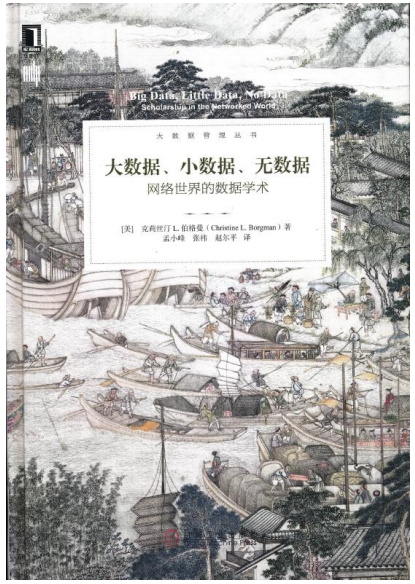
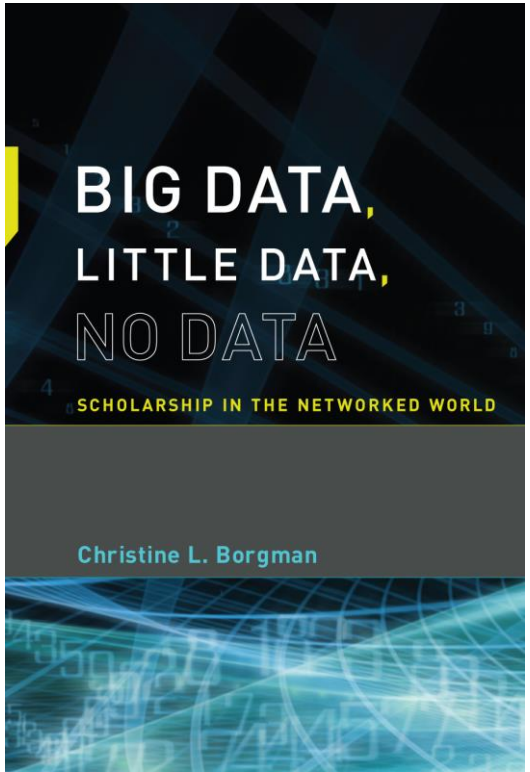
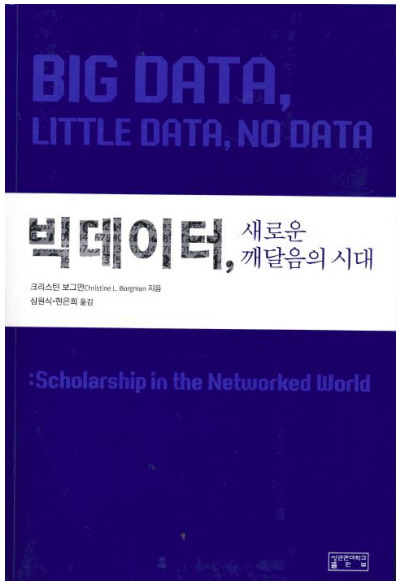
Big Data, Little Data, or No Data?
Scholarship, Stewardship, and Humanities Research

Christine L. Borgman
Distinguished Research Professor & Presidential Chair in Information Studies, Emerita
Director, UCLA Center for Knowledge Infrastructures, @scitechprof

Centre for Contemporary and Digital History (C²DH)
University of Luxembourg, **8 April 2021**







A call to action for the humanities (2009)

- What are data?
- What are the infrastructure requirements?
- Where are the social studies of digital humanities?
- What is the value proposition for digital humanities?

Royce Hall, UCLA

Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly*, *3*. http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Two cases

1. Whose text, whose mining, and to whose benefit?

Borgman, C. L. (2020). Whose text, whose mining, and to whose benefit? *Quantitative Science Studies*, *1*(3), 993–1000. https://doi.org/10.1162/qss_a_00053

2. Digital data archives as knowledge infrastructures

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888–904. https://doi.org/10.1002/asi.24172

What are data?

Open Access / Open Data Policies

- European Research Council
- Research Councils of the UK
- Australian Research Council
- U.S. Federal research policy
- Individual countries, funding agencies, journals, universities

National Science Foundation
WHERE DISCOVERIES BEGIN

JAPAN SCIENCE AND TECHNOLOGY AGENCY

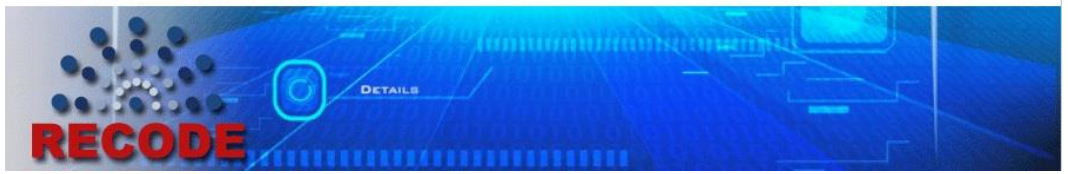
中央研究院
ACADEMIA SINICA

Australian Government
National Health and Medical Research Council

erc European Research Council
Supporting top researchers from anywhere in the world

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Policy RECommendations for Open Access to Research Data in Europe



Data Archiving and Networked Services

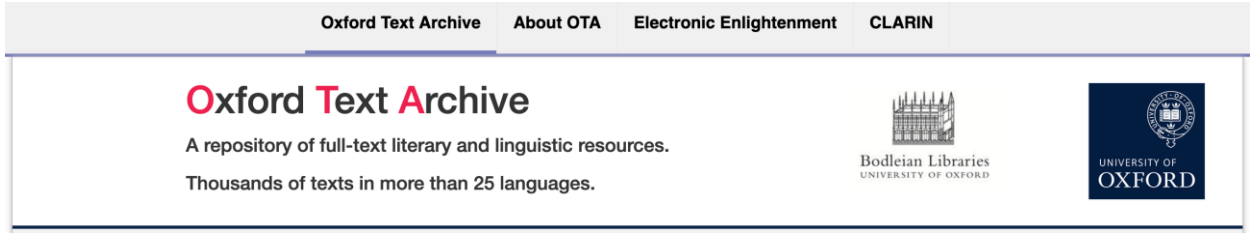
DANS

Open Data Practices

HARVARD Dataverse

- Link datasets to journal article or publication
- Deposit datasets in a digital data archive
- Publish data documentation
  - Research protocols
  - Codebooks
  - Software
  - Algorithms
- Cite data and software

