

**<philtag>-Workshop vom 13/14 October 2008 in Trier  
"Communicating eHumanities: Archives, Textcentres,  
Portals"**

# Research Data Curation in eHumanities



*Excellent Information Services for Excellent Research*



Dr. Heike Neuroth  
Max Planck Digital Library Berlin  
Göttingen State & University Library  
[neuroth@sub.uni-goettingen.de](mailto:neuroth@sub.uni-goettingen.de)

# Motivation

- **TextGrid** - A Community Grid for the Arts and Humanities  
[Federal Ministry of Education and Research, BMBF]



- **eHum** - National eHumanities Infrastructure Project  
[German Research Foundation, DFG]

Deutsche  
Forschungsgemeinschaft

**DFG**

# TextGrid Goals

- Establishment of a “**Virtual Research Infrastructure**”
- TextGrid: a **generic platform** for scientific text data processing
- **Service Grid** - Toolbox for collaborative work
- **Data Grid** - Virtual archive(s) for data curation



# TextGrid Tasks

- Roadmap for community building (difficult!)
  - (Business/Organizational) models for sustainability
  - New functionality:
    - new forms of collaboration and cooperation
    - new, reliable standards; easy exchange of data and research results; new forms of linking up researchers, services, and content
    - new impulses on the conduct of research
    - long-term preservation of research data
- **TextGrid: contributes to the preservation and the keeping alive of our Cultural Heritage**

# National eHumanities Infrastructure

- **Developing Germany's eInfrastructure for the humanities**
- **Conceptualizing in 2008**
  - Analysis of over 30 initiatives, organizations, projects (e.g. Digital Humanities, European Resource observatory for the Humanities and Social Sciences, JISC e-Infrastructure-Programme 2006-2009 ...)
- **Goals and main objectives**
  - to explore the current provision of Germany's eInfrastructure for the humanities
  - to help define its future development

# Digital Curation as part of eInfrastructure

- Data Curation can be defined as follows:  
The activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future.' Data curation can also include managing vast data sets for daily use; updating it to keep it readable, etc.  
<http://www.dcc.ac.uk/FAQs/data-curator>
- Discovery & re-use of eHumanities research data

# Research Data

- Means data as facts, observations or experiences on which an argument theory or test is based. Data may be numerical, descriptive or visual. ...  
[records.curtin.edu.au/recordkeeping\\_manual/glossary.html](https://records.curtin.edu.au/recordkeeping_manual/glossary.html)
- Includes laboratory notebooks, as well as any other records that are necessary for the reconstruction and evaluation of reported results of ...  
[www.egr.unlv.edu/CAS/glossary.htm](http://www.egr.unlv.edu/CAS/glossary.htm)
- Data that is related to a published research finding that was produced under a federally sponsored award to a nongovernmental entity, that is used by the federal government in the development of an agency action that has the force and effect of law as it relates to property. ...  
[www.twc.state.tx.us/business/fmgc/fmgc\\_appa\\_glossary.doc](http://www.twc.state.tx.us/business/fmgc/fmgc_appa_glossary.doc)

# Situation in Germany

- *"[Für Forschungsdaten] sehen alle Wissenschaftseinrichtungen einen dringenden Handlungsbedarf hinsichtlich der systematischen Sicherung, Archivierung und Bereitstellung dieser Daten für die Nachnutzung durch Dritte."* (Allianz-Initiative Digitale Information, Juni 2008, [http://www.mpd.l.mpg.de/news/Allianz\\_Schwerpunktinitiati.pdf](http://www.mpd.l.mpg.de/news/Allianz_Schwerpunktinitiati.pdf))

- ALEXANDER VON HUMBOLDT-STIFTUNG
- DEUTSCHER AKADEMISCHER AUSTAUSCHDIENST
- DEUTSCHE FORSCHUNGSGEMEINSCHAFT
- FRAUNHOFER-GESELLSCHAFT
- HELMHOLTZ-GEMEINSCHAFT DEUTSCHER FORSCHUNGSZENTREN
- HOCHSCHULREKTORENKONFERENZ
- LEIBNIZ-GEMEINSCHAFT
- MAX-PLANCK-GESELLSCHAFT
- WISSENSCHAFTSRAT

# But how, because ...

- Which data?
- Quality Control?
- Who is responsible?
- Do we have an infrastructure: technical, organisational? Who is paying?
- Can we trust repositories, infrastructures?  
Sustainability?

