WHAT DO WE NEED TO PRESERVE ON UNIVERSITY LEVEL? ASK THE LOCAL SCIENTIST

Heike Neuroth & Stefan Strathmann
Research & Development
Göttingen State and University Library, Germany
neuroth@sub.uni-goettingen.de

iPRES 2007

Digital Preservation: Sustainable Programs and Best Practices

October 11-12, 2007, Beijing, China

www.digitalpreservation.de

nestor

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Listed on UNESCO Archives Portal

TOC

- **nestor WP5: training & education**
- Cooperation with the ETH Zurich, CH
- University of Göttingen
- Introduction to the survey
- First results
- Outlook

nestor WP5: training and education

- MoU between 8 university partners (nestor and educational institutions) from Germany, Austria and Switzerland

➢ Presentation tomorrow given by Prof. Achim Oßwald

nestor: Memorandum of Understanding

Funded by

Federal Ministry of Education and Research

Memorandum on the long-term accessibility of digital information in Germany

Digital information has become an integral part of our cultural and scientific heritage. We are increasingly confronted with scientific findings, historical events and cultural achievements presented in electronic form.

The rapid pace of technical change, however, is causing data carriers and data formats to age quickly. The result is an acute threat to the long-term usability of digital objects which serve as sources for science and research.

Long-term digital preservation allows digital objects to be accessible for use by future generations. It needs to be anchored in the social context of the national information, research and cultural policy, and the global integration of science and research. The

http://www.langzeitarchivierung.de/downloads/memo2006-e.pdf

TOC

- nestor WP5: training & education
- **Cooperation with the ETH Zurich, CH**
- University of Göttingen
- Introduction to the questionnaire
- First results
- Outlook

Cooperation with the ETH Zurich

- Cooperation between

  ETH Zurich and nestor/SUB Göttingen
- Common development of a survey regarding the Digital Preservation needs **on university level**
- Both institutions are among the biggest university libraries in their respective countries

TOC

- nestor WP5: training & education
- Cooperation with the ETH Zurich, CH
- **University of Göttingen**
- Introduction to the survey
- First results
- Outlook

Georg-August-Universität Göttingen - facts

- Founded 1737
- Employees: 13.346
- Scientific staff: 3.069
- Professors: 404
- Students: 24.607
  (All figures from 2006)

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

- 13 different departments, covering the full scientific spectrum, including e.g. natural sciences and humanities
- Related research institutions like the Max Planck Institutes (e.g. Max-Planck-Institute for Biophysical Chemistry…), the German Primate Centre
- Joint computer centre (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen GWDG)

TOC

- nestor WP5: training and education
- Cooperation with the ETH Zurich, CH
- University of Göttingen
- **Introduction to the survey**
- First results
- Outlook

The survey - aims

Examination of:

- Awareness and needs for Digital Preservation on university level
- Legal requirements, agreements, etc.
- Timeframes for digital preservation (short/medium/long time preservation)
- Affected formats
- Amount of data to be preserved
- Metadata use and needs
- Current practise and structures
- …

The survey – institutional involvement

To broaden the acceptance of the survey many relevant departments where included in advance:

- The Presidential Board of the University
- The Göttingen State and University Library (SUB)
- The joint computer centre (GWDG)
- The IT department of the medical faculty

The survey – audience

- Invitation sent by e-mail
- Sent to the faculties, the institutes, full and associate professors and a broad spectrum of individual scientific employees
- About 4.000 recipients
- 306 answers seem to be little feedback but this is due to the fact that in most cases only the head of the institute answered