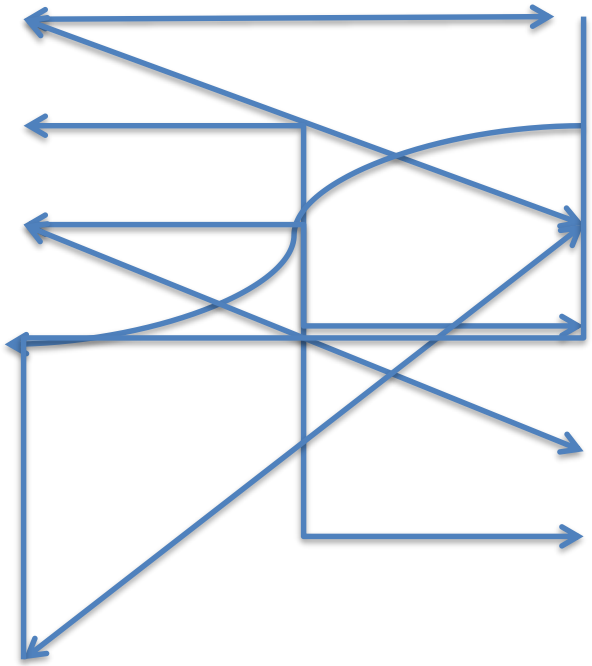
ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
A PARTNER IN SOCIAL SCIENCE RESEARCH



UNIVERSITY OF CALIFORNIA | dash

Publications <–> Data

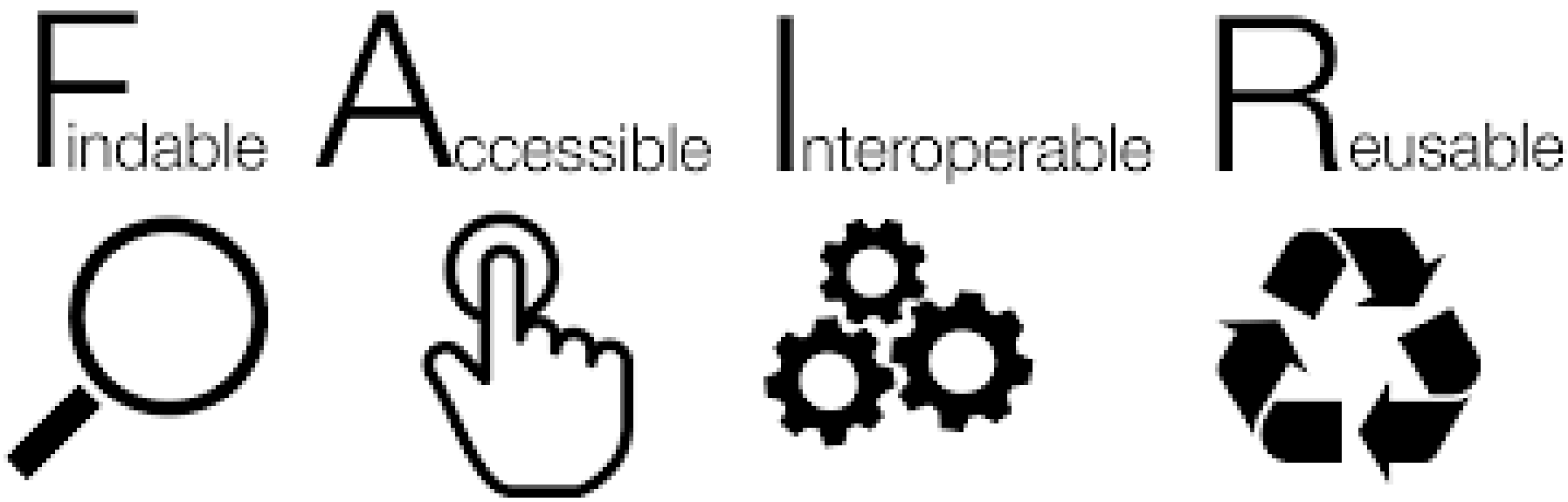
- Article 1
- Article 2
- Article 3
- Article 4



- Article n

- Dataset time 1
- Dataset time 2
- Observation time 1
- Visualization time 3
- Community collection 1
- Repository 1

Data Stewardship: The Ideal

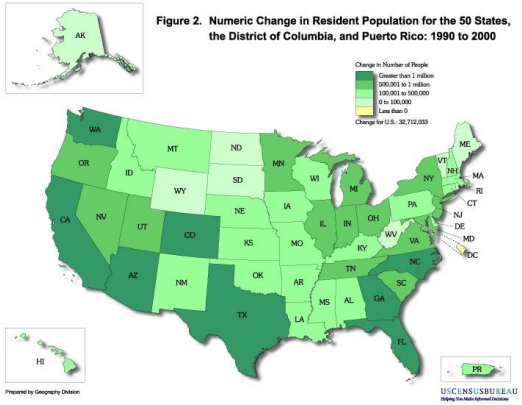


Wilkinson, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, http://dx.doi.org/10.1038/sdata.2016.18

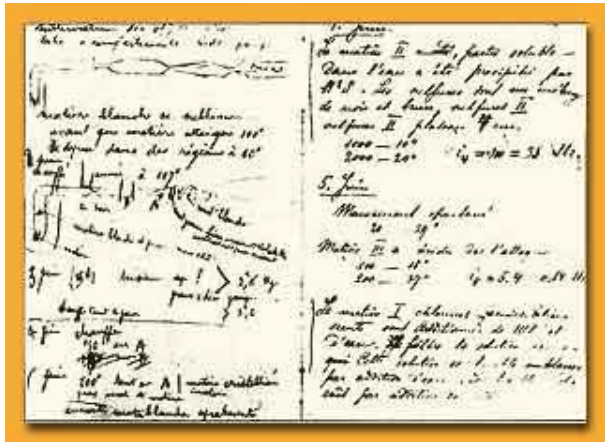
8

Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

Pisa Griffin



http://www.census.gov/population/cen2000/map02.gif

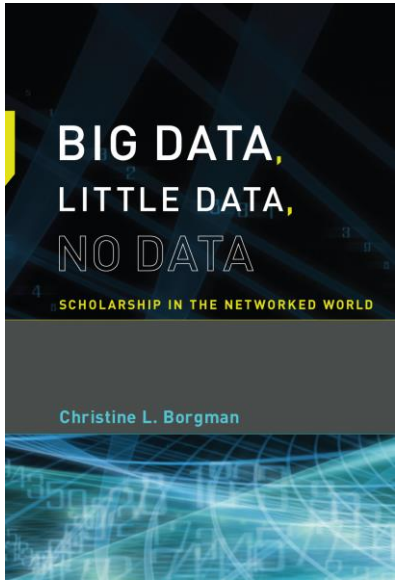


Marie Curie's notebook aip.org

Date:1/2.07.75 Place:Sakaltutan
Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. eS came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. {much money went} Has a tractor.

