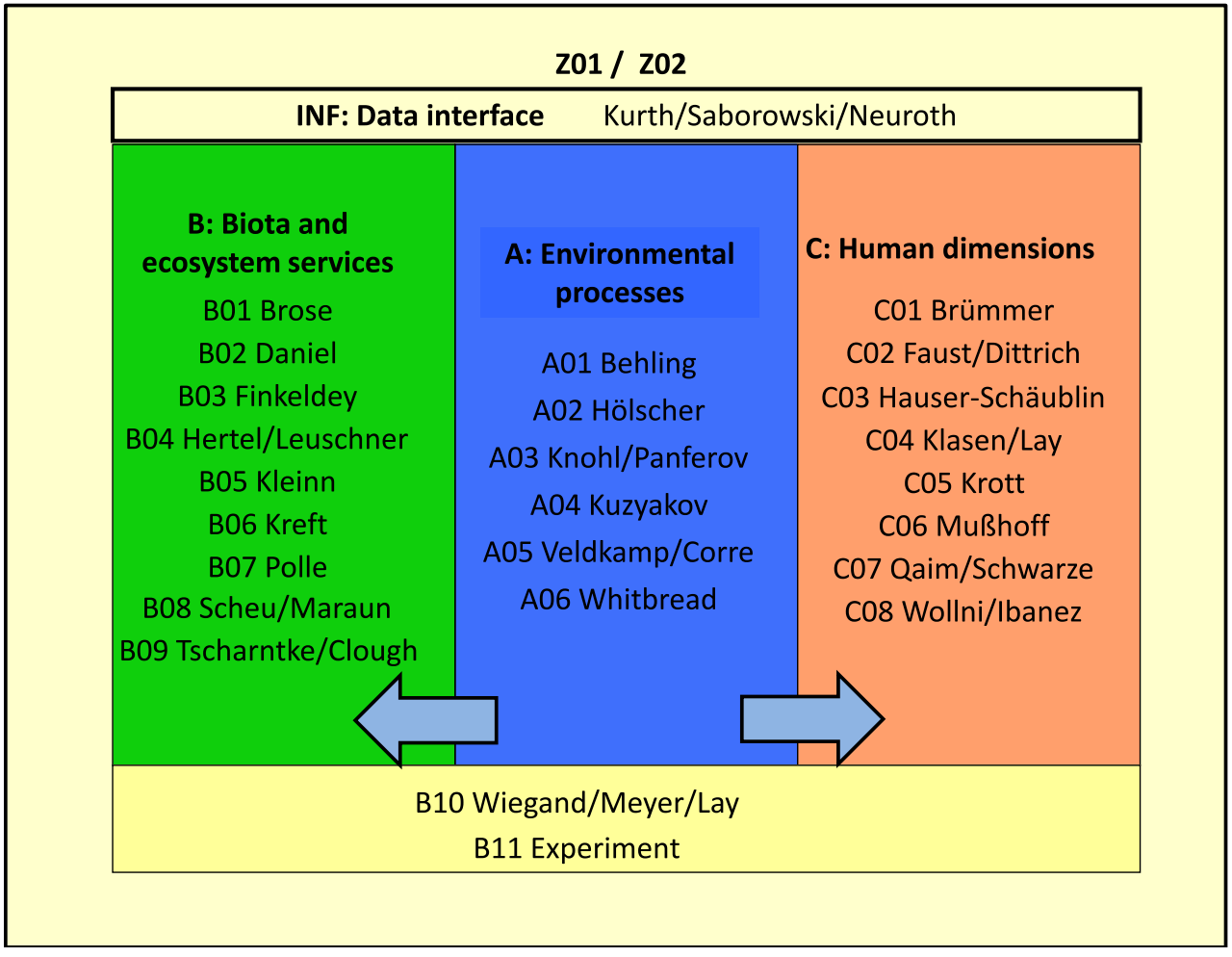
30 PhD students,

15 PostDocs.

6 technical support staff.

Altogether about 250 persons involved.

Project Structure



2. Some challenges will be not discussed here:

- Language (everything in English)
- Distance and Access:
  - Internet connections
  - Data transfer
  - Field conditions
  - Reliability
  - (Excessive VIEWSTATE information bloats data exchange)
- Culture (Questions of trust and security)
- … ?

Data challenges: Large unstructured data sets

Not clear if this can and should be changed. Examples:

1. Collection of Plant Photographs and Places (1000s; 8 GB and 31 GB resp.)
   No time to catalogue separately, and probably not useful either.
2. Collection of bird song recordings (Size unknown)
   Not to be catalogued separately (for now), goal is to identify the birds.
   "We have round the clock recordings for a whole year's time…"
3. Collection of audio recordings of sociological interviews (13 GB)
   Not to be published because of sensitive person-related information
4. Plant database with collection of more than 80.000 images
   (raw and processed: 251 GB and 209 GB resp.)

