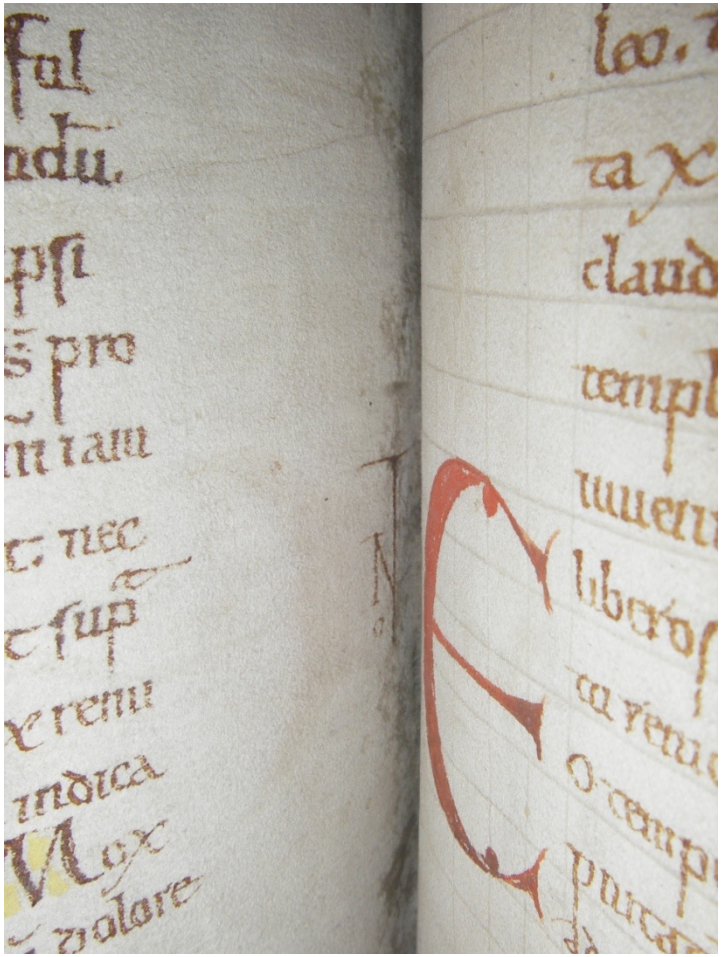
Cultural heritage digitisation

concepts
strategies
levels



Mats Dahlström

January 20, 2022

Concepts

- turning something from analog to digital
- digitalisation – digitisation

- digitalisering – digitisering

- digitalisat

WHY DIGITIZE?

Motives for digitizing cultural heritage material

- Preservation – reducing the wear and tear on the originals
- Make the material more accessible
- Make the material usable
- Make the material re-usable
- Develop and share technology and competence
- Enable new research
- Present the material in new ways and in new contexts
- Administrative, economical or political reasons
- Prestige and propaganda

The long chain of digitization

- Planning and budget
- Selecting, collecting, managing
- Document analys and preparation
- Strategies
- Tools (selecting, acquiring, developing)
- Text and image capture (scanning, digital camera)
- OCR and other technologies
- Text encoding and other optimization, e.g. indexing
- Testing, validating
- Metadata
- Selecting delivery media and distribution forms/channels
- Publication and distribution
- Marketing
- Evaluation, documentation and reports
- Archiving, migrating
- Maintenance