Date:July1980 Place:Sakaltutan
Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuß; one with a driver from Süleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin değil. { not sharp - i.e.? not profitable} I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak {dolmuß stop} from Beledive and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php



What are knowledge infrastructures?

Knowledge Infrastructures

"Robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds"

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

http://knowledgeinfrastructures.org

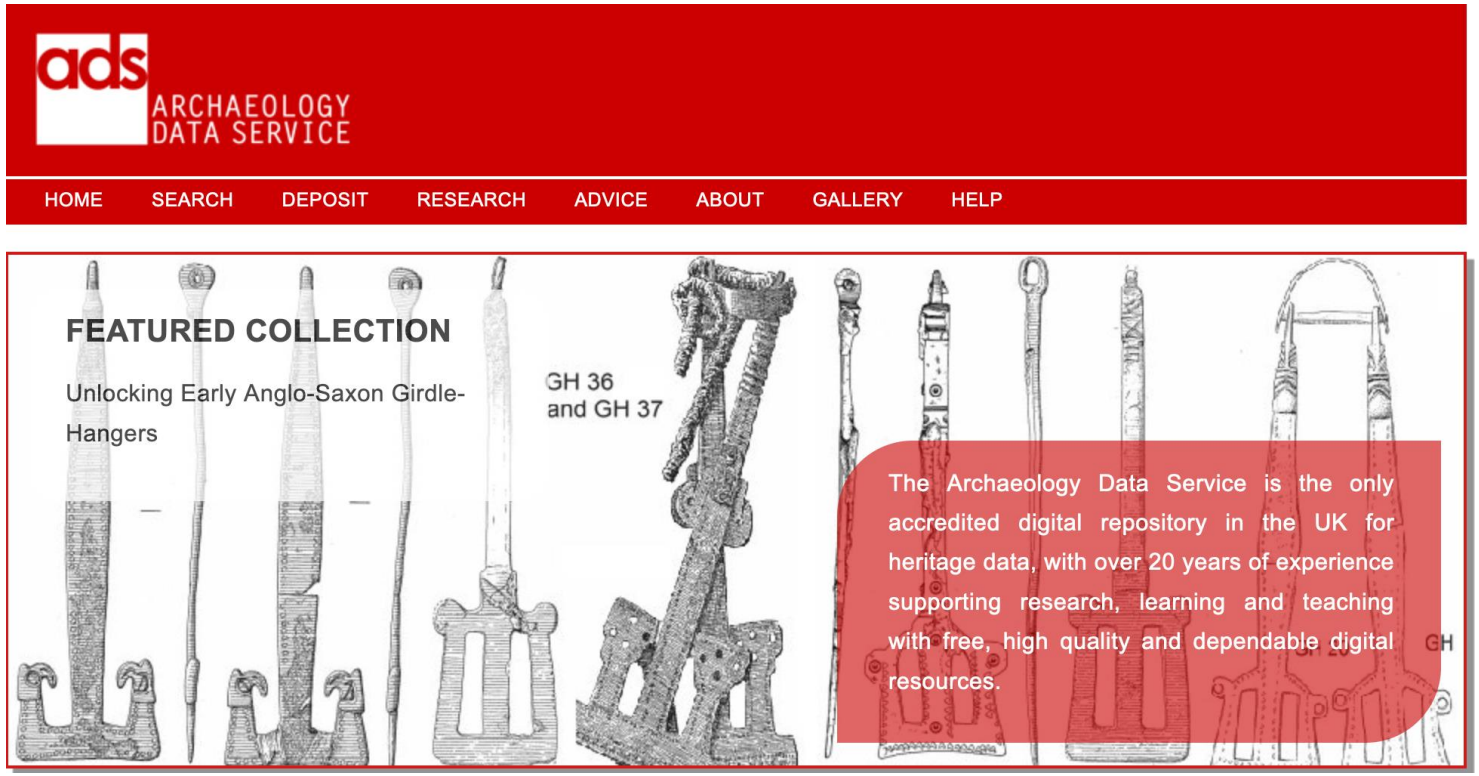
Case 1: Data Mining

Whose text, whose mining, and to whose benefit?

Borgman, C. L. (2020). Whose text, whose mining, and to whose benefit? *Quantitative Science Studies*, *1*(3), 993–1000. https://doi.org/10.1162/qss_a_00053

