

# Digital Source Criticism

How to deal with digital metasources  
in the age of big data?

Andreas Fickers / University of Luxembourg / C<sup>2</sup>DH

# The challenge

- To face the growing gap between the fast development of digital sources and tools for doing history and the rather slow appropriation and critical reflection of digital tools and techniques by the academic community
- We need an "update" of classical hermeneutics (Schleiermacher, Dilthey, Heidegger, Gadamer, Ricoeur, Habermas) to the digital age

# A matter of epistemological urgency!

“History as a field of enquiry is standing on the edge of a conceptual precipice. Historians need to be thinking about the radical impact of the digital turn in historiography and historical methodology in a critical and engaged manner.”

- Weller, Toni, ‘Introduction: History in the digital age’, in: Toni Weller (ed.), *History in the Digital Age* (London: Routledge, 2013).

# Digital Hermeneutics

- Hermeneutic reflection needs to include all phases of a research / learning process:

- Searching
- Documenting
- Analyzing
- Presenting

**Digital Source Criticism** is an essential part of digital hermeneutics

# Digital sources as new “epistemic thing”

- “The methodological consequences of using digitized and born-digital primary sources, particularly in comparison with traditional printed sources is, of course, one of the most sensitive epistemological issues for today’s historian.”

– Serge Noiret, “The digital historian’s craft and the role of the European History Primary Sources (EHPS) Portal”, in *Archivi & Computer. Automazione e Beni Culturali*, 19 (2009) 2/3, pp. 5-41.

# Digital sources as new “epistemic thing”

- Dealing with digital sources / information / data is at the very heart of the historical practice / profession
- It touches the disciplinary nature of history as “science” which is based on:
  - The scientific ideal of objectivity
  - The originality of documentary sources
  - The professional habitus of university trained historians
- It forces us to rethink the basic skills and literacy of our discipline
  - Information retrieval & management
  - Documentation and interpretation
  - Argumentation and storytelling

# **6 questions, which historians need to critically engage with:**

- How does digitization affect the concept and function of archive / archiving?
- What new heuristics of search are needed in the age of internet and “big data”
- How to develop a critical methodology of digital source critique?
- What new historical questions can new digital tools and techniques produce?
- What are the possibilities of digital storytelling in history? Who is our public?

# **6 questions, which historians need to critically engage with:**

- How does digitization affect the concept and function of archive / archiving?
- What new heuristics of search are needed in the age of internet and “big data”
- How to develop a critical methodology of digital source critique?
- What new historical questions can new digital tools and techniques produce?
- What are the possibilities of digital storytelling in history? Who is our public?



# **6 questions**, which historians need to critically engage with:

- How does digitization affect the concept and function of archive / archiving?
- What new heuristics of search are needed in the age of internet and “big data”
- How to develop a critical methodology of digital source critique?
- What new historical questions can new digital tools and techniques produce?
- What are the possibilities of digital storytelling in history? Who is our public?

# **6 questions**, which historians need to critically engage with:

- How does digitization affect the concept and function of archive / archiving?
- What new heuristics of search are needed in the age of internet and “big data”
- How to develop a critical methodology of digital source critique?
- What new historical questions can new digital tools and techniques produce?
- What are the possibilities of digital storytelling in history? Who is our public?

# **6 questions**, which historians need to critically engage with:

- How does digitization affect the concept and function of archive / archiving?
- What new heuristics of search are needed in the age of internet and “big data”
- How to develop a critical methodology of digital source critique?
- What new historical questions can new digital tools and techniques produce?
- What are the new possibilities of digital storytelling?

# **6 questions**, which historians need to critically engage with:

- How does digitization affect the concept and function of archive / archiving?
- What new heuristics of search are needed in the age of internet and “big data”
- How to develop a critical methodology of digital source critique?
- What new historical questions can new digital tools and techniques produce?
- What are the new possibilities of digital storytelling?
- Does digital history enable a new public engagement with history?