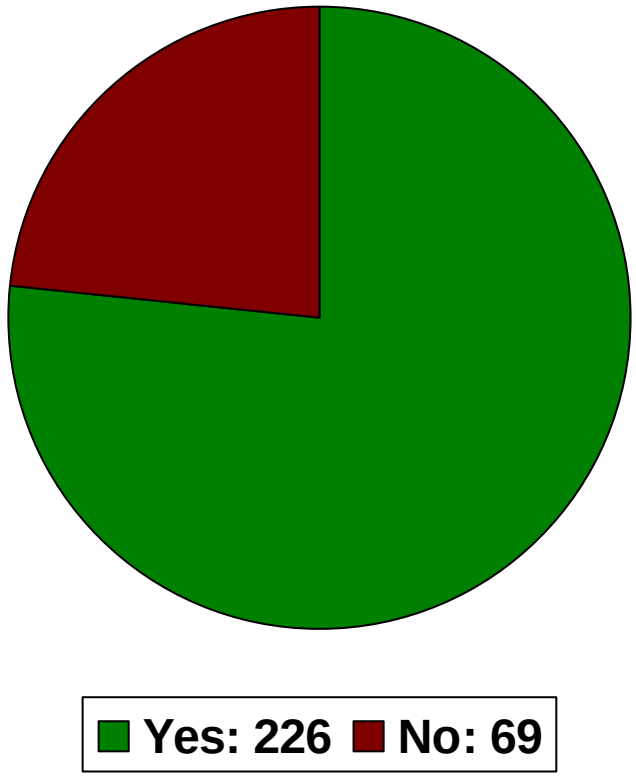
Survey - feedback

University Administration	**8**
University Institutions (examination offices, archive, language center)	**12**
Associated Institutions (Max Planck Society, German Primate Center)	**17**
Natural Science, Mathematics, Informatics	**140**
Medical Science	**27**
Humanities & Theology	**51**
Law, Economic Sciences, Social Sciences	**46**
Anonymous	**5**
Total	**306**

Survey – interest in DP

Is your department interested in DP?

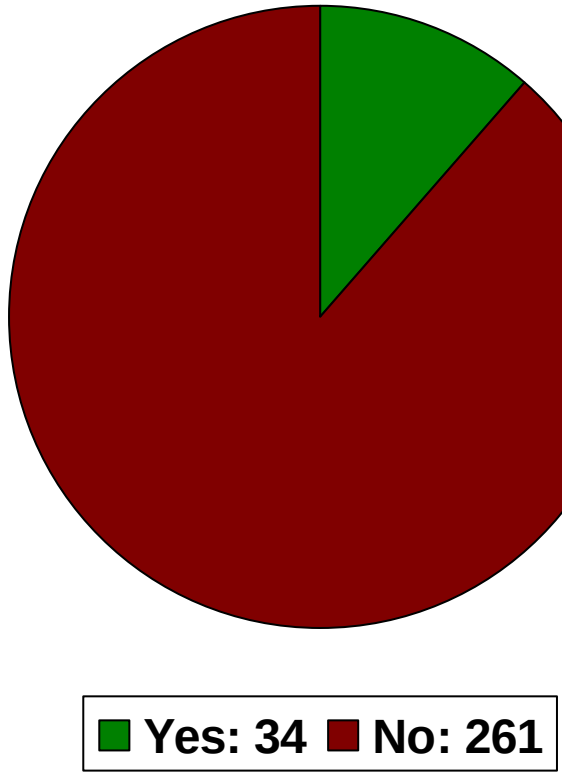
YES: 77 %



There is a huge interest in digital preservation …

Survey – DP plans

Are there any plans for the DP in your area?