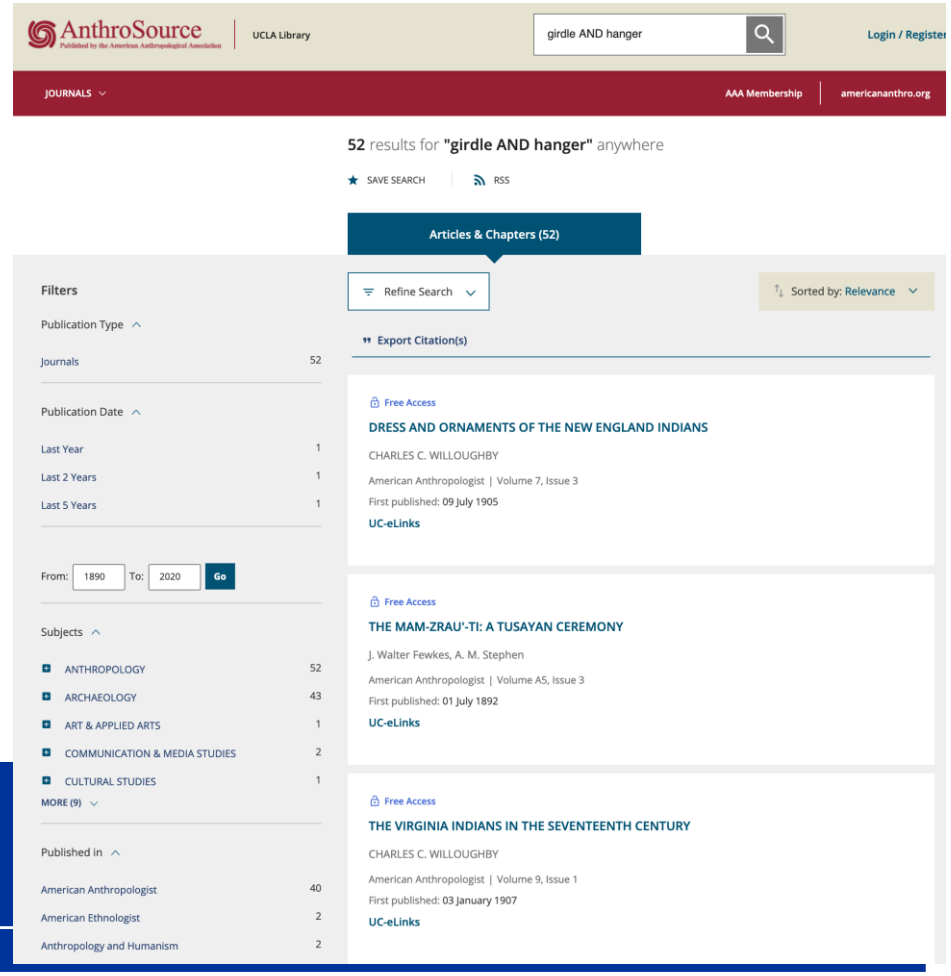
Text Data Mining in the Humanities

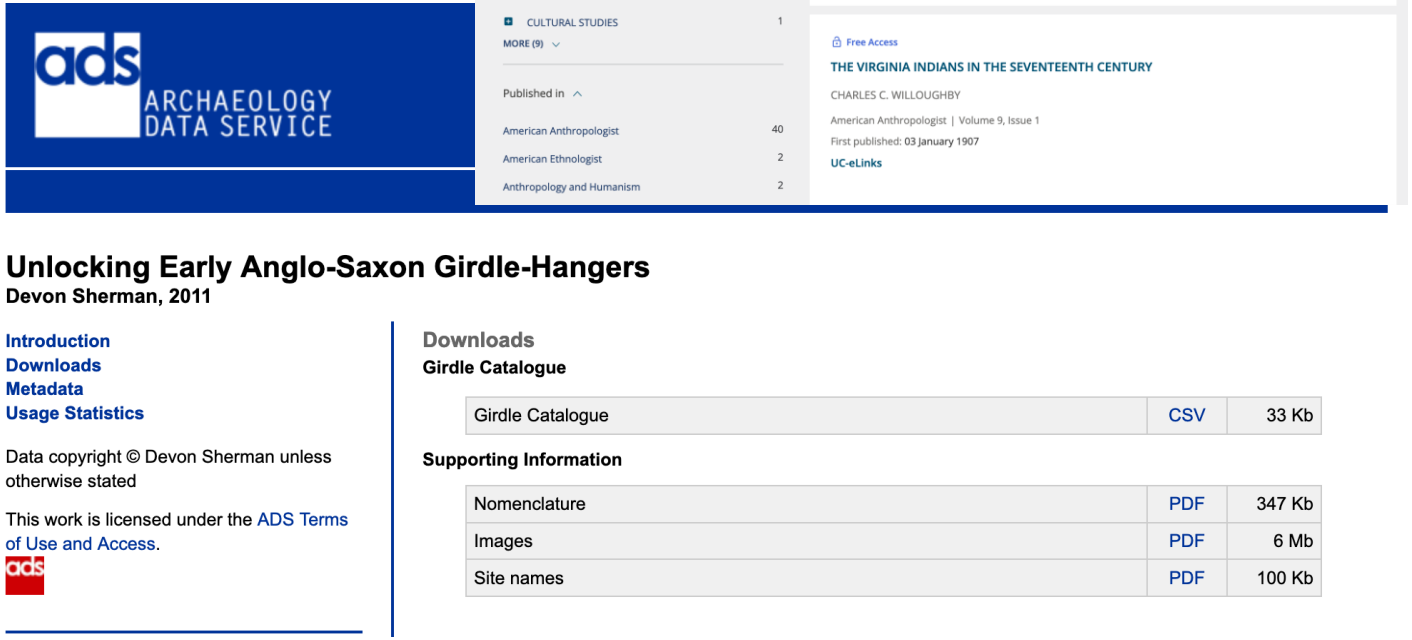
- Research problems
  - Historical
  - Cultural
  - Linguistic…
- Data sources
  - Publisher databases
  - Archives
  - Administrative records
  - Open corpora
  - Scraping websites…



Searching vs. Mining

- Searching scholarly content
  - Keywords, authors, titles
  - Retrieve and read
  - “Consumptive use”
- Mining scholarly content
  - Query to extract dataset
  - Retrieve and analyze
  - “Non-consumptive use”

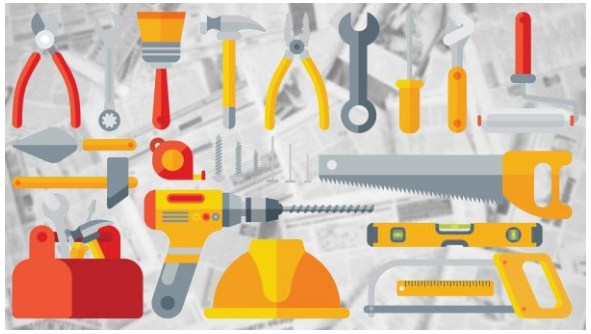




Locus of mining methods

- Local, client-side
  - Download dataset
  - Compute with local tools
- Remote, server-side
  - Submit query to database
  - Compute on server
  - Download analysis
  - Privacy preservation
  - Intellectual property protection

