Prerequisites and suitability for preservation and accessibility of research raw data

Thomas Severiens

Institute for Science Networking, Oldenburg, Germany

www.isn-oldenburg.de

Background of the presented data

- Basing on an expertise for the project „nestor“ „on the state of research data and raw data from research activities: Prerequisites and suitability for preservation and accessibility“ by Prof. Dr. Eberhard Hilf and Thomas Severiens written in 2004.
- Nestor: www.langzeitarchivierung.de
- Expertise: nbn-resolving.de/urn:nbn:de:0008-20051114018

Some words about „nestor“

- Funded by the German Ministry for Education and Research in 2003 until 2006
- Run by
  - State Library of Bavaria
  - German Federal Archive
  - Computer- and Media-Service of Humboldt-University in Berlin
  - German National Library
  - Institute for Museum Research
  - Göttingen University and State Library
- Goal:
  - Build a network of expertise in Long-Term Storage of Digital Resources

Method of the the expertise

- Online Survey sent to 327 Institutions and Persons known to be active in production of primary data (276 in Germany, 51 abroad)
- Workshop with those experts giving high quality answers or „surprising“ responses
- Interviews (by phone or along other meetings) for clarification and enrichment of the data source
- Result: 61 usable data sets

Survey

- Survey contained 8 blocks of questions:
  1. Producers vs. Consumers of primary data
  2. Genre of produced data
  3. Cooperation with external instances
  4. Provision: horizont, business modell
  5. Experience as consumers
  6. Experience with usage of old primary data
  7. National infrastructure activities
  8. Futher comments

http://www2.hu-berlin.de/nestor/questionnaire/q2.php

Results

- 73% of the 61 answers from institutions actively producing primary data
- 27% from other institutions

Results

- Genre of primary data produced (randomly ordered):
  - gene sequencing data
  - diffraction data
  - cosmological simulation data
  - images, movies
  - nmr data
  - geophysical measurement data
  - fusion & plasma data
  - solar radio spectra
  - numerical simulation data
  - population health statistical data
  - astronomical images
  - linguistic data (audio of spoken language)
  - results of questionnaires
  - marine data
  - seismic data
  - weather data
  - high energy collider data

Results

- 97.8% of primary data is stored in binary formats
- But only 91% of the data sets contain a self description, in the data sets
- For the rest: „every student of our field learns...“ or „described in published papers“ or „all colleagues know how to handle...“

Results

- 45% of the institutions share primary data with other institutions of external scientists, or would be willing to do so.
- 28% provide access only under restricting conditions: not during first 6-12 months...
- 27% of the institutions will not allow access for external consumers.

Results

- 45% of the Institutions (offering 85% of the primary data) see problems coming up from DRM usage to, resulting in restricting „self-access“ in the future.
- Overlap with those, who see requirement to include special requirements for LTA into national copyright law.

Results

- Outsourcing of LTA:
  - 60% of institutions do not see any problems
  - 40% see problem of data protection (especially from medical research and social science)

Results

- Selection criteria after 10 years:
  - 33% of institutions declare all data to be relevant for LTA
  - 20% declare none of their own primary data to be of any relevance after 10 years
  - 33% only a strict selection is of any relevance, but non of these institutions has a list of selection criteria.

Results

- Selection criteria after 30 years:
  - 20% of institutions declare all data to be relevant for LTA
  - 27% declare none of their own primary data to be of any relevance after 30 years
  - 33% only a strict selection is of any relevance, but non of these institutions has a list of selection criteria, but come up with first ideas...
    - Used in reviewed publications
    - General cultural or political interest
    - Astronomical data

Results

- Asking for a list of selection criteria, about 80% did answer not to have such a list... The other 20% did not answer this question.

Results

- 68% prefer to put the primary data into the open access after a period of 30 years (many of them prefer to publish the data much earlier!)
- 5% do will not agree to put their data into the open access (medical and social science)

Results

- 54% would prefer to have national guide lines for selection citeria and implementation etc., but several would prefer to develop guidelines as disciplines
- 20% are strictly against such guideline

- Technical and conceptional support would help regarding:
  - **File formats**, policies, user interfaces, **data management systems**, persistent identifiers, **rights situation**

Critical Outview

- No guidelines for selection available
- No separation of LTA from daily archiving
- Primary data is written onto tapes at the day of measurement, stored but not archived or long-term archived
- Expected amount of data: 1,000 – 2,000 TByte annually for Germany only.
- Noone knows, how often old data is restored...
- Missing national/european policy.