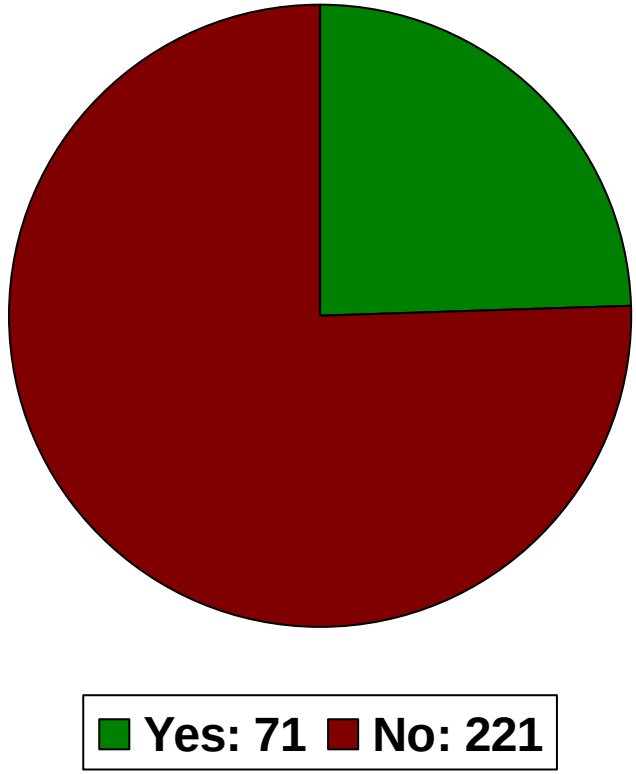
YES: 12 %

… but at the moment there is little planning!

Survey – legal requirements

Do you know about any legal requirements, agreements, etc. regarding DP in your area?

YES: 24 %



Knowledge of mandatory policies like the one from the German Research Foundation (DFG) is not very widespread …

Policy example - DFG

Empfehlung 7

Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden.

Erläuterungen

Ein wissenschaftliches Ergebnis ist in aller Regel ein komplexes Produkt vieler einzelner Arbeitsschritte. In allen experimentellen Wissenschaften entstehen die Ergebnisse, über die in Veröffentlichungen berichtet wird, aus Einzelbeobachtungen, die sich zu Teilergebnissen summieren. Beobachtung und Experiment, auch numerische Rechnungen, sei es als eigenständige Arbeitsmethode, sei es zur Unterstützung der Auswertung und Analyse, produzieren zunächst "Daten". Vergleichbares gilt in den empirisch arbeitenden Sozialwissenschaften.Experimente und numerische Rechnungen können nur reproduziert werden, wenn alle wichtigen Schritte nachvollziehbar sind. Dafür müssen sie aufgezeichnet werden.

Jede Veröffentlichung, die auf Experimenten oder numerischen Simulationen beruht, enthält obligatorisch einen Abschnitt "Materialien und Methoden", der diese Aufzeichnungen so zusammenfaßt, daß die Arbeiten an anderem Ort nachvollzogen werden können. Wiederum gilt Ähnliches in der Sozialforschung mit der Maßgabe, daß es immer mehr üblich wird, die Primärdaten nach Abschluß ihrer Auswertung durch die Gruppe, die die Erhebung verantwortet, bei einer unabhängigen Stelle zu hinterlegen.