https://miro.medium.com/max/1000/1*7oOowGYxS7CvhcXTguRPyw.png



https://www.poynter.org/wp-content/uploads/2018/12/newstools.png

https://upload.wikimedia.org/wikipedia/commons/8/8a/ContractLaw.jpg

Stuart Miles: FreeDigitalPhotos.net

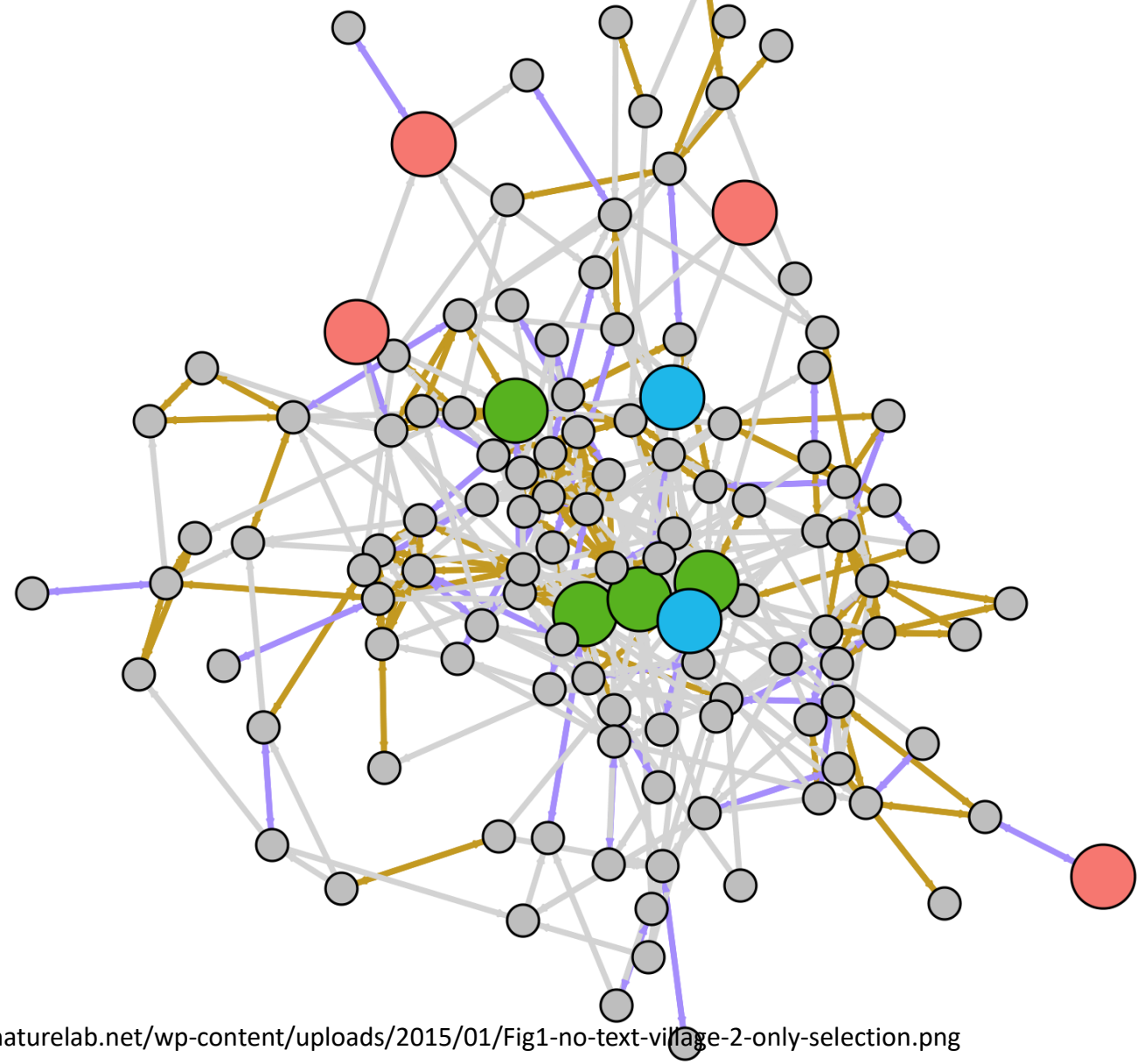
https://miro.medium.com/max/1000/1*7oOowGYxS7CvhcXTguRPyw.png

University Data Resources

- Library databases
- Primary scholarly materials
- University archives
- Research data stores
- Academic personnel dossiers
- Learning management systems
- ID cards: library, health, recreation, dorms, food service, credit, debit, transportation…
- Community outreach
- Conferences and workshops
- Food services
- Staff personnel
- Sensor networks
- Security cameras
- Network traffic
- Street traffic …

Borgman, C. L. (2018). Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier. *Berkeley Technology Law Journal*, *33*(2), 365–412. https://doi.org/10.15779/Z38B56D489

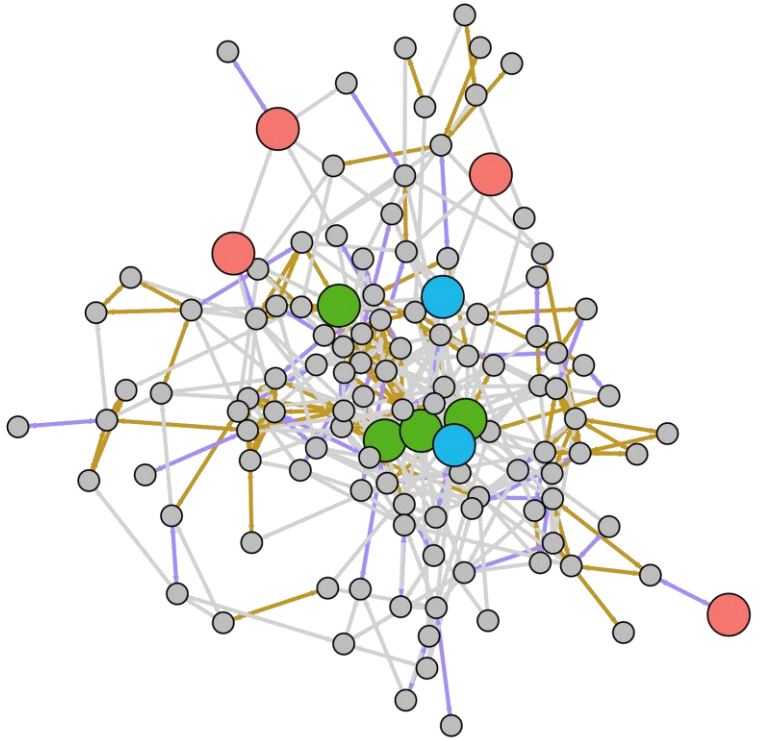
Networks of data



http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png

Summary: Case 1

- Reading vs. Mining
- Knowledge infrastructures
  - Access to data resources
  - Rights and responsibilities for using data
  - Tools and technologies
  - Legal and policy matters
    - University contracts
    - Copyright and ownership of data
    - Privacy and data protection



http://humannaturelab.net/wp-content/uploads/2015/01/Fig1-no-text-village-2-only-selection.png

Borgman, C. L. (2020). Whose text, whose mining, and to whose benefit? *Quantitative Science Studies*, *1*(3), 993–1000. https://doi.org/10.1162/qss_a_00053

Case 2: Digital Data Archives

Digital data archives as knowledge infrastructures

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888–904. https://doi.org/10.1002/asi.24172

DANS as a Digital Data Archive

- Netherlands government funding; about 50 staff
- Partners in international research infrastructure projects
- 50+ years of social science and humanities data
- Published datasets in EASY: 32,000+
- About 3.5 million files; total 5TB

Data Archiving and Networked Services

DANS

DANS is an institute of KNAW and NWO

KNAW

NWO

Driven by data

Note: figures as of 2018

Research Questions

1. What are the roles of digital data archives in knowledge infrastructures?
2. How do stakeholder roles in digital data archives vary?
   a. Contributors to a data archive?
   b. Consumers of a data archive?
   c. Archivists who manage the data archive?

