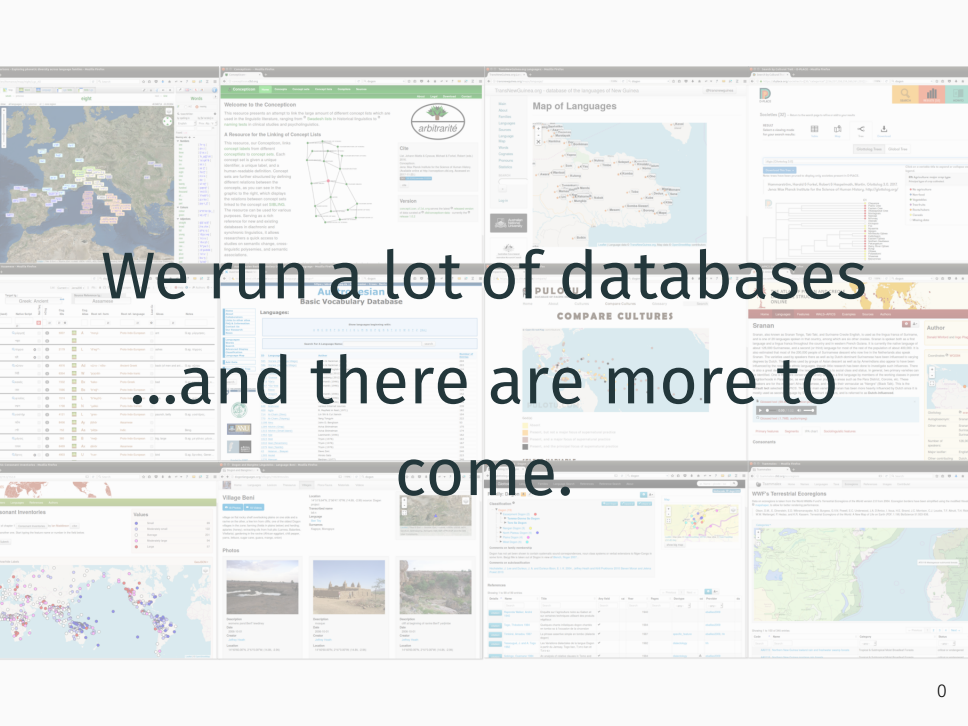# Sustainable Databases

putting the base back into basic research

Robert Forkel

Matchmaking Workshop, December 6/7, 2017, Berlin
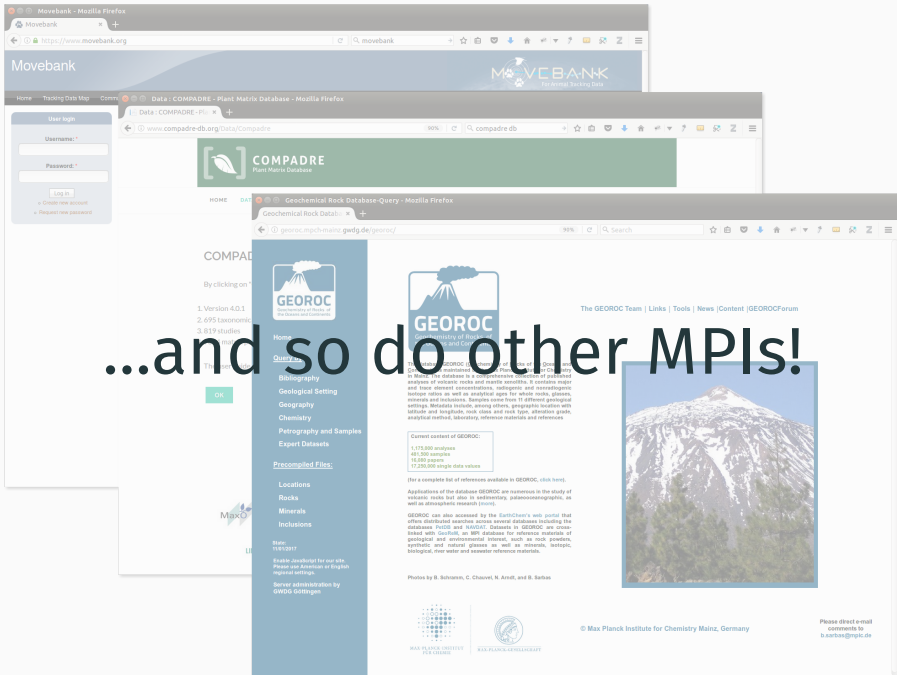
Max Planck Institute for the Science of Human History

We run a lot of databases
…and there are more to come.

- we run on databases
    - biggish collections of structured cross-cultural or cross-linguistic data
    - typically curated as sets of text files
    - released versions published online
- and we're trying to change the field
- so at some point – hopefully – the field will run on databases

...and so do other MPIs!

Clearly, databases play an important role in how MPIs conduct basic research.

Some examples:

- Compadre DB at MPI for Demographic Research contains matrix population models for hundreds of species
- Movebank at MPI for Ornithology provides animal tracking data
- GEOROC at MPI for Chemistry collects published analyses of rocks in a structured format.

Databases are research results that others may build on.

But are databases sustainable?

- Often, today, they seem to just die.
- The principle of "re-use keeping data alive" breaks down, when reuse is disconnected from data curation (e.g. happens in a different discipline)
- We still have to understand better what sustainability for databases means:
  - different service levels?
  - graceful degradation of service?
- But an Excel file stamped with a DOI should not be an acceptable vision for the future.
- There must be a middleground between bit-stream preservation and dumping bespoke web application on unsuspecting data centers.

They should be!

- The CLLD project dealt with these issues ...
- ...and made it through at least one director "turning into a pumpkin".
- Some economies of scale can be realised, but more would be possible on a higher organisational level
- Standardization helps - but isn't easy.
- Tabular data and a way to link to media files goes a long way! Some building blocks are there:
    - CSVW: Tabular data goes a long way toward generic databases ...
    - GWDG CDSTAR: ...in particular when complemented by multimedia
    - MPDL's DOI service to make databases citeable

So what?

But many issues should be addressed on MPG level, not for individual institutes:

- licensing
- domains
- service levels

other issues would profit from MPG-wide shared practices, e.g.

- a recommended technology stack
- would simplify hiring and training

So what?

Increasingly funding agencies ask for data management plans.

Correspondingly, research institutions should increasingly provide fitting building blocks for such plans.

Even better would be an approach integrating these building blocks with organisational boundary conditions.

- Working group for a data(base) management plan for the MPG?
- Do we need a **BAD** – "Beratender Ausschuss für Datenbanken"?