Interim solution: All considered as unstructured data. Create metadata set for whole collection, store data on separate server, access only through data management group.

For BExIS: Large BLOBs should not be saved inside data base cells but in the file system to remove size restrictions and ease the management.

Data challenges: Unwieldy structured data sets 1

Slow response

Example: Upload of meteorological data

Actual upload is fairly quick (compressed as zip in the range of 5 MB), but processing might take long (e.g. 2 hours for 215462 rows)!

No feedback in between, only "Please wait while processing...". The processes should be separated, at least for large data sets:

1. Upload data and provide positive response if completed, close upload browser window
2. Process data, import into database
3. Send message after completion (positive or negative with error message, catch possible errors!)

Hours of waiting with open browser is not acceptable!

Data challenges: Unwieldy structured data sets 2

Missing data without notification, only lack of confirmation:

The import may be incomplete, e.g. only 7381 rows of 20556 present.

Import never visually stopped, no error messages and no confirmation.

Obviously related to large data set, but not clear when or why this happens, but needs to be changed, and a clear error message is needed.

Data challenges: Unwieldy structured data sets 3

Data corrupted?

This may need further investigation, in the same data set there are not only data missing, but also different!

Upload:		Download:
…		…
30.03.2014 01:30 1002 …	⟷	30.03.2014 01:30 1002 …
30.03.2014 02:00 1002 …	⟷	30.03.2014 03:00 1002 …
30.03.2014 02:30 1001.7 …	⤫	30.03.2014 03:00 1001.7 …
30.03.2014 03:00 1001.7 …	⤫	30.03.2014 03:30 1001.7 …
30.03.2014 03:30 1001.3 …	⟷	30.03.2014 03:30 1001.3 …
…		…

Metadata challenges: Structure

The metadata in BExIS 1 consist essentially of **one complex block of data**, created using EmptyTemplate.xml for structure, testEditMetadata2.xsl as processing script and four schema description files. JavaScript is used to modify the interface (add fields etc.). Help information is included in the schema description files. This makes altogether for an extreme cumbersome situation with respect to adjusting the setup to specific needs.

Metadata description for the BExIS data sets consists essentially of the following parts:

A. Personal information (Name, address, affiliation,…)

B. Project information (Titel, PI, subject,…)

C. Description of experiment or measurement (methods used, theory, subject,…)

D. Data information (structure of data, units, variables,…)

For each researcher, A and B are essentially fixed, C may vary and D usually varies widely.

Metadata challenges: Structure 2

We need a metadata system that is flexible enough to use and reuse these building blocks. The first two part probably may only be shown, to be edited only on request.

To add a new dataset, a new description has to be created.

In BExIS 1 for this **either**

- an existing metadata set can be reused

or

- the description of the data can be read from the user's data sheet.

This is a very unfortunate alternative, since this information is really independent from each other.

Since the entry form for the variable description is quite awkward, this leads to the recommendation to read the variable information from an Excel sheet and recreate the metadata description using cut & paste.

Metadata challenges: Download and Upload

There are several reasons why download and upload of metadata should be a symmetrical process:

- to facilitates the reuse of metadata,
- to allow an efficient change of metadata using other tools than the web interface,
- to allow the transfer form one system to another.

While the download of the XML form of the metadata is essentially sufficient, the option to upload metadata is sorely missed, either for a new set or to change the metadata in place.

(Care must be taken to not destroy the structure of existent data sets.)

Metadata challenges: Interface

We found the user interface hard to handle and very hard to adapt.

The basic setup is complicated, with explanations embedded as <xsd:annotation> tags into schema definition files.

Controls were very limited, only text, text areas and selections were available. Even for a yes/no question a drop down list had to be used.

To add radio buttons (for alternatives) and checkboxes (for multiple options) we defined a complicated setting that involved

- the XML form (EmptyTemplate.xml),
- the schema definition files (schema.xsd et al.) and
- the XSL transform script (testEditMetadata2.xsl).

In similar vain we managed to hide parts of the metadata, but failed to rearrange the setting or to change the interface in response to the layout of the data sheet.

It seems that a much more flexible and easier manageable setting is needed that allows administrators an easy customization of the entry form.