# New Digital Hermeneutics

- We need a number of skills / multi-modal literacy combining:
  - Searching: **algorithmic criticism**
  - Documenting: **digital source criticism**
  - Analysis: **tool criticism**
  - Presentation: **interface criticism**

# 1) Digitization and archiving

- How does digitality / digitization affect the concept of archive?
- How do we deal with the change from the “age of scarcity” to the “age of abundance” (Roy Rosenzweig)
- Epistemological implication: how does digitization (retro-digitization as well as digital born) affect the ontological status of “sources”?

# 1) Digitization and archiving

- Since 19<sup>th</sup> century: archival science has developed best practices for selection, inventarization, conservation, retrieval of archived sources
- Promoted the concept of “record” as standardized description model for a variety of “documents”
  - Mainly administrative documents
- Professional (historical) archives emerge in parallel with the professionalization of history as a scientific discipline
  - Both archives and history as a discipline play a crucial role in the “invention of tradition” : creation of historical master narratives of “nations”
  - “Critical” source editions: tools for training of historians

# 1) Digitization and archiving

- Archives developed into institutions with high symbolic (“treasures”) and social (“gatekeepers”) capital;
- Incorporate / embody professional expertise and “trust”: archival “records” have been “proofed”
- Appraisal of documents as “sources” is core expertise of archivists: appraisal based on the concept of “provenance”
  - Records originating from the same “source” should be kept together to form one “corpus” (“respect de l’ordre”)
  - They should not be merged / interfiled with other records from other sources (“respect de fonds”)



# 1) Digitization and archiving

„Documents as evidence are ontological entities whose evidentiary origins lie in their belonging to taxonomic or indexical regimes or to looser discursive or conversational regimes (...) ‚Facts‘ occur through the infrastructuralization of documentary techniques and technologies.“

Ronald E. Day, *Indexing it all. The subject in the age of documentation, information, and data*. Cambridge, Mass.: MIT 2014, p. 4f.

# 1) Digitization and archiving

Despite heavy “interference” of archivist / archive in the creation of records / files:

- myth of the archive as “neutral”,
- records as “innocent” and
- archivists as “objective”

# 1) Digitization and archiving

- Archives as institutions wield power
  - Over administrative, legal, fiscal accountability of governments, corporations, and individuals
- Archives as records wield power
  - Over the shape and direction of historical scholarship, collective memory and national identity
- Archivists as keepers of archives wield power
  - Over through active management of records, their appraisal, selection, description, preservation and use.

# 1) Digitization and archiving

- „Archives have always been about power, whether it is the power of the state, the church, the corporation, the family, the public, or the individual. Archives have the power to privilege and to marginalize. They can be a tool of hegemony; they can be a tool of resistance. They both reflect and constitute power relations [...] They are the basis for and validation of the stories we tell ourselves, the story-telling narratives that give cohesion and meaning to individuals, groups, and societies.“
  - Joan M. Schwartz / Terry Cook: ‘Archives, Records, and Power: The Making of Modern Memory’, in: *Archival Science* 2 (2002), p. 13.

# 1) Digitization and archiving

- Epistemological consequences of ontological change from “sources” to “document” to “data”:
  - Digitization fractures the “control zone” of classical archives / libraries: we need to rethink power relations between users and owners of databases
  - Changing nature of “archives” asks for new search strategies and techniques (“heuristics of search”) and a critical reflection on the availability / quality of “metadata” (contextual information)

# 1) Digitization and archiving

New questions / challenges:

- What to do with traditional concepts of source critique / archival logic such as “originality” or “authenticity”?
- How to test / qualify the “integrity” of digital sources? (problem of “authenticity”)
- How to deal with the logics / practices of archiving (“respect des fonds” / “respect de l’ordre”) in full text searchable databases?

## 2) Heuristics of search

- How to find relevant information in the age of abundance?
- Joshua Sternfeld: „Historical Understanding in the Quantum Age“
  - 2010: Library of Congress signed agreement with Twitter to preserve all 170 billion of the companys tweets created between 2006 and 2010
  - To continue to do so on a daily basis: roughly half a billion of tweets a day!
  - Not to count retweets & „favourited“ tweets...

