Big Data Challenges in the Humanities and Social Sciences

Peter Doorn
Director, Data Archiving and Networked Services (DANS)
The Hague, The Netherlands
peter.doorn@dans.knaw.nl

Session: Humanities & Social Sciences Applications
Volumes of data production in the humanities and social sciences
  - Data produced by people, social processes, and digitization
  - Small, semi-big and big data
Not only ‘big data’ (size) challenges, also methodological, preservation and access challenges
  - Data not *intended* for research
The grid and e-scholarship in the HSS
  - promises, results, deceptions
Examples:
  - Comparing classification in Wikipedia and the UDC
  - Archiving data on the grid
Removing barriers to e-scholarship
Traditional data production in the humanities and social sciences (HSS)

- **Humanities**
  - archaeology: excavations and surface surveys
  - history and cultural studies: digitized/transcribed
    - archival sources
    - library holdings (books and other texts, with images)
    - museum holdings (artworks, images with descriptions)
  - linguistics: human speech (audio/video + transcriptions)

- **Social and behavioral sciences**
  - social sciences:
    - social surveys
    - qualitative interviews (audio/video + transcriptions)
    - censuses and registration data
  - psychology: experiments

*Driven by data*
Big data production in the HSS

• Born digital
  • administrative processes: government administrations
    • taxation, population registers, school data, traffic flows
  • commercial processes: business and financial transactions
    • banking, sales (goods, real estate), stock exchange
  • socially produced: social networks
    • Twitter, Wikipedia, Facebook, YouTube, Flickr
  • personal devices: GSM, GPS

• Mass digitization
  • images
  • OCR of images: text & numbers
  • audio-visual

Driven by data
SSH: big data challenges

- Data generated by people tend to be small
- Data generated by social processes, transactions, administrations and personal devices tend to be big
- Growing number of big data projects in SSH
  - Millions of digitized books ("Culturomics")
  - Analysis of twitter feeds:
    - Sentiment analysis to predict markets and economic trends
    - Linguistic analysis
  - Traffic flows using GPS

Driven by data
However, more typical for the humanities is “semi-big” data

- Collaborative work: bringing together work from many scholars
  1. Historical shipping
  2. Historical censuses
  3. Clio-Infra
  4. Holocaust studies
  5. Dendrochronology

Driven by data
1. Historical shipping

Bringing together shipping records from projects over the decades: South Chinese Sea Trade (1681-1792); Dutch-Asiatic Shipping (1602-1795); Climate of the World Oceans (weather observations from ships’ logs, 1750-1854), Atlantic Connections, Trans-Atlantic Slave Trade, etc.
2. Historical Censuses since 1795

- Census digitization projects since 1996
- Collaboration with Statistics Netherlands
- 40,000+ pages of tables turned into numbers
- Images of the original source books
- Up to 60,000 visitors per year
3. Clio-Infra: historical data on worldwide economic growth & inequality

- Data collection from thousands of sources from all over the world by hundreds of specialists
- Solving massive problems of data interpretation, cleaning, linking, harmonization, comparison...

From source to database: example on age data about Ceylon, 1770

Jan Luiten van Zanden
4. Holocaust studies

Holocaust Researchers Catalog 42,500 Nazi Ghettos, Camps; Numbers Are 'Unbelievable'
5. Digital Collaboratory for Cultural Dendrochronology

Data collections of ‘old wood’ for The Netherlands

- Private sector in The Netherlands (6000 BC-present):
  - > 2000 research projects
  - > 20,000 measurement series of 13,000 trees (60% dated)

- Private sector and universities in Germany:
  - Archaeology: e.g. Dorestad
  - Cultural heritage: many objects from The Netherlands and Flanders
  - Architectural history: North and East NL, Amsterdam
Big Data challenges in the HSS

- Getting access to the data
  - privacy issues
  - commercial interests
- Big Data sources are not intended for research
  - what do the data represent or measure?
  - how to match traditional to new Big Data?
- Methodological
  - theory-driven questions
  - data-driven possibilities
- Size and computational challenges
  - scholars in the HSS are working in cooperation with e-scientists
  - Long-term archiving of Big Data

Historians are used to working with sources not intended for research, but ‘Big Data’ is 21st century…
“BIG GRID is crucial to the success and continuity of many Dutch research communities, covering important areas such as life sciences, astronomy, particle physics, meteorology, and climate research, water management, to name just a few.

However, the very nature of the new infrastructure, a multidimensional collaboration enabler and accelerator, allows for direct participation of also social sciences, humanities, and even addresses communities in administrative domains, like digital academic repositories.”
European Research Infrastructure projects in SSH about grid

- **CESSDA**: “grid technologies for facilitating the merging of distributed data sources”
- **DARIAH**:
  - “grid services for an open semantic architecture facilitating arts and humanities research
  - need for ‘easy’ interfaces for humanities scholars, services need to be usable without the complexities of the grid infrastructure”
- **CLARIN**: grid technology for
  - “access to guidance and advice through distributed knowledge centres
  - access to repositories of data with standardized descriptions, processing tools ready to operate on standardized data”
Germany: TextGrid

- Grid-enabled workbench to process, analyse, annotate, edit and publish XML-encoded textual data for academic research
- Connect to the D-Grid Integration Platform (DGI) via TextGrid-specific middleware components
- Demonstrate the efficiency of the grid-enabled tools in the areas publishing, processing, retrieval, and linking
- Semantic TextGrid: semantic methods for processing text assets, and for interweaving texts and dictionaries
Germany: TextGrid