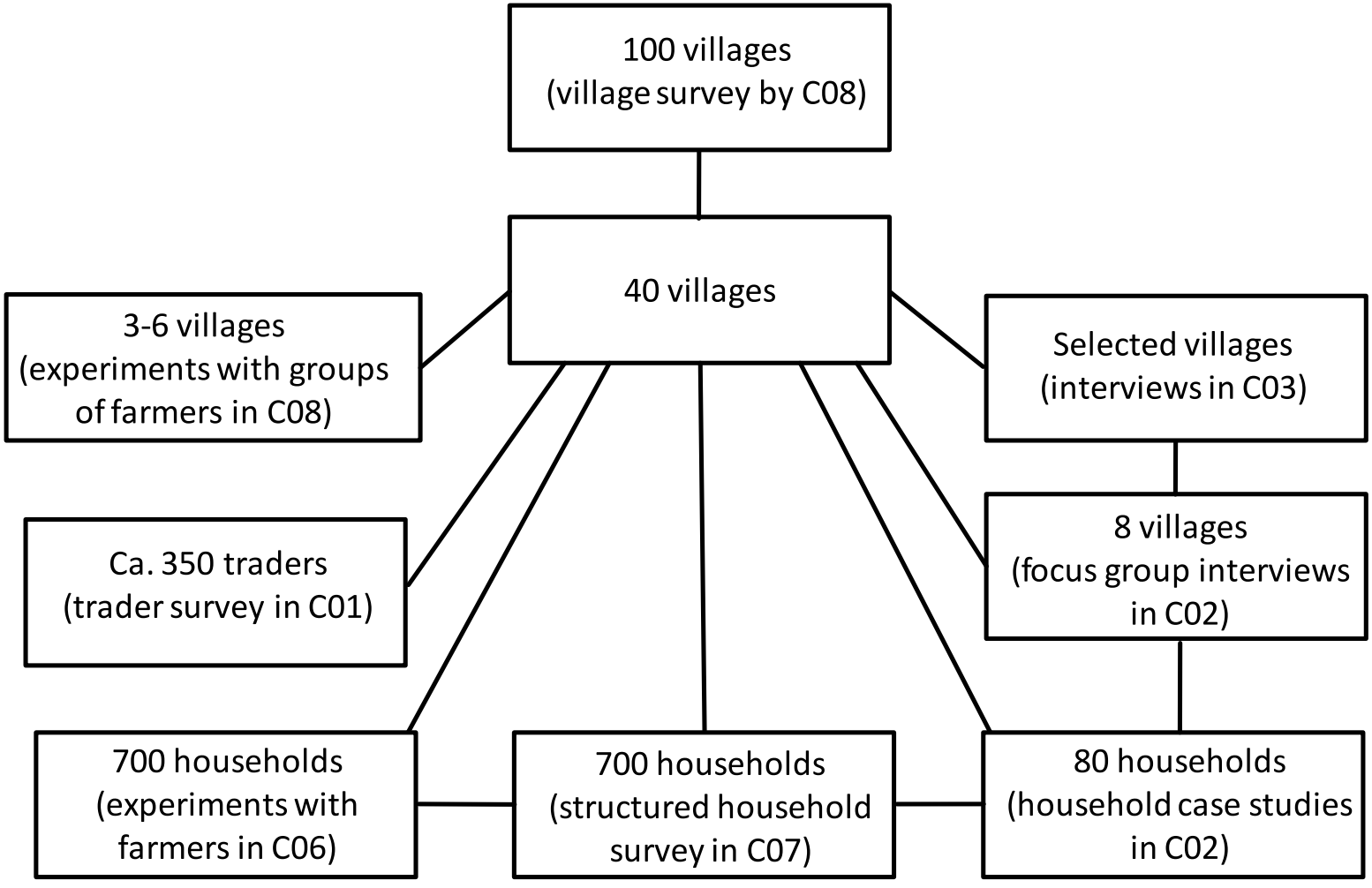
(Along the way the system proved to be very sensitive to the insertion of comments, which didn't help either.)

Metadata challenges: Contents

The wide range of topic in the CRC 990 requires quite different kinds of descriptive metadata.

E.g. the focus on biodiversity information is not well suited for the socioeconomic research in our C branch and does not fit the meteorological data and the plot diary data we use as general information sources. Thus the C part of the metadata described above has to be fully customizable, with only basic information necessary for a search across areas.

Example: Sociological Research. Taxa?



General considerations

Publication of data

We would like to make parts of the system publicly available to display the ongoing research and the progress.

- A search interface for metadata (and probably additional options to optimize the appearance of the metadata).
- Publication of data sets with persistent identifiers for reference in articles.

For this probably another entry point without focus on log in might be needed.

Usability of website

The usage of the website is somewhat hampered by the system: since all the tables use form data, you cannot open links in a new window (e.g. to compare different metadata).

Thank you for your attention!

Thomas Fischer
SUB Göttingen
Project INF in CRC 990
fischer@sub.uni-goettingen.de