Deutsche Forschungsgemeinschaft

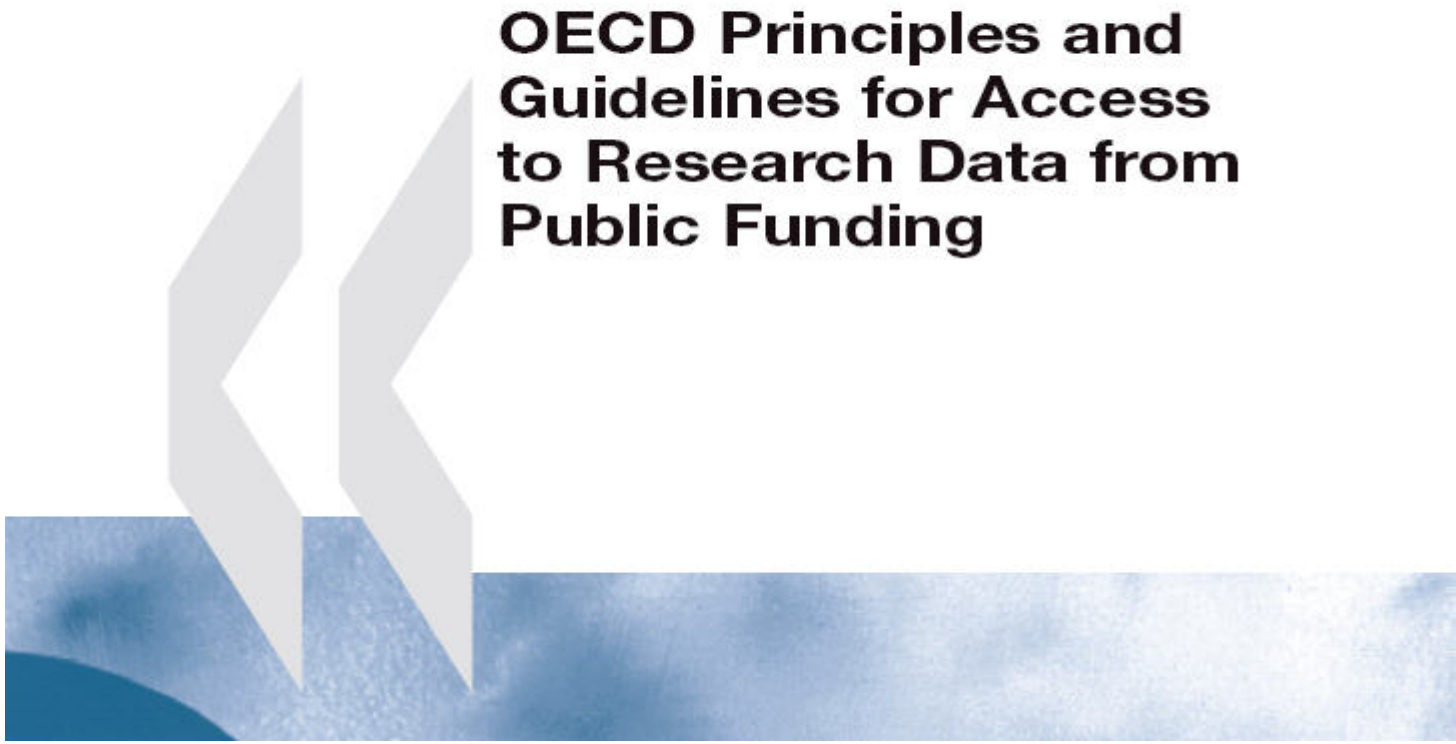
http://www.dfg.de

DFG

Recommendation n.7 regarding preservation of primary scientific data, part of *policy for saveguarding good scientific exercise* issued by German Research Foundation, DFG (1998)

http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

Policy example - OECD



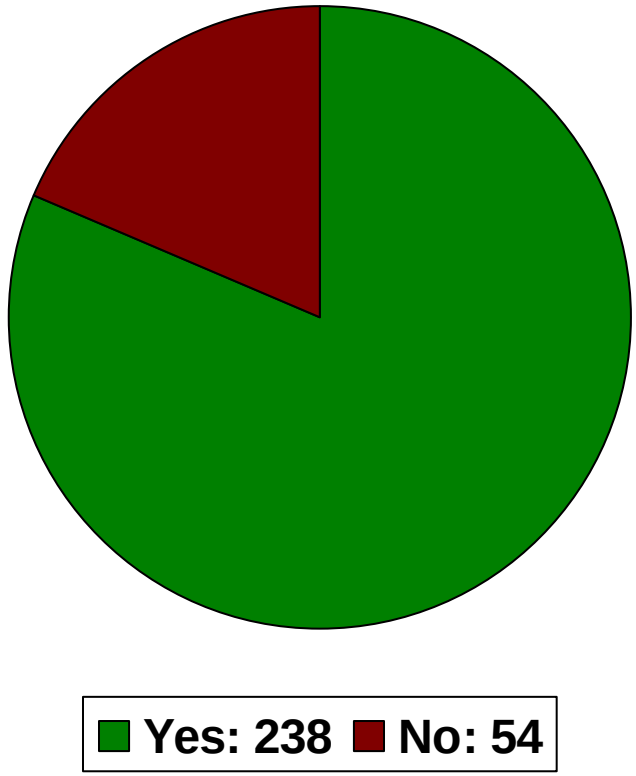
[2007]

http://www.oecd.org/dataoecd/9/61/38500813.pdf

Survey – DP needs

Are there digital data which should be preserved for the long time in your area?