## 2) Heuristics of search

- Joshua Sternfeld: „Historical Understanding in the Quantum Age“
  - Doing history in the „quantum framework“ needs a basic reflection on the concept of „scale“ and „appraisal“
  - Tension between „Newtonian physics“ and „Quantum physics“ = tension between „classical history“ and „digital historiography“
  - „When you move into the realm of the vast, traditional laws governing historical understanding begin to breakdown“



## 2) Heuristics of search

- “Google-Syndrom” (Peter Haber): has turned the classical, deductive mode of information retrieval in history upside down:
  - Instead of going from the general to the specific (handbooks, monographs, articles, sources), we go from the specific to the “similar”
  - From specific search (opacs, Bool operators) to *browsing*
  - Statistical evidence (based on algorithms and semantic genealogy) instead of thematic / problem based relevance

## 2) Heuristics of search

- Google has invented a new business model based on the logic of “linguistic capital”
  - Google algorithms are turning “words” (search terms) into economically exploitable linguistic subsets
  - Commodification of language is producing a new language (creole): “new algorithmic dialects”
    - Frederic Kaplan: “Linguistic Capitalism and Algorithmic Mediation”, in: Representations 127 (summer 2014), pp. 57-63.
    - Max Kemman et al.: “Just google it”!

## 2) Heuristics of search

- Search engines as “black boxes”?
  - Searching for the “normal”?
  - Looking for deviance?
  - Discovering “patterns”
  - Establishing hierarchies or relational structures?
    - David Gugerli, *Suchmaschinen. Die Welt als Datenbank*. Frankfurt a.M.: Suhrkamp 2009.
- Search engines are not neutral! They don’t retrieve information, but co-produce information by ranking and indicating the “importance” of information!
  - Need “algorithmic criticism” (Stephen Ramsey /*Reading Machines* 2011)

## 2) Heuristics of search

- Challenges / problems:
  - you only search for / find what is made findable / visible!
    - Only 2 % of cultural heritage is digitized!
    - Large parts of the web are “invisible” (non indexed or hidden sites); key issue of referencing / tagging in semantic web; total dependency on (hidden) logic of algorithms
    - High failure rate of automatic search software (OCR / especially with retro-digitized sources)
  - Suggested “objectivity” or “neutrality” of visualizations of search results (hierarchies, rankings,...) is biased (visual evidence!)

windenergie

### Result filters

Clear selection

### Term statistics chart

Clear selection

**MILIEUBERICHT** 20-2-1990  
Wekelijks programma met nieuws over het milieu. Deze aflevering aandacht voor windenergie. In Nederland bestaat een aantal windmolenverenigingen, die mbv windenergie elektriciteit opwekken voor eigen gebruik. De leden verdienen zelfs geld door...

**Schone Inlas** 20-12-2010  
Het is schoon, milieuvriendelijk en duurzaam en het ziet er ook nog eens prachtig uit: windmolens. De Nederlandse regering is er zo enthousiast over dat er de komende jaren flink in wordt geïnvesteerd. Maar is windenergie wel zo milieuvriendelijk al...

**ENERGY SURVIVAL** 11-12-2006

## 2) Heuristics of search

- Points of reflection:
  - Temptation of “quick and easy”: democratisation or de-professionalization?
  - “Algorithmic turn”: need for basic statistical / mathematical literacy in humanities; search algorithms are not neutral! “Algorithmic criticism” should be an essential part of digital literacy!
  - Often uncritical use of statistical mechanisms for quantitative analyses (“big data” = “trust in numbers” ? (Ted Turner)
  - “statistical correlation” or “visual evidence” (data visualization) is not “scientific” / “historical” relevance

### 3) Digital Source Critique

- Shift from “source” to “document” to “data” changes the ontological status of historical sources
- Digitization as process of coding and re-coding impacts on indexical relationship between source and historical reality
- Instead of “sources” we should speak of meta-sources:
  - “a set of structured information, modelled, passed on to the computer and processed by it”

J.-Ph. Genet, « Source, métasource, texte, histoire », in: F. Bocchi et P. Denley (éd.), *Storia & Multimedia*, Bologna (1994), pp. 3-17.