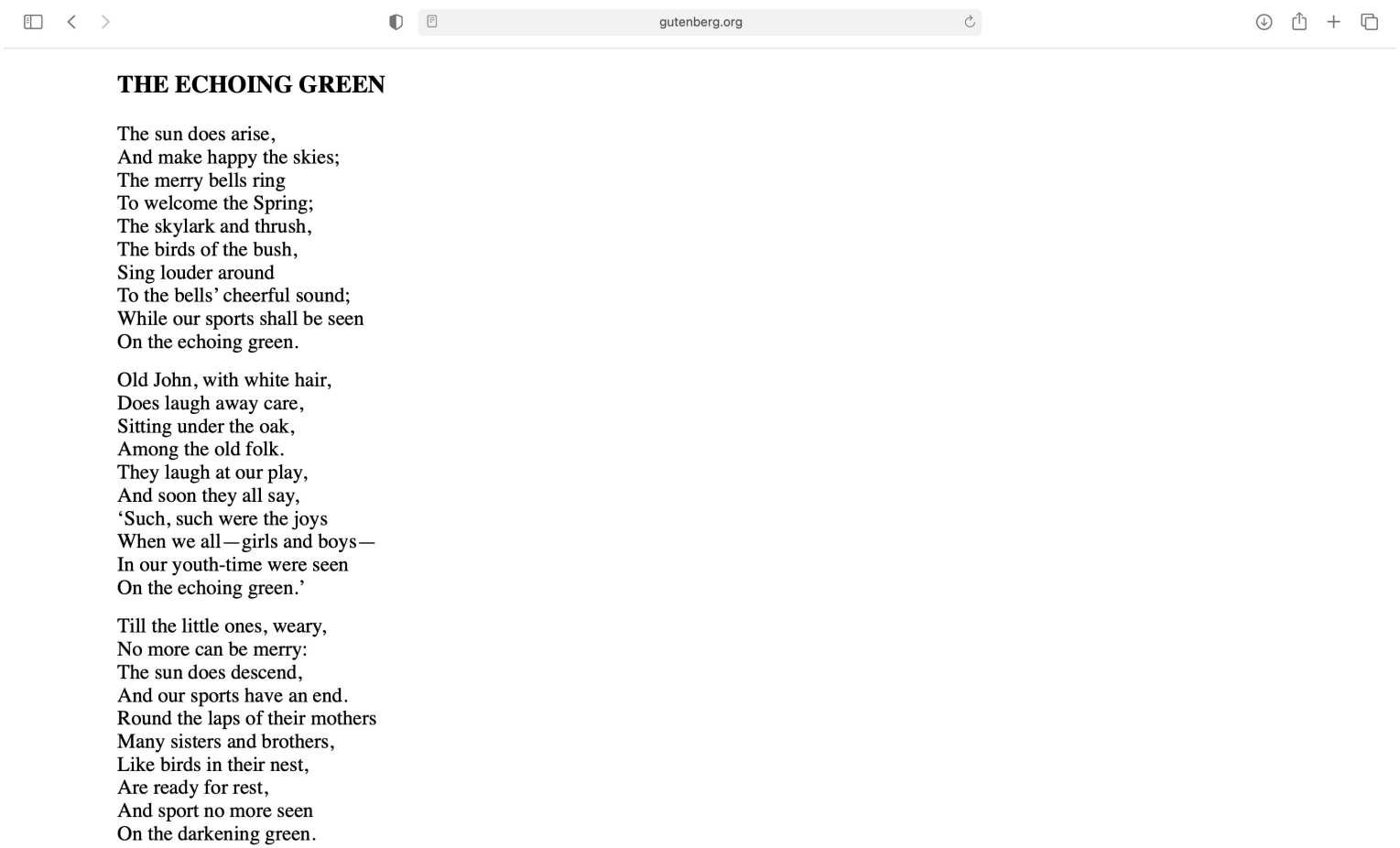
Examples of overall strategies

- source media / new media “fidelity”
- long term / short term digitization
- first hand / second hand
- proactive / reactive digitization
- conservation / restoration digitization
- text digitization / image digitization

