Reproducibility of Information Science Research

Werner Kuhn
Center for Spatial Studies
Department of Geography
University of California, Santa Barbara

Products of Research

The products of research (in information science) are:

1. **Questions** and hypotheses (i.e., statements of research problems and solution ideas)
2. **Software** (i.e., executable code in some programming language(s))
3. **Data** (i.e., evidence, in some file or database format(s))
4. **Narratives** (i.e., publications or reports, establishing the context).

The narratives guide the reproduction of some data by (often different) software.

Research Questions

Reproducing research requires **understanding** what the question(s) and tested answers were. This necessitates

- **clarity** of the questions, hypotheses, and narrative
- **testability** of the hypotheses
- preferably: **falsifiability** of the hypotheses.

Claim: reproducibility mainly fails through violations of these **standard science** requirements. Addressing this issue is likely to have most impact.

Software

Software specifies experiments. These follow requirements of regular experimental science, including specifications of

- controlled, fixed, and observed **variables**
- the experimental **method** (i.e., the code)
- the complete **environment** in which the experiment was run
- all **results** of the experiments, whether these are desirable or not.

Claim: **open source** is neither required nor sufficient for reproducibility. The programming language implementation and processor specifics are often closed or undocumented. Furthermore, executable code is almost never understandable on its own, and open source tends to come with less traceability than commercial software.

Data

Data are experimental results. They need to be

- **available** for inspection (open data)
- **interpretable**, i.e., with explicit semantics preventing misunderstandings
- **reproducible** in the sense of confirming the published results.

Claim: **open data** and **explicit semantics** are essential for reproducibility. Linked data is ideal for both, because it is the minimal data model and can make semantics explicit.

But: open data in itself, even with semantics, is **far from sufficient** for reproducibility.