# Digital sources as “matasources”

- We therefore need to critically reflect on the whole “life cycle” of metasources
  - “Il s’agit de tenir compte des ‘entours’ du document, des éléments paratextuels (présent sur le site web), situationnels (concernant la mutation du dispositif socio-technique) ou historiques”
    - Matteo Treleani, Mémoires audiovisuelles. Les archives en ligne ont-elles un sens? Montréal: Presses de l’Université de Montréal 2014, p. 30.



# Life cycle of digital sources



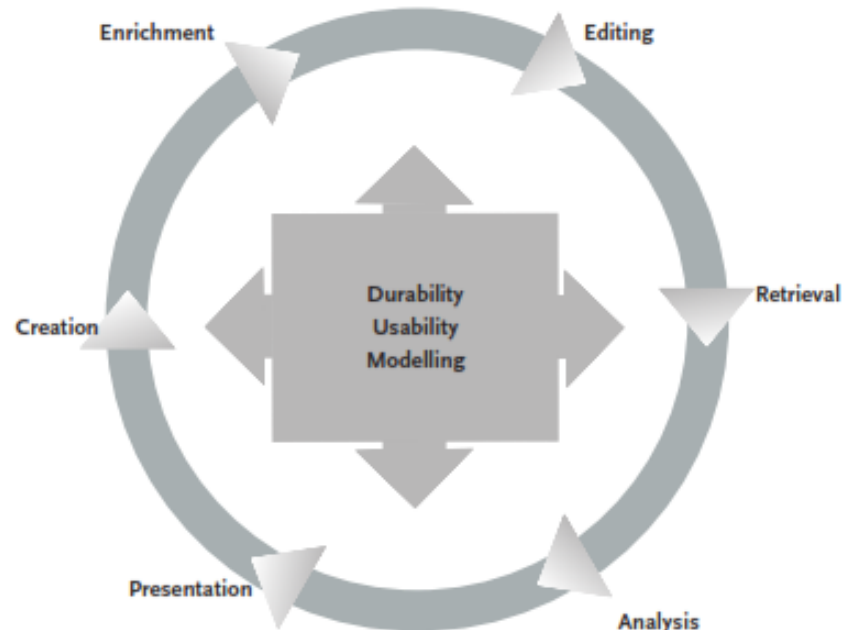
Onno Boonstra, Leen Breure, Peter Doorn: *Past, Present and Future of Historical Information Science* (Amsterdam: NIWI-KNAW, 2004).

# 3) Digital Source Critique

- Concepts of “original” and “authenticity” are obsolete
- We should concentrate on the relativistic nature and multidimensional definition of “data” and critically assess the integrity of data (6 V’s)
  - Volume? (lots of data or big data)
  - Velocity? (speed of accumulation and dynamic nature)
  - Variety? (mashing-up of heterogeneous data-types)
  - Validity? (amount of bias / noise)
  - Veracity? (correctness and accuracy)
  - Volatility? (persistence and longevity)
- Carl Lagoze, “Big Data, data integrity, and the fracturing of the control zone”, in: *Big Data & Society* 1 (2014), pp. 1-11.

# Dream: meta-data on data integrity

- Volume? (lots of data or big data)
- Velocity? (speed of accumulation and dynamic nature)
- Variety? (mashing-up of heterogeneous data-types)
- Validity? (amount of bias / noise)
- Veracity? (correctness and accuracy)
- Volatility? (persistence and longevity)



Protocol: automatic “tracking” of re-codings, attached to the digital “source”

# 3) Digital Source Critique

- Big problem (especially with “digital born” sources): missing metadata
  - Basic information needed for historical contextualisation (who, when, where) is often missing (audio-visual sources: YouTube)
  - Metadata are indispensable for setting up larger databases / linking of catalogues (“semantic interoperability”); standards are needed (for example “Dublin core”)
  - Pelle Snickars: “If content is king, context is its crown”  
<http://journal.euscreen.eu/index.php/view/article/view/jethc006/6>)