# Which Data?

- Huge range of heterogenous data: video, sound, text, 3-D objects, images, raw data ...
- Huge range of heterogenous file formats: not yet standardized formats like ISO Standards in earth sciences, lot of proprietary formats ...
- All data? Selection mechanismen needed ... - which data are valueable ...not only today but also for next generations ...
- ...

# Quality Control?

- Format Validation
- Metadata: minimal requirements for context (bibliographic) information, technical and preservation metadata
- Research data: DFG evaluators as mechanismen? If evaluation process is positive DFG might be able to finance curation ...
- ...

# Who is responsible?

- Scientist for selection of data which need to be preserved
- Data Curator for organizing whole process
- (Technical) Provider for data hosting
- Funder for long-term policy (e.g. evaluators agree to spend additional money for curation of data)
- Publisher/Aggregator for making data accessible
- ... Data Centres, Institutions, and USER ...

See also UK research data feasibility study, July 2008:

[http://www.ukrds.ac.uk/UKRDS%20SC%2010%20July%2008%20Item%205%20\(2\).doc](http://www.ukrds.ac.uk/UKRDS%20SC%2010%20July%2008%20Item%205%20(2).doc)

# eInfrastructure?

- Organizational: Network, loose cooperation, contracts, identification of main players (scientists, content, standards experts, technical experts, data curators ..)
- Financial: cost estimation (e.g. also for consulting?), role of funders, ...
- Others: guidelines, best practice, recommendations (e.g. TEI baseline encoding)
- Partnership: Close cooperation between humanitists and infrastructure people – to establish a new research culture ....

# Trust?

- Sustainable infrastructure/network: guarantee?  
E.g.:
  - Skills of persons involved
  - Sustainable institution (libraries are there since several hundreds of years, data centres like AHDS?)
  - Longterm funding, not depending on one funding strategy
  - Investment in having and building more/new expertise (technical, organisational, general in curation)
  - Appropriate quality mechanisms, external audit and certification method? Zertificate?
  - ...

# Scientists

- Why motivation for „open data“
  - Scientists might loose controll over „*their* data“?
  - When is the right time for publication of research data?
  - Time consuming to „curate“ data, scientists have no time
  - Impact factor for research data?
  - Incentive system for „open data“?

See also „Open Access 2.0: Freier Zugang zu Forschungsdaten“,  
9/10 October 2008 in Berlin

<http://open-access.net/de/austausch/openaccesstage/programm/#c856>

# First Ideas

- Need of a national policy, across disciplines (including „hard“ sciences) and special policies for disciplines in humanities
- Cooperative network among humanitists
- Identification of competence centres (standards, technology, data types like text/video etc., content provider, tools/services, resource providers for storage/CPU/tools etc., ...)
- ...

# Where we are: Time is perfect now!

- Good chance in Germany, „Allianz Digitale Information“ will publish a first general policy end of this/beginning of next year
- National working group „Research Data“ with representatives from DFG, MPG, Helmholtz, Leibnitz, HRK ...
- IT infrastructure is in place: German Grid Initiative (but?)
- International coopertion: is excellent (Interedition, Dariah, TextGrid, DFG/NEH etc.)

# <philtag>-Workshop vom 13/14 October 2008 in Trier "Communicating eHumanities: Archives, Textcentres, Portals"

**Thank you for your attention!  
Questions, comments?**



*Excellent Information Services for Excellent Research*



Dr. Heike Neuroth  
Max Planck Digital Library Berlin  
Göttingen State & University Library  
[neuroth@sub.uni-goettingen.de](mailto:neuroth@sub.uni-goettingen.de)