Preservation of Scientific Data (in Natural Sciences)

Thomas Severiens
Institute for Science Networking
at the Carl von Ossietzky University
Oldenburg, Germany
Severiens@ISN-Oldenburg.de

Overview

- Introduction
- Primary Data
- Running Implementations and Developments
- Requirements and Status
  - Volume of Data to be preserved
  - Requirements by the Users
- Aspects of a Business Model for Preservation
- Conclusions

Introduction

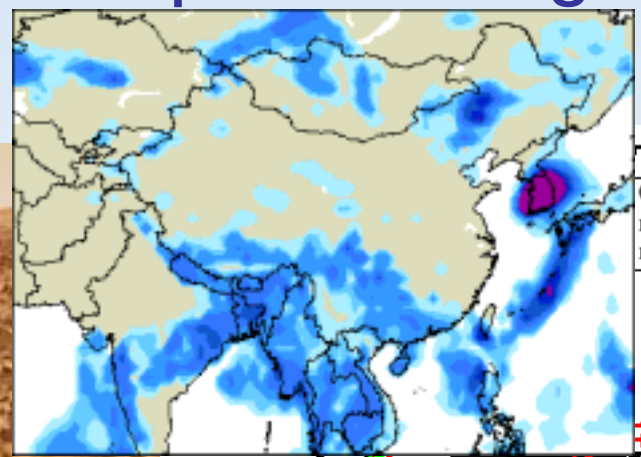
- ISN – Institute for Science Networking
- Survey on status of Preservation of Primary Data in Germany (and its neighbours)

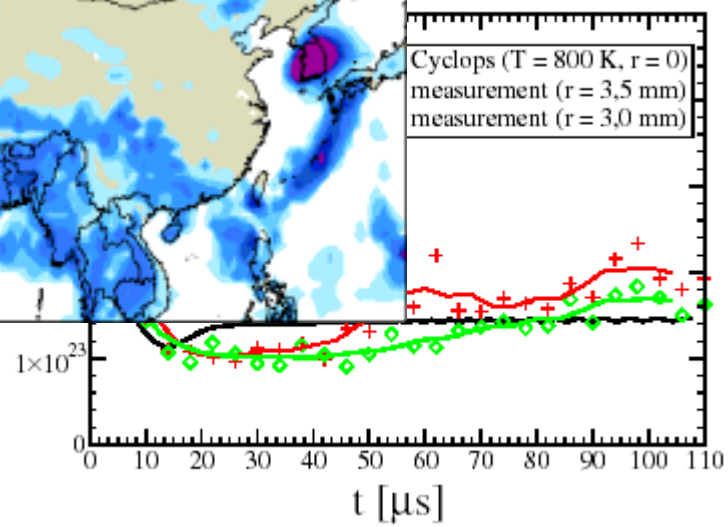
Primary Data 1

- Examples:
  - Weather observation data
  - Space observation data
  - Accelerator detector data
  - Surveys (-> next presentation by R.van Horik)
  - Data in Medicine
  - Genetic sequence data
  - Data in crystallography
  - (Sound and Video)

Primary Data 2

- ...are binary encoded streams of pure data, mostly.
- ...mostly, are not saved in XML format, but are optimized for processing in science workflow.





Primary Data 3

- ...build the basis for all scientific work and publication,
- ...are very expensive of even impossible to reconstruct
  - Neutrino flow during a supernova in our neighbourhood
  - Weather observation data
  - Measuring data form high energy colliders already broken down
  - Medicine data giving information on long term development in health of special groups over several centuries

Primary Data 4

- ...often contain information, which is still undiscovered
  - Example: Radius of the proton
- ...are the key to identify scientific falsifications of articles
  - Schön's articles would never have been published, if he had to publish the primary data of the experiments

Status of Preservation Preparation

- Volume: in Germany about 1.000 TByte every year
- Format: 70% binary encoded, 20% ASCII encoded, 10% XML or similar
- Self-description or Metadata: about 65% contain within the stream or extra file within archive
- Media: DLT (60%), DAT, CD-Rom, ...

Status of Preservation Preparation

- Access: all institutions do allow access for colleagues from science, most do not allow access for commercial reasons.
- Institutionalisation: still all institutions run their own archiving system.
- Selection strategies: not developed at all.
- Business Model: all institutions: "Preservation is of high public interest, so the government (=tax payers) should pay."

Status and Implementations

- Weather Service (Deutscher Wetterdienst): Has to preserve all data by law. Runs distributed computer pools. Offers access to the raw data, earns money with computed data.
- World Data Centers: world-wide network of (mostly) global observation institutions (52 institutions in 12 countries). Since 1956. Share data in their archives to keep it available and alive. Implemented standards and auditing system.

Conclusions 1

- Survey showed: process of shaping the awareness on requirements of and for long term preservation of primary data is in very early stage in most fields (except from astronomy, high energy physics, and global observation).
- Primary data are key use case within every preservation implementation, because here preservation helps to save much more money than it ever will cost, even on the short time scale.

Conclusions 2

- Experience, Expertise, and Standards are available on selected fields, but should be published to get their knowledge in the broad and to develop open implementations.
- Technical work of preservation is done by many institutions in parallel. In many fields without any standards at all. Synergy effects could be used for the benefit of all.
- Preservation is the job for experts on this field in co-operation with experts on the datatype.
- Preservation of primary data must be part of a globally co-operating network.

References

- Nestor project:
  - www.langzeitarchivierung.de
- Deutscher Wetterdienst (German weather service):
  - www.dwd.de
- World data centers (overview site):
  - www.ngdc.noaa.gov/wdc/wdcmain.html

Thank you for your attention!

Thomas Severiens
Institute for Science Networking
at the Carl von Ossietzky University
Oldenburg, Germany
Severiens@ISN-Oldenburg.de