## 4) Tool criticism

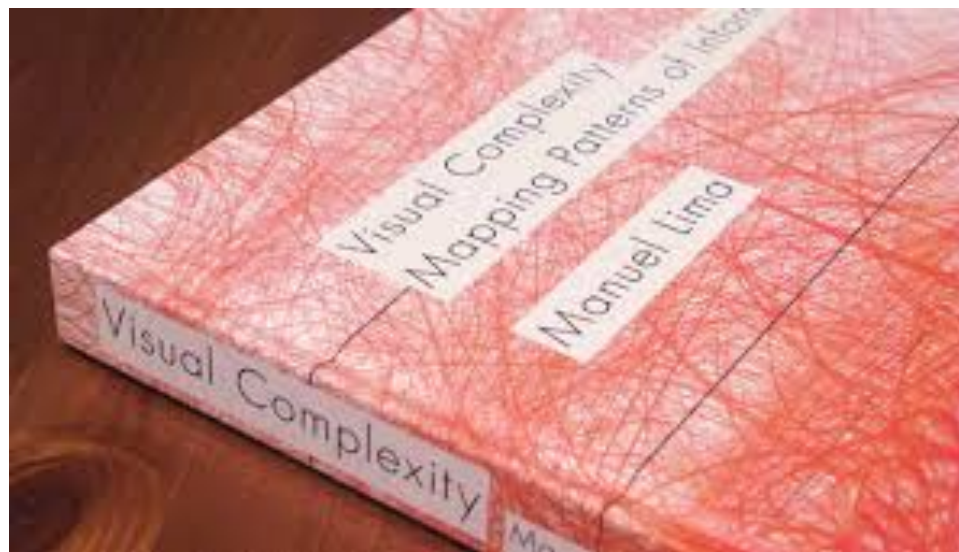
- Does the future of history lie in the computer-based analysis of “big data”?
- Great hopes:
  - “text mining” & visualisation of semantic relations
  - Software for optical recognition & sound analysis
  - Simulation of actor networks, structures, ...
- From “analysing data” to “doing things with data”
  - Jim Mussel: Move from digital history 1.0 to 2.0

## 4) Tool criticism

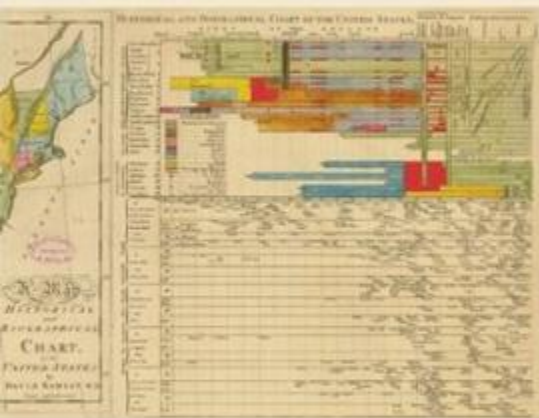
But: visual evidence is not historical relevance!

“If displays of data are to be truthful and revealing, then the design logic of the display must reflect the intellectual logic of the analysis (...) Clear and precise seeing becomes as one with clear and precise thinking”

Edward Tufte: *Visual Explanations*, p. 53.



## *Cartographies of Time*



## *A History of the Timeline*

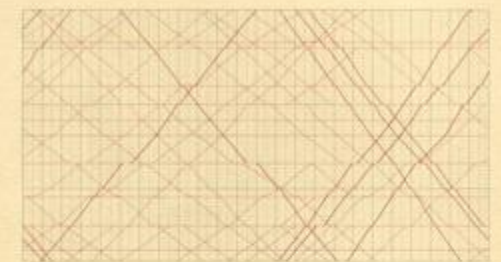
Daniel Rosenberg and Anthony Grafton

EDWARD R. TUFTÉ

## VISUAL EXPLANATIONS



IMAGES AND QUANTITIES. EVIDENCE AND NARRATIVE



## The Visual Display of Quantitative Information

EDWARD R. TUFTÉ

# Challenge: “Doing things with data”

- *Digital historian 1.0*
  - Must be able to understand why data performs as it does (digital source criticism; problematizing authenticity)
- *Digital historian 2.0*
  - Requires a more advanced understanding of the affordances of the digital (manipulating data from multiple resources, modelling their relationships)



# “Doing things with data”

- “In manipulating data from multiple resources, modelling their relationships and so exposing facets hitherto unrealized, the historian moves from simulation to simulacra, to validating representations against reified originals to producing analyses of phenomena, objects and relationships that belong to the past.”