Data Archiving and Networked Services

DANS

UCLA Center for Knowledge Infrastructures

Research Methods

- Ethnography
- Mining transaction logs
- Interviews
  - DANS data contributors
  - DANS data consumers
  - DANS archivists and staff

A. Scharnhorst, C. Borgman, DANS 2014

DANS Interviews 2015-2016

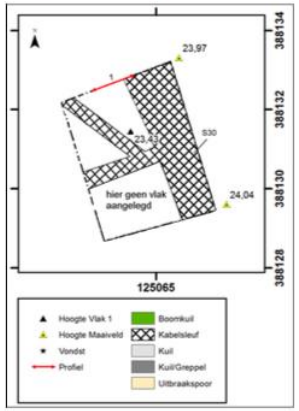
Stakeholders	Number of Interviews	Domain Expertise	Occupation
Data contributors	9	Archaeology, History, Paleogeography (3) Labor Economics (1) Linguistics (1) Oral Histories (1) Information Science (1) Theology (1) Plant Biology (1)	Academic staff (4) Cultural institution staff (3) Private company staff (1 interview with 2 staff) Unaffiliated (1)
Data consumers	8	Archaeology, History (6) Political Science, Sociology, Public Administration (3)	Academic staff (3) Cultural institution staff (2) Citizen scientists (1) Students (2)
DANS staff	10	Archaeology and humanities (6) IT Development (4)	Archivists Project Managers IT developers

Findings: DANS data

Afbeelding 8 (links). Proefsleuf nr. 1, Profiel nr. 2. Foto (Fotonummer 3) genomen naar het zuiden.

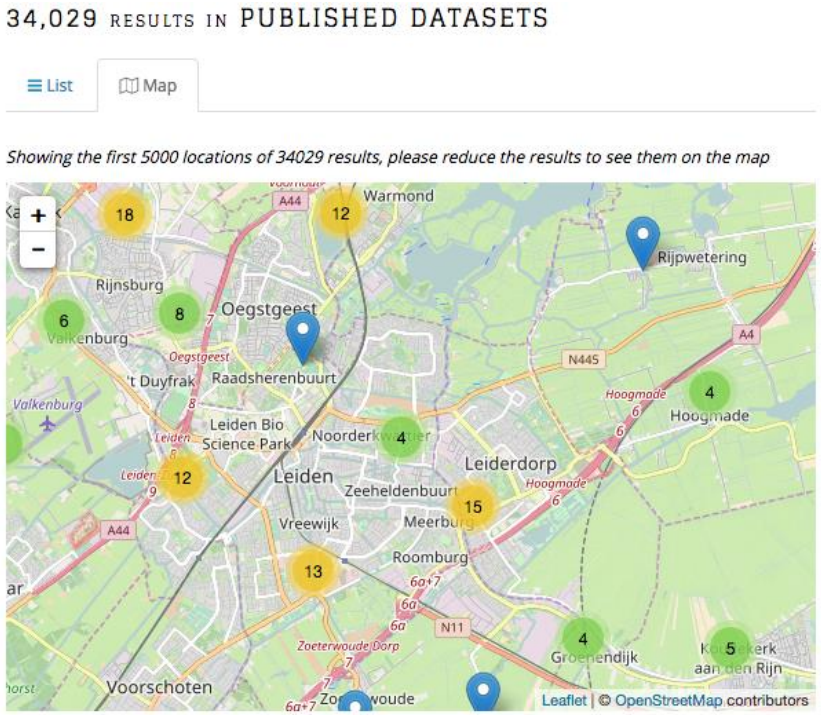
Wilgen, L. R. van (SOB Research) (2014).DANS. https://doi.org/10.17026/dans-z5y-tdb6

Afbeelding 46. Proefsleuf nr. 7, Vlak 1. Duidelijk zichtbaar is de verstoring door kabels en leidir 13) genomen naar het noordoosten.

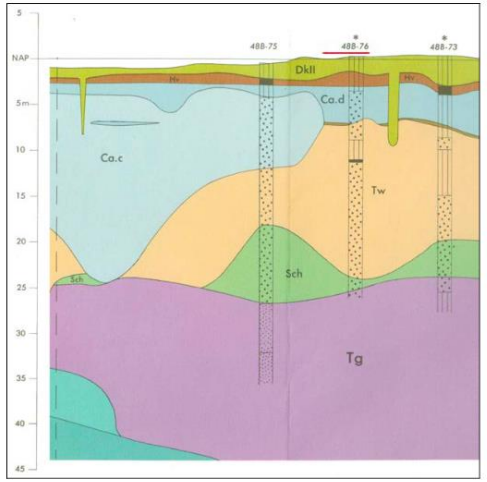


Afbeelding 47. Proefsleuf nr. 7, Vlak 1. Schaal 1: 100.

Bosch, J. E. van den (SOB Research) (2012). DANS. https://doi.org/10.17026/dans-2cg-tg88



https://easy.dans.knaw.nl/ui/?wicket:interface=:1:1:::



Afbeelding 6. De globale ligging van het plangebied (rood gemarkeerd), geprojecteerd (zie Afbeelding 5) van de Geologische Kaart van Nederland.

Structured and unstructured data

ARCHEOLOGISCH BUREAUONDERZOEK EN INVENTARISEREND VELDONDERZOEK D
GRONDBORINGEN BOUWLOCATIE WEELWEG 1, GAPINGE, GEMEENTE VEERE

1. Inleiding

1.1 Planontwikkeling

Het archeologisch onderzoek is uitgevoerd in het kader van de vergunningverlening voor de uitbreiding van een ligboxenstal aan de Weelweg 1 te Gapinge (Gemeente Veere). Het plangebied beslaat een oppervlakte van circa 1200 m . De belangrijkste te voorziene bodemverstoringen betreffen de aanleg van funderingen en daarnaast de aanleg van een mestkelder met een oppervlakte van circa 830 vierkante meter en een aanlegdiepte tot circa 2.3 beneden het maaiveld.

Afbeelding 1. Ligging van het plangebied (rode stip) in Nederland.