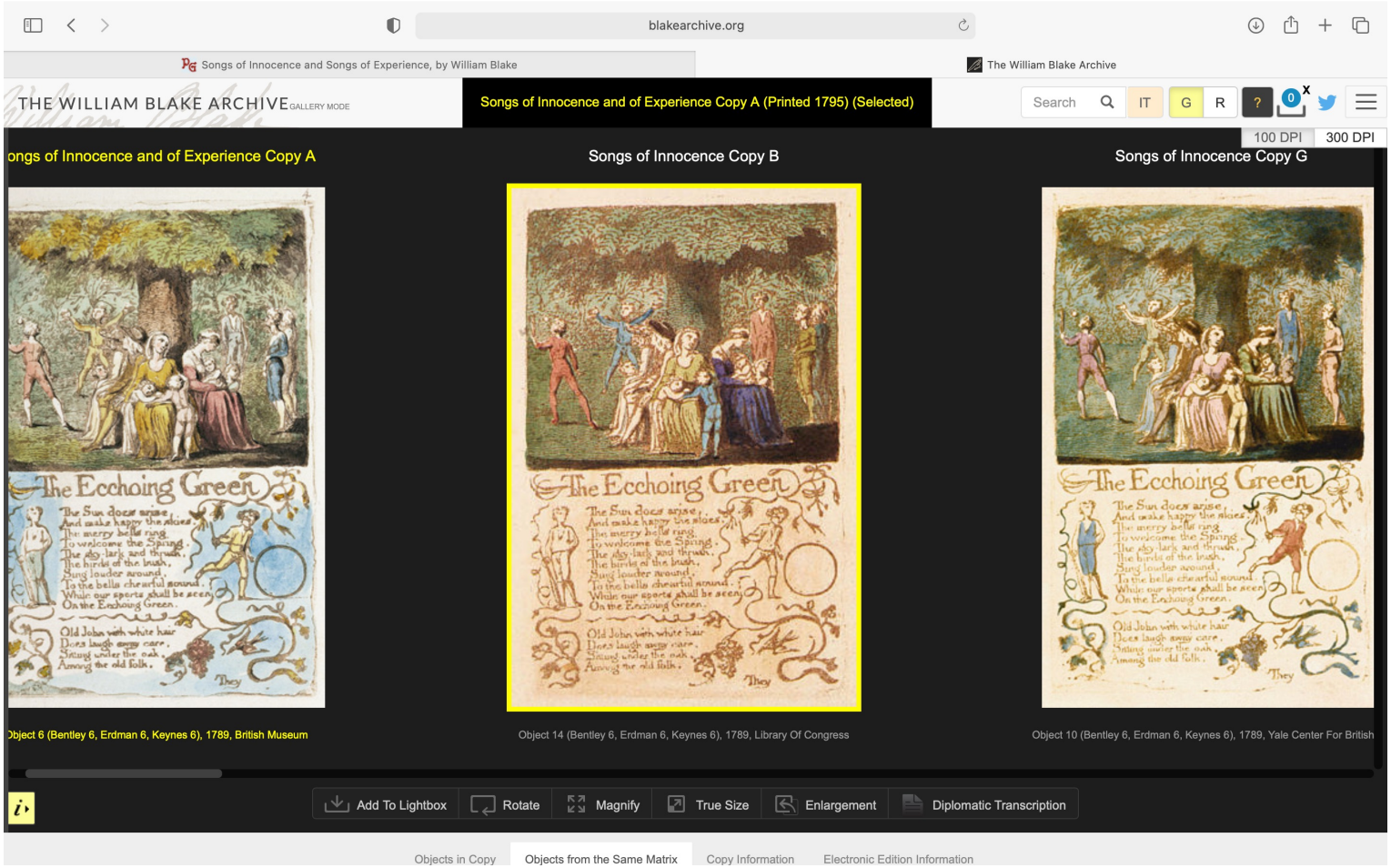
Text digitization versus image digitization

Linguistic text	Graphical and material artefacts (document)
Machine readable transcriptions	Digital facsimiles
Text capture (OCR)	Scanning, photography
Text encoding	Image editing
Text files	Image files
Might have a small ”facsimile” dimension	Often concentrated to the document parts which carry text

Faces of William Blake –
Project Gutenberg (PG)