YES: 82 %



The majority has data for DP
Most institutes declare a need for long-term DP

Survey – DP timeframes

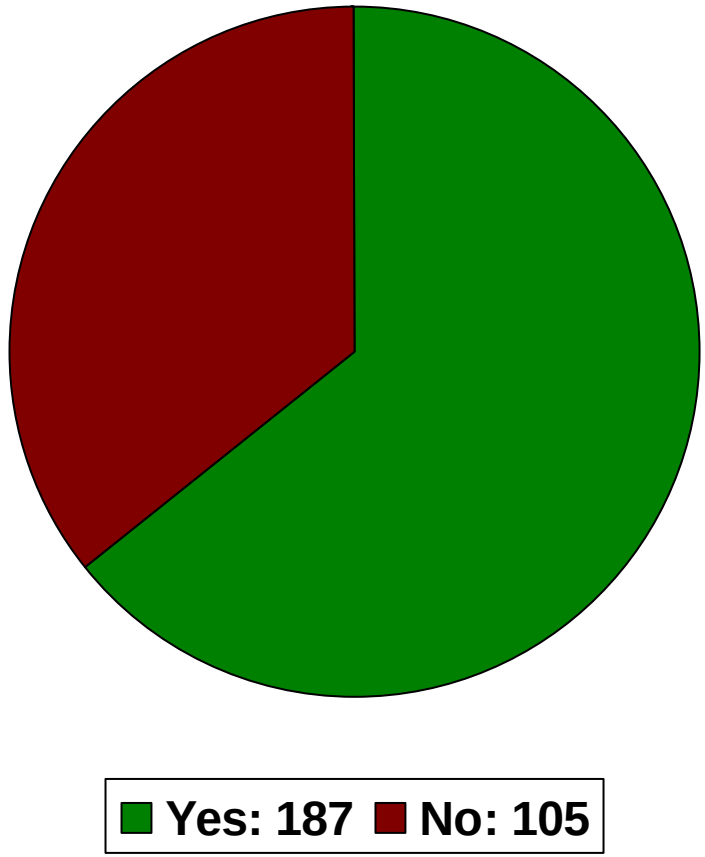
- Most of the digital data are meant for long-term preservation
- Despite that, there is a preferred period of time depending e.g. on the format

Examples

- 5 years and less: websites
- 5 - 10 years: programs
- more then 10 years: databases, primary data

Survey – standard formats

Are the data which should be preserved in standard formats? Are the formats well known and/or documented in the subject community?



YES: 64 %

It seems that nearly two-third of the data are in a „well known“ format (e.g. PDF, TXT, PPT)