– Jim Mussel: ‘Doing and Making. History as Digital Practice’, in: Toni Weller (ed.) *History in the Digital Age* (London: Routledge, 2013), p. 91.

- Doing things with data forces us to recognize / problematize the constructed nature of evidence (statistical, visual, semantic)!

## 4) Tool criticism

### – Problem of sustainability

- Short life time of projects; focus on development of new tools, but not on their use (didactical / pedagogical training)
- Dominance of “technological solutionism” (Evgeny Morozov)

### – Important:

- Joint ventures between developers (technician) and users (historian): developing “inter-language” in trading zone of DH
- Promote sustainability (embedding into existing infrastructures like DARIAH) and avoid reinvention of the wheel

## 4) Tool criticism

- Manfred Thaller: only use tools which you understand!
- Andreas Fickers: don't be afraid of playing around with new tools (experimental mind-set / “thinkering”); but: always in a reflexive mode: we need „tool criticism“ as yet another indispensable skill of digital literacy
- We should learn to combine the quantitative methods of „distant reading“ with the heuristic tradition of „close reading“
- We need to make critical assessment and discussion of possibilities & limitations of tools explicit when publishing our research results (// oral history)

## 4) Tool criticism

- Copy & paste culture / download-culture:
  - Produce individual research archives (thousands of digital photographs...)
  - How to develop a strategy for inventarization, categorization, description, annotation of “meta-sources” in our own databases?
  - Challenge of information management (dealing with abundance) is reproduced on an individual level
- History / historiography as “metadating meta-dates”
  - Creating own databases allows flexible categorization and “interpretative flexibility” in production of historical causalities

# 5) Interface criticism

- More and more, “data” of humanities research is produced and represented in digital forms and environments (Internet / open access policy)
- Human-computer interaction (both on soft- and hardware level) is central element in knowledge production
- But: we rarely reflect on the importance of interfaces in the man-machine communication / interaction
  - Desktop philosophy and intuitive user designs create “user illusion”
  - Interface culture is based on concealment of technology (black-boxing)

# 5) Interface criticism

- Increasingly complex visualization of our data / research results asks for a critical reflection of the “visual evidence” of the “appresentation” (Karin Knorr-Cetina) of our knowledge
- We need a better understanding of the “commodity layer” and the “mechanism layer” (David Berry) of continuous interfaces
- Graphs and charts reify statistical information and produce a “look of certainty” (Johanna Drucker)
  - “If displays of data are to be truthful and revealing, then the design logic of the display must reflect the intellectual logic of the analysis (...) Clear and precise seeing becomes as one with clear and precise thinking.”
    - Edward Tufte: *Visual Explanations*, p. 53.

# 5) Digital storytelling

- Big questions:
  - What to do with mass of digitized sources?
  - What new possibilities of historical storytelling do they offer / demand?
- New possibilities:
  - “enhanced publications” / digital editions / e-books
  - Online journals / blogs / twitter
  - Contemporary history: videoessays / podcasts / virtual exhibitions