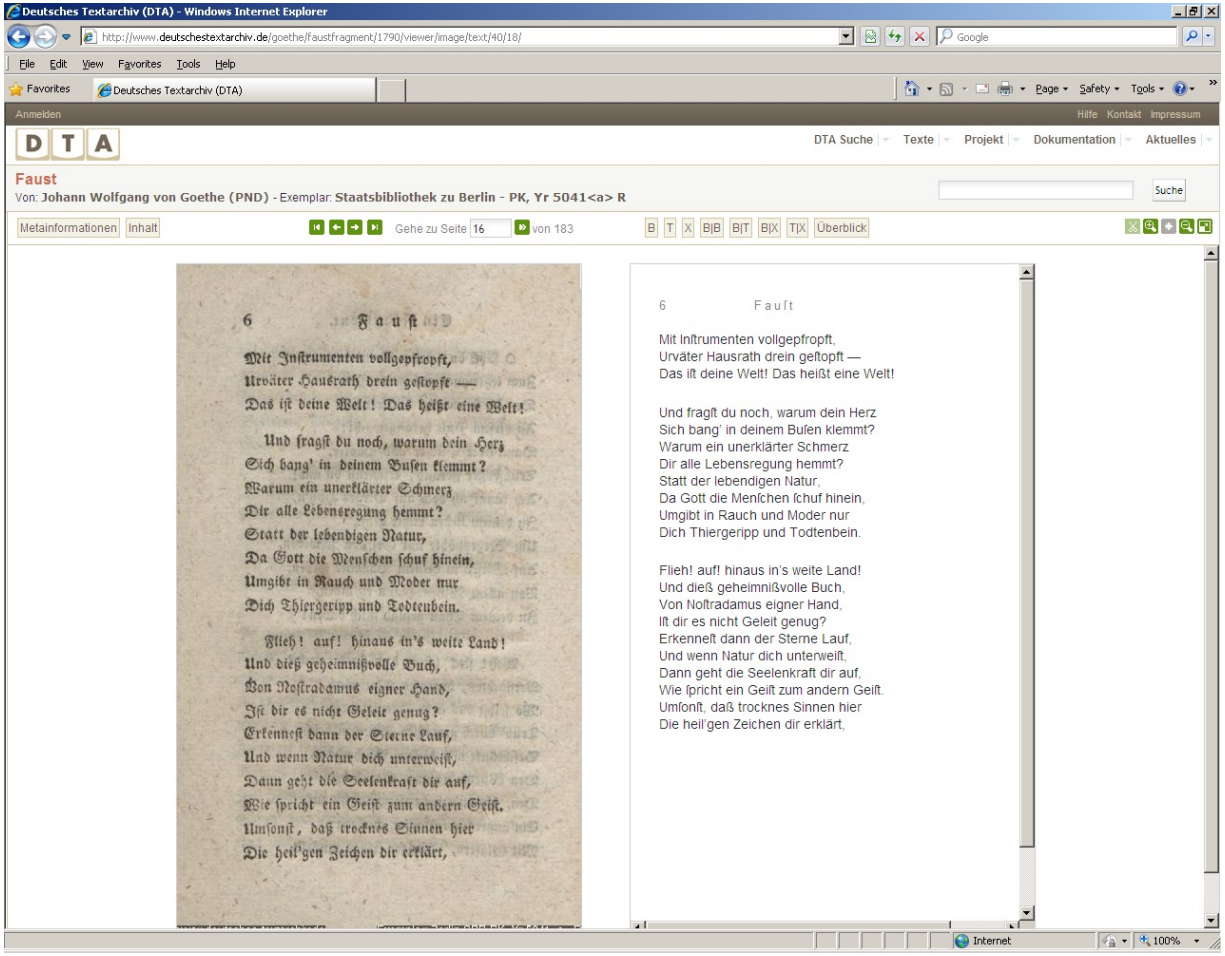


Faces of William Blake –
The William Blake Archive (BA)