1.2 Archeologisch onderzoek

Op de Archeologische verwachtingsadvieskaart voor het grondgebied Walcheren (WAD, 2008) wordt ter plaatse van het plangebied een zone weergegeven met een middelhoge archeologische verwachting

«« Back to list

Overview | Description | Data files (1)

Download | View details

Dataset Contents / original

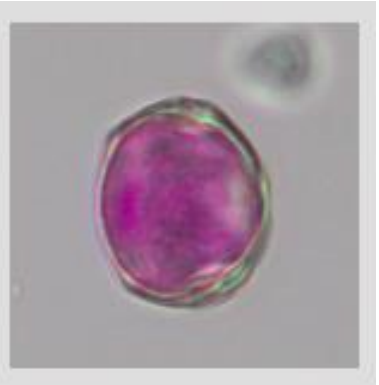
Dataset Contents
original

	Name	Size	Accessible
	Veere-Rapport Weelweg 1-120919.pdf	2605111	Yes

https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:66597/tab/2

Nested datasets with file access control

DEVELOPMENT OF AN AUTOMATIC POLLEN CLASSIFICATION SYSTEM



«« Back to list

Overview | Description | Data files (602)

Download | View details

Dataset Contents / original / Pollen600 / Alder

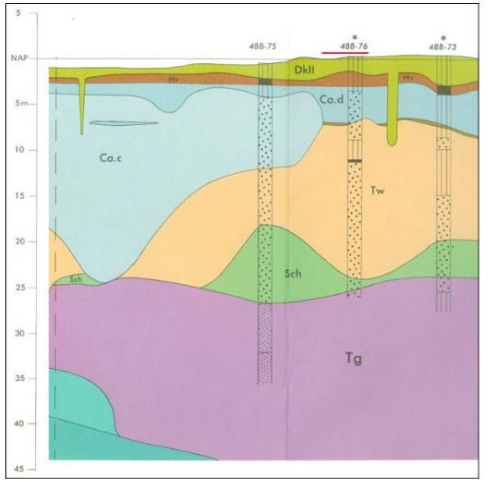
- Dataset Contents
  - original
    - Pollen600
      - Alder
      - Birch
      - Hazel
      - Mugwort
      - Sweet Grass

Name	Size	Accessible
100Alder.TIF_particle_.tif	41126	Yes
101Alder.TIF_particle_.tif	43186	Yes
102Alder.TIF_particle_.tif	44352	Yes
103Alder.TIF_particle_.tif	42874	Yes
104Alder.TIF_particle_.tif	42432	Yes
105Alder.TIF_particle_.tif	46342	Yes
106Alder.TIF_particle_.tif	40934	Yes
107Alder.TIF_particle_.tif	43394	Yes
108Alder.TIF_particle_.tif	35566	Yes

https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:62234/tab/2

Who contributes data to DANS?

- Archaeologists
  - Companies that conduct site surveys
  - University researchers
- Researchers
  - Public policy, economics, geography
  - Sociology, linguistics, oral history…
- Staff on behalf of data collectors
  - Institutional librarians
  - Department heads



Afbeelding 6. De globale ligging van het plangebied (rood gemarkeerd), geprojecteerd (zie Afbeelding 5) van de Geologische Kaart van Nederland.

Bosch, J. E. van den (SOB Research) (2012). DANS. https://doi.org/10.17026/dans-2cg-tg88

Why do they contribute data to DANS?

- Meet legal requirements
- Get credit for data
- Share data with others
- Control access to their data
- Preserve data for long term

Afbeelding 8 (links). Proefsleuf nr. 1, Profiel nr. 2. Foto (Fotonummer 3) genomen naar het zuiden.

Wilgen, L. R. van (SOB Research) (2014).DANS. https://doi.org/10.17026/dans-z5y-tdb6

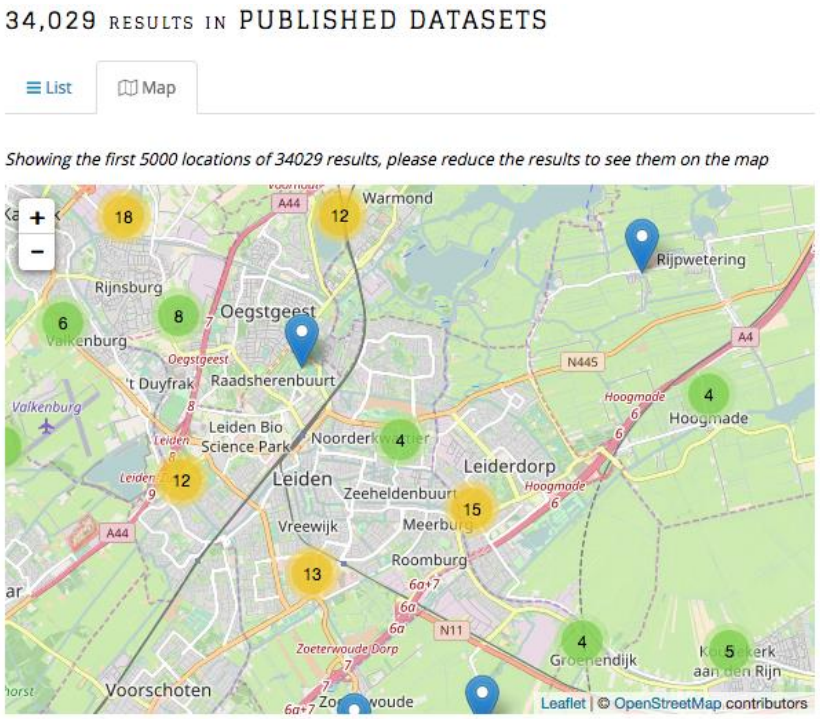
Who consumes data from DANS?

- Archaeologists
  - Companies that conduct site surveys
  - University researchers
- Researchers
  - Public policy, economics, geography
  - Sociology, linguistics, oral history…
- Students, teachers, visiting scholars
- Local history guides, genealogists
- Museum curators, amateur scientists...

http://www.little-dixie.lib.mo.us/?page_id=42

How do users search in DANS?

- Browse topics
- Browse spatial regions
- Search for place names
  - Archaeology sites
  - Building sites
- Few known item searches
- Some scripted searches
- Open access datasets are most used



https://easy.dans.knaw.nl/ui/?wicket:interface=:1:1:::

How are DANS data used?