## 6) Digital Storytelling

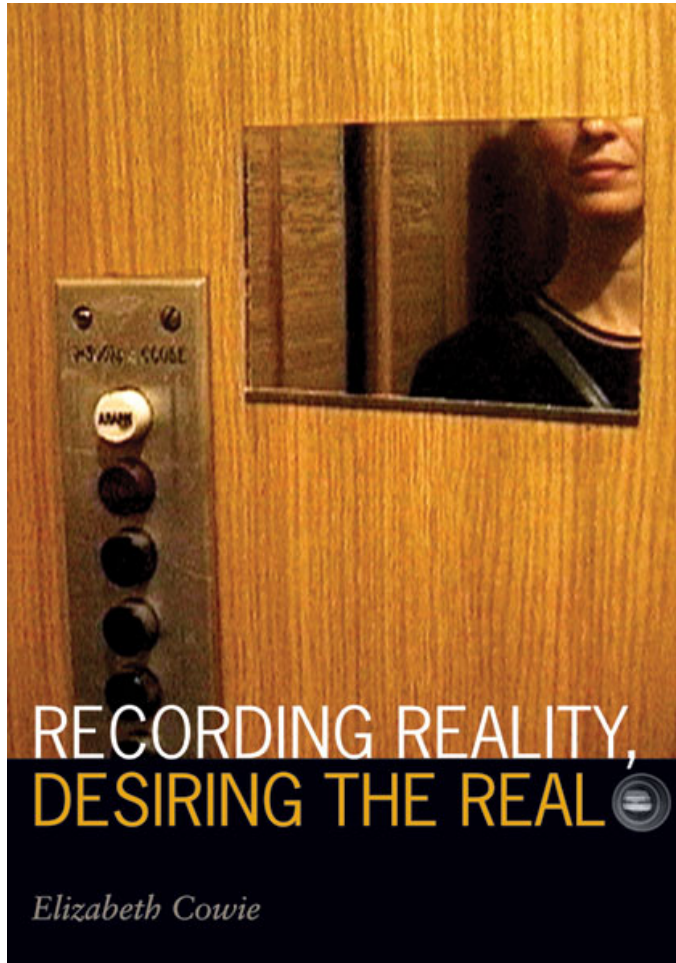
- So far:
  - Text (book, article) = remains standard medium for historical narration (with footnote as symbol of scientific authority)
  - But: making annotations in audio / audiovisual sources should become as natural as adding footnotes to a scholarly text
  - Challenge: go beyond “illustrative” use of audio/visual sources; they should become part of your argumentation



## 6) Digital Storytelling

- Historians of the digital age need to be able to understand and interpret both the codes *and* conventions of mediated representations of the past
- They need to *combine media and digital literacy* and incorporate specific skills trainings in their curricula

## 6) Digital Storytelling

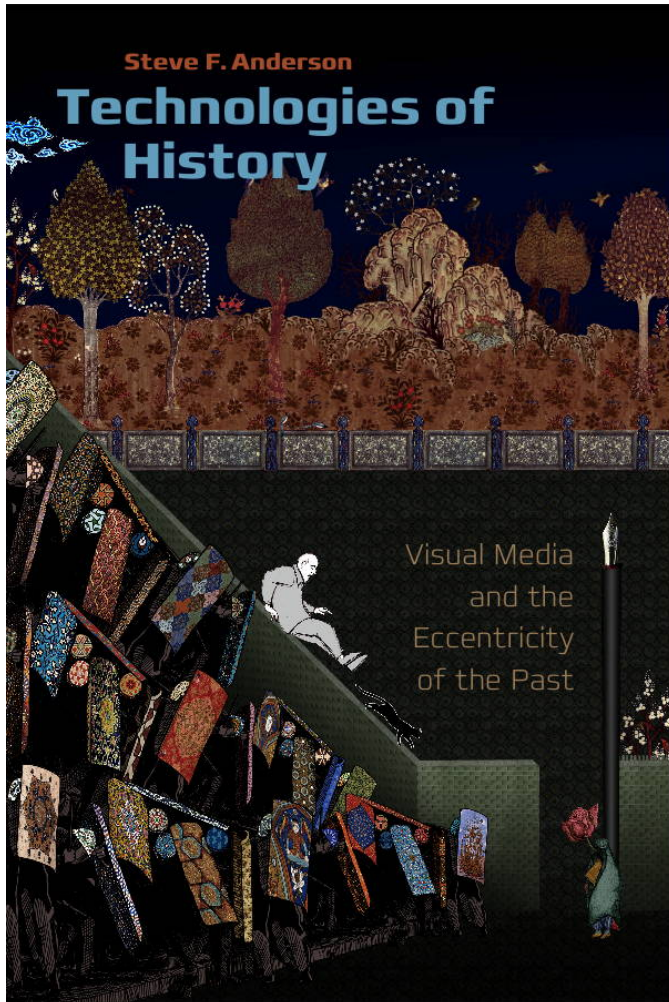


- Media literacy: understanding the language of images, sounds, complex multi-media representations
- “Photography and cinematography address two distinct and apparently contradictory desires. There is the desire for reality held and reviewable for analysis [...] and there is a desire for the real not as knowledge but as image – as spectacle” (p. 2)