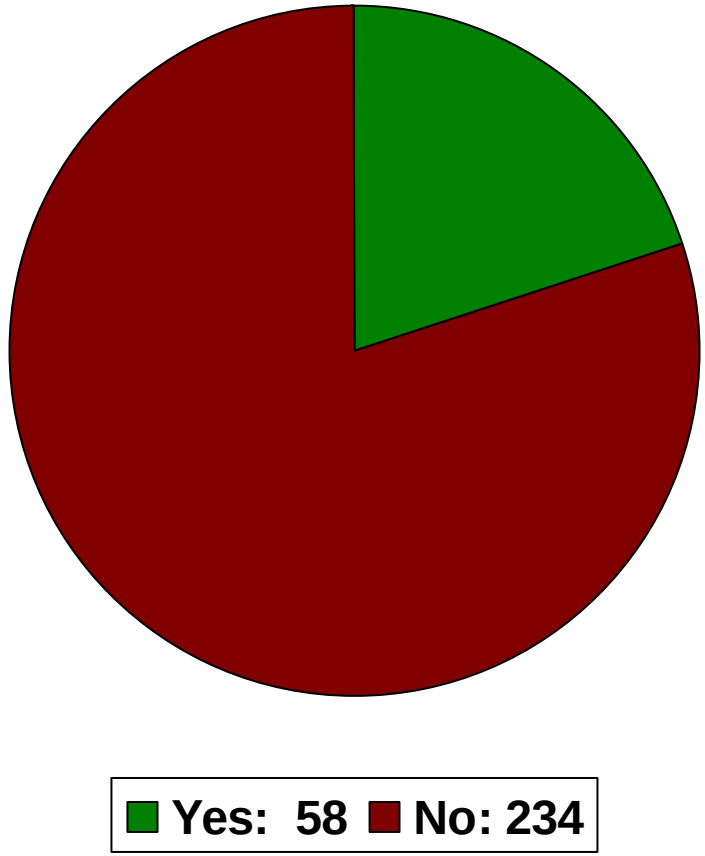
Survey - formats

- In total, 96 different formats were mentioned:
  - Very common formats are widespread across all subjects
    (PDF, JPG, TIFF, TXT, PPT ...)
  - Other formats regarding special subjects are mentioned only very rarely
    (spatial data, GIS data, remote sensing data, ...)
  - Several proprietary formats are in use
    (Brain Products, MAQSIMA, Turboveg, ...)

Survey – technical metadata

Are there any technical metadata about the data?

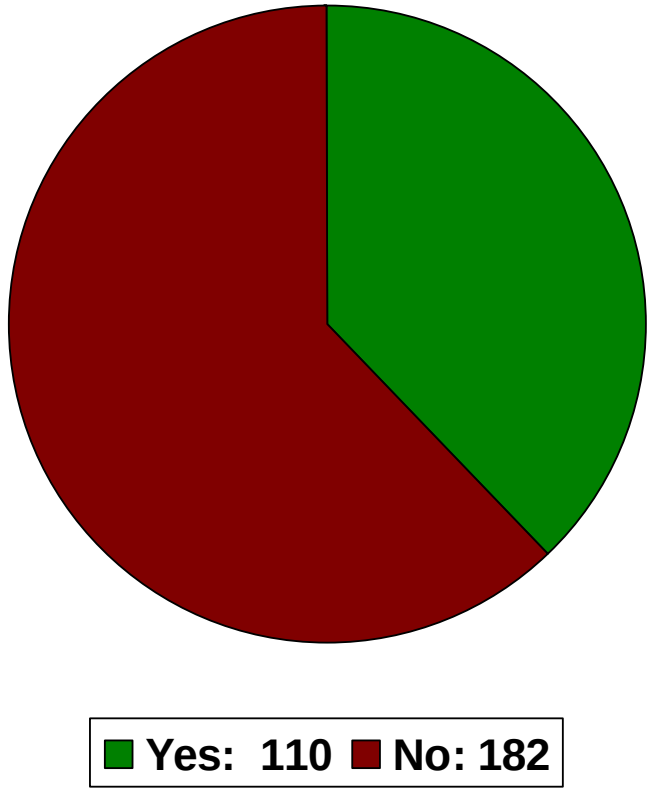
YES: 20%



Most files have no technical metadata included …

Survey – descriptive and administrative metadata

Are there any descriptive or administrative metadata about the data?



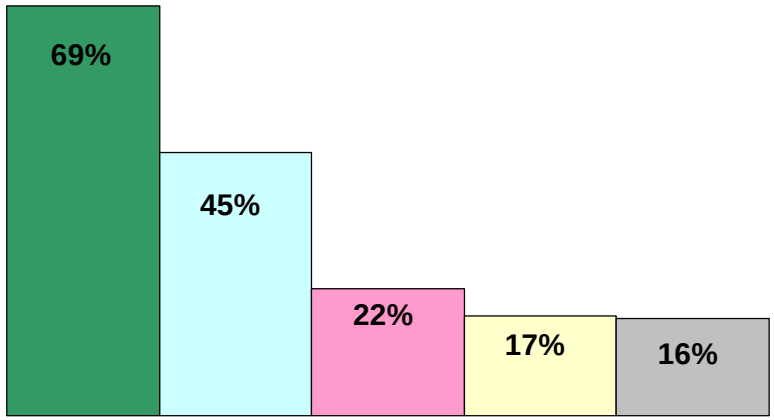
YES: 38 %

… and descriptive data is often missing, too

➡ High effort is needed to add them!