- Downloaded for later use
- Reports are read
- Geographic regions assessed
- Comparisons to other data
- Create new products, e.g., local history guidebooks

http://www.intlconnect.illinois.edu/jobsearch

Visible roles of archivists

- Acquire data
  - Work with contributors
  - Seek useful data
- Disseminate data
  - Assist users in searching
  - Outreach to communities
- Staff help desks

Invisible roles of archivists

- Curate data
  - Ingest, clean, verify anonymity
  - Migrate/transfer to other formats
  - Describe, document, add metadata
  - Sustain access for 10 years
- Apply expertise
  - Subject domain experts
  - Metadata and cataloging experts
  - Design and software engineering
  - Statistics and data analysis

#addValue

Summary: Case 2

Data Archiving and Networked Services

DANS

- Archive content is community driven
- Archives are a source of trust in datasets
- Archive uses and users are diverse
- Access to data is mediated by infrastructure stakeholders
  - Archivists
  - Archive managers
  - Contributors who release or retain control over data
  - Institutions who promote or constrain access

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888–904. https://doi.org/10.1002/asi.24172

Conclusions

A call to action for the humanities (2009)

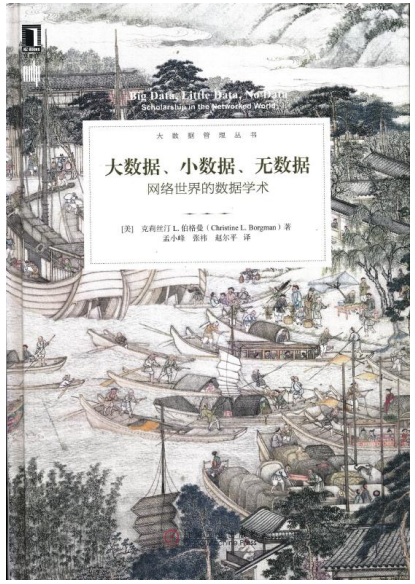
- What are data?
- What are the infrastructure requirements?
- Where are the social studies of digital humanities?
- What is the value proposition for digital humanities?

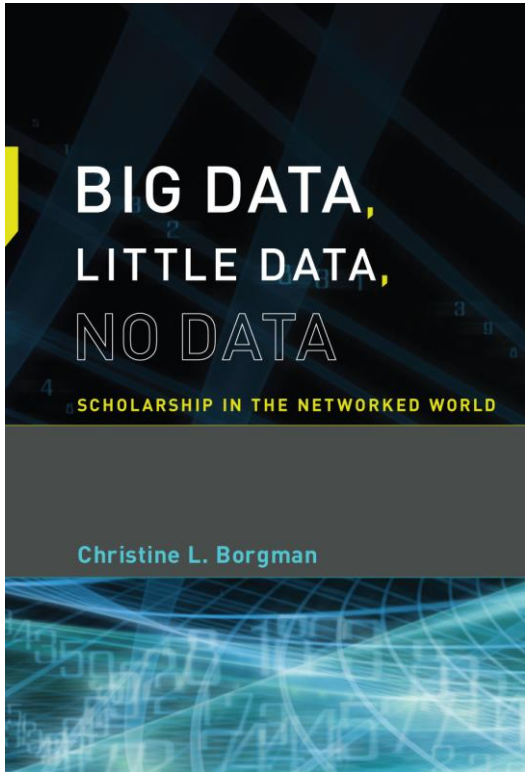
Royce Hall, UCLA

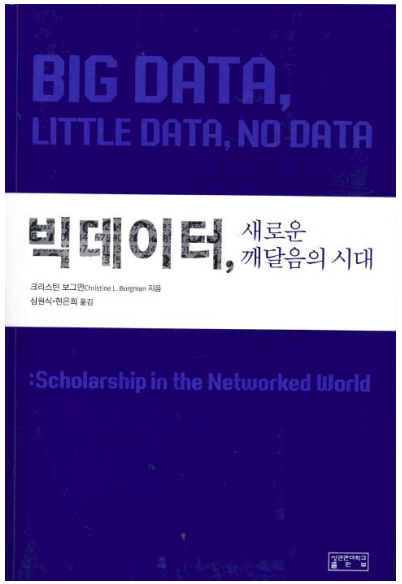
Borgman, C. L. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly*, *3*. http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html

What are data?

- Representations … used as evidence of phenomena…
- Text, images, sounds, symbols, maps, music…







What are the infrastructure requirements?

Knowledge infrastructures: "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds"

- Technical infrastructures
- Scholarly practices
- Governance models
- Policy frameworks

Royce Hall, UCLA

Where are the social studies of digital humanities?

- Social studies of science is a robust field
- DANS study is a rare case in humanities
- Research on scholarly practice
  - Identify innovative methods
  - Identify benefits and risks
  - Develop new methods and tools
  - Develop new policies and practices
  - Transfer insights between communities

Royce Hall, UCLA

What is the value proposition for digital humanities?

- Opportunities
  - Cultural data studies
  - Open data, open questions
  - Primary materials in digital form
  - Global questions, local computing
  - FAIR principles promote data reuse
- Challenges
  - Build robust knowledge infrastructures
  - Invest in digital data archives
  - Produce reusable data
  - Advance research policy

Royce Hall, UCLA

Acknowledgements

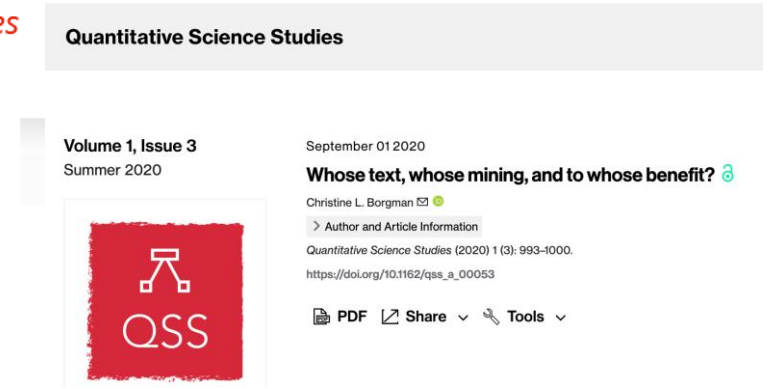
UCLA Center for Knowledge Infrastructures

Bernadette Boscoe
Peter Darch
Milena Golshan
Irene Pasquetto
Ashley Sands
Michael Scroggins
Morgan Wofford

Cotsen Institute for Archaeology
- Willeke Wendrich
- Deidre Whitmore

Data Archiving and Networked Services
DANS

Henk van den Berg
Ingrid Dillo
Peter Doorn
Andrea Scharnhorst
Herbert van de Sompel
Andrew Treloar
Sally Wyatt



Data Mining with Limited Access Text: National Forum (2018)

QSS special issue editors
- Loet Leydesdorff
- Stasa Milojevic
- Ismael Rafols

NSF

Alfred P. Sloan Foundation 1934

For a full list of CKI participants, collaborators, and publications, see https://knowledgeinfrastructures.gseis.ucla.edu/