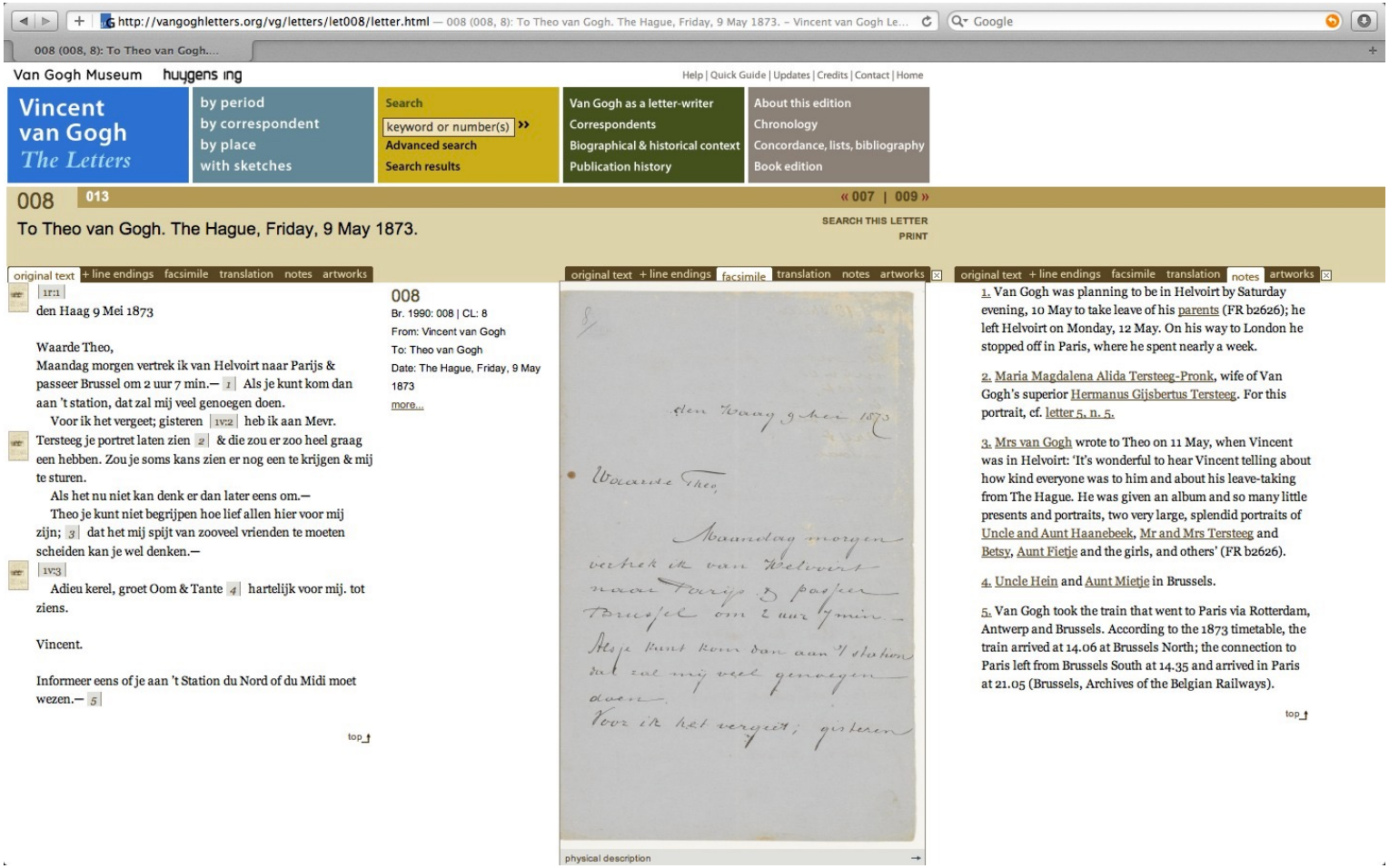


The Literary Bank, Strindberg section

Digital image and text mode side by side (Deutsches Textarchiv)



The van Gogh letters



mass digitization

	Primarily automated
	Breadth
	Minimized editing and metadata
	Clone
	Random and totality
	Text
	Quantitative