Survey - administrative descriptions

- ***By which means do you or your staff members currently ensure the availability of your/their data?***



- ***Staff-members themselves are responsible for their data.***
- ***When leaving, staff members have to indicate, where and in which form their data is stored.***
- ***For the group / the institute / the department there exist clear guidelines on the safeguarding of data.***
- ***We have appointed a responsible person for the group / the institute / the department who takes care of the safeguarding of data which will be relevant in the long run.***
- ***Additional information or other approaches***

Actually, the staff themselves are responsible for preserving the data …

Survey – amount of data

- The amount of data which needs to be preserved is hard to predict because the interviewees could give only rough estimates
- The data for long-term preservation amounts to **several hundred terabytes**
- Most of the data is generated in natural science, mathematics and informatics

Survey – amount of data

	< 5 years	5-10 years	> 10 years
Biology	12 TB	10 TB	4,3 TB
Physics	***9 TB***	***102 TB***	***102 TB***
Medicine	5,3 TB	3 TB	0,3 TB
Geology	4 TB	0,5 TB	6,2 TB
Forest Sciences and Forest Ecology	3,3 TB	3,3 TB	2,5 TB
Chemistry	1,8 TB	0,3 TB	1,2 TB
Humanities, Social Sciences	1,5 TB	1,5 TB	1,5 TB
Other	1,8 TB	1,2 TB	0,2 TB
Total	**38,7 TB**	**121, 8 TB**	**118,2 TB**

TOC

- nestor WP5: training & education
- Cooperation with the ETH Zurich, CH
- University of Göttingen
- Introduction to the questionnaire
- **First results**
- Outlook

Survey – first results

- There is a huge interest in digital preservation…
- … but as for now there is little planning
- Knowledge of mandatory policies like the one from German Research Foundation (DFG) is not very widespread …
- The majority has data for DP
- Most files don‘t have technical metadata…
- … and descriptive data is often missing, too
- In most cases the staff is responsible for preserving the data …
- ***There is no common policy!***

Survey – first results

- There is an urgent need for DP at university level
- There is a great discrepancy between the estimated need and the practical implementation of DP
- The knowledge about the legal requirements, agreements, etc. is not widely spread
  All scientists have to sign an agreement about data curation for 10 years – but don’t refer to that

Survey – first results

- The formats in use are very heterogeneous and often not well documented
- There is a general lack of metadata (technical MD, administrative MD and descriptive MD)
- The metadata used are often not standardised

Survey – first results

- A sustainable practise in DP is missing!
  - The data preservation/curation depends on the staff itself
  - Only very few departments have clear rules/agreements regarding the preservation issues
  - In case of leaving the department the staff often has to declare where/how he stored his data

TOC

- nestor WP5: training & education
- Cooperation with the ETH Zurich, CH
- University of Göttingen
- Introduction to the questionnaire
- First results
- **Outlook**

Outlook

- Detailed study about the preservation needs at Göttingen University
- Iterative development of central DP and data curation services for the Göttingen University
- Preservation planning must consider all aspects of a university (administration, research, teaching etc.)
- A working Group with members of all units (including scientists) of the university needs to be established
- **There is a huge need for a (decentral) preservation officer!**

Outlook

- A roadmap has to be developed, considering various aspects
  - scientific
  - legal
  - political
  - financial
  - technical
- It has to be agreed upon by all key players at the university

Thank you very much for your attention!

Comments? Questions?

Heike Neuroth & Stefan Strathmann
Research & Development
Göttingen State and University Library, Germany
neuroth@sub.uni-goettingen.de

Acknowledgment:
Pascalia Boutsiouci, ETH Zurich (Switzerland)
nestor Team at SUB Göttingen