TextGrid VRE: Repository + Lab

Driven by data

But also results...
“NCeSS investigates how innovative and powerful computer-based infrastructure and tools, developed under the UK e-Science programme, can benefit the social science research community”

- Examples of grid-projects:
  - Mixed Media Grid (MiMeG): generate tools and techniques for social scientists to analyse audio-visual qualitative data and related materials collaboratively
  - SABRE software has been specifically designed for the statistical analysis of multi-process random effect response data, using parallel processing
This website is archived and is not maintained. For up to date information on eResearch, go to our NEW Manchester e-Research Centre Site, MeRC http://www.merc.ac.uk
Dutch Big Data project from e-humanities

- Subject: organization of knowledge
- Comparison of designed classification system (UDC) with a socially grown knowledge system (Wikipedia)
- Multidisciplinary research group, including DANS researcher Andrea Scharnhorst
- Big social data set (dump of Wikipedia: 2,8 TB)
  - Mine the data to extract the page and category link changes over time
  - Create complex visualizations
- Computational support by BiG Grid team: Tom Visser, Coen Schrijvers and Ammar Benabadelkader

Cheng Gao
Krzysztof Suchecki
Andrea Scharnhorst
Alkim Almila Akdag Salah
• Data collection and data processing ➔ awareness of the value of preserving data for re-use:
  • for validating the results of earlier research
  • for comparative analysis
  • answering new research questions with existing data (secondary analysis)

• Emergence of data archives:

<table>
<thead>
<tr>
<th>Era</th>
<th>Social Science Data Archives</th>
<th>Text Archives for Linguistics and Literary Studies</th>
<th>Historical Data Archives</th>
<th>Archaeology Data Archives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960s</td>
<td>ICPSR, ZA, UKDA Steinmetz</td>
<td>Oxford Text Archive</td>
<td>NHDA, HDS, IPUMS</td>
<td>ADS, EDNA</td>
</tr>
<tr>
<td>1970s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1980s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Archiving and Networked Services in The Netherlands

- Created in 2005, combining the forces of:
  - Steinmetz Archive for the social sciences (1964)
  - Netherlands Historical Data Archive (1989)

Institute of Dutch Academy and Research Funding Organisation (KNAW & NWO) since 2005

Mission: promote and provide permanent access to digital research information

First predecessor dates back to 1964 (Steinmetz Foundation), Historical Data Archive 1989

Driven by data
DANS services

**EASY:** Electronic Archiving System for self-deposit

**Data Seal of Approval**

**NARCIS:** Gateway to scholarly information in The Netherlands

Persistent Identifier
URN:NBN resolver

Driven by data
A visual representation of the DANS data collection

21303 unique datasets

2527 datasets assigned to multiple categories:

Source: DANS EASY per 2012-01-20

Driven by data
Datasets in DANS EASY according to size and access

Number of datasets according to size

- 1.8% of datasets > 2 GB
- 2.8% of datasets > 1 GB

Datasets according to access

- 37% Open
- 49% Closed
- 12% Restricted
- 3% Group

Driven by data
Typical for HSS: large numbers of relatively small datasets and files

- Currently 25,000 data sets in DANS archives
  - Every data set consists of 1+ data files, up to 25,000+
  - Most data sets are small (97% < 1 Gb)
  - For example, the entire population census of 1960 (>11 million records) fits on one CD-ROM (< 700 Mb)
- Total number of files 2 million
- Total storage volume ca. 70 Tb
- Archiving is more than storing: management!
- Archival management of large volumes of files: slow and problematic on normal servers
  - Copying of the whole archive to the grid is not trivial
  - Mass conversions (e.g. thumbnails of images)
  - Data integrity control (checksums)
  - Compressing the data

Driven by data
Archiving experiment on Big Grid

- Grid middleware not very suitable for our archiving purposes
- Use case:
  - How can you be sure that what you store on the grid is valid?
  - Giving proof of data integrity is a requirement of ISO standard 16363 for trusted digital archives
- Advantages of grid storage:
  - Fast access to grid worker node
  - Hierarchical storage manager: eg. efficient automated backup procedures
  - Shared facility is efficient and economically attractive
Grid-archiving experiment

- Experiment with five digital archives, containing a total 290,341 files, grouped over a total of 1695 .tar-files of 5 GB each (c. 8.5 TB)
- Carried out by Jan Just Keijser (NIKHEF)
- Three-phase workflow:
  - Create checksums
  - Create tarballs (.tar files)
  - Upload tarballs to the grid
  - Download .tar file
  - Compress it to a .tar.gz file
  - Upload compressed tarball
  - Download and unpack .tar.gz file
  - Calculate checksums
  - Send checksums back and compare

Trust is essential - one checksum mismatch detected: disk failure on grid worker node!
Conclusion: removing e-Humanities and e-Social Science barriers

- Acceptance of grid technology by HSS community is low and slow: “my laptop has enough processing power”
- Grid is still perceived as “complicated”
- Researchers are not aware of:
  - data management issues
  - the research potential of “Big HSS Data”
- Demonstrator projects are still needed:
  - Social scientists need to focus more on the analytical potential of “Big Social Data”
  - “Culturomics” in humanities
- DANS can help to make that accessible, although we are not only driven by data, but also by... demand!
Thank you for your attention

- peter.doorn@dans.knaw.nl
- www.dans.knaw.nl