## 6) Digital Storytelling

- Historians should engage in new forms / strategies of non-linear and transmedia storytelling
- Historians should explore the democratic potential of “data-base histories” and non-linear narratives

# Steve Anderson: database histories



- “histories comprised of not narratives that describe an experience of the past but rather collections of infinitely retrievable fragments, situated within categories and organized according to predetermined associations.” (p. 122)

# Henry Jenkins: transmedia storytelling

- “A transmedia story unfolds across multiple media platforms with each new text making a distinctive and valuable contribution to the whole. In the ideal form of transmedia storytelling, each medium does what it does best—so that a story might be introduced in a film, expanded through television, novels, and comics; its world might be explored through game play or experienced as an amusement park attraction”.
- Henry Jenkins, *Convergence Culture: Where Old and New Media Collide*. New York Univ. Press (2006), pp. 95–96.

# 7) Shared & public history

- Historians should engage in new ways of sharing resources, collaborative writing and storytelling to enable new forms of digital scholarship and dissemination
  - Example: Writing History in the Digital Age
  - “Born digital, open peer review, open access”
  - <http://dx.doi.org/10.3998/dh.12230987.0001.001>



## 7) Shared & public history

- The Internet / Web has affected the power relation between „professional“ and „amateur“ historians
  - Greater „visibility“ of local, community, bottom-up history archives and activities
  - „Grey literature“, sources, and objects of private collectors, enthusiasts, „nerds“ become interesting „data“
  - Possibility of public engagement / crowdsourcing = democratization of history!

## 7) Shared & public history

- Tension between „memory“ and „history“:
  - Eye/Ear-witness as enemy of historian?
  - Just as „oral history“ movement tried to give a voice to underrepresented actors in history (workers, women, minorities, migrants, etc.), public history projects might help to broaden the scope of traditional historical scholarship and to promote a dialogical approach to collective memory



# Conclusion: digital history as „trading zone“?

- We live in a phase of “creative uncertainty”:
  - Problem of communication: “in-betweenness” of vocabularies (jargon; creole; inter-language)
  - Problem of methodologies: tension between heuristics of “close reading” and “distant reading”
  - Problem of epistemology: co-construction of new epistemic objects & new strategies of constructing scientific evidence (from causality to correlation)

# Conclusion: digital history as „trading zone“?

- Digital hermeneutics as heuristic of “in-betweenness”:
  - „trying to locate a hermeneutics at the boundary between mechanism and theory (...) Algorithmic criticism proposes that we channel the heightened objectivity made possible by the machine into the cultivation of those heightened subjectivities necessary for critical work“.
  - Stephen Ramsey: *Reading Machines. Toward an Algorithmic Criticism*. University of Illinois Press 2011, p. x.

# Conclusion: digital history as „trading zone“?

Requires skills trainings and multi-modal literacy:

“We need database literacies, algorithmic literacies, computational literacies, interface literacies. We need new hybrid practitioners: artist-theorists, programming humanists, activist-scholars; theoretical archivists, critical race coders. We need new forms of graduate and undergraduate education that hone both critical and digital literacies. We have to shake ourselves out of our small, field-based boxes so that we might take seriously the possibility that our own knowledge practices are normalized, modular, and black boxed in much the same way as the code we study in our work.”

Tara McPherson, ‘U.S. Operating Systes at Mid-Century: The Interwinning of Race and UNIX’, in L. Nakamura and P. Chow-White (eds), *Race after the Internet* (New York: Routledge, 2012), p. 35.

# Conclusion: digital history as „trading zone“?

“The current challenge facing the discipline of history is not in creating ever bigger sets of data and developing new tools, important as these are. The real challenge is to be consciously hybrid and to integrate ‘traditional’ approaches in a new practice of doing history”.

- Gerben Zaagsma: ‘On digital history’, *BMGN / Low Countries Historical Review* 128 (2013) 4, p. 17.

**If “hybridity is the new normal”, we should be prepared for both – the “old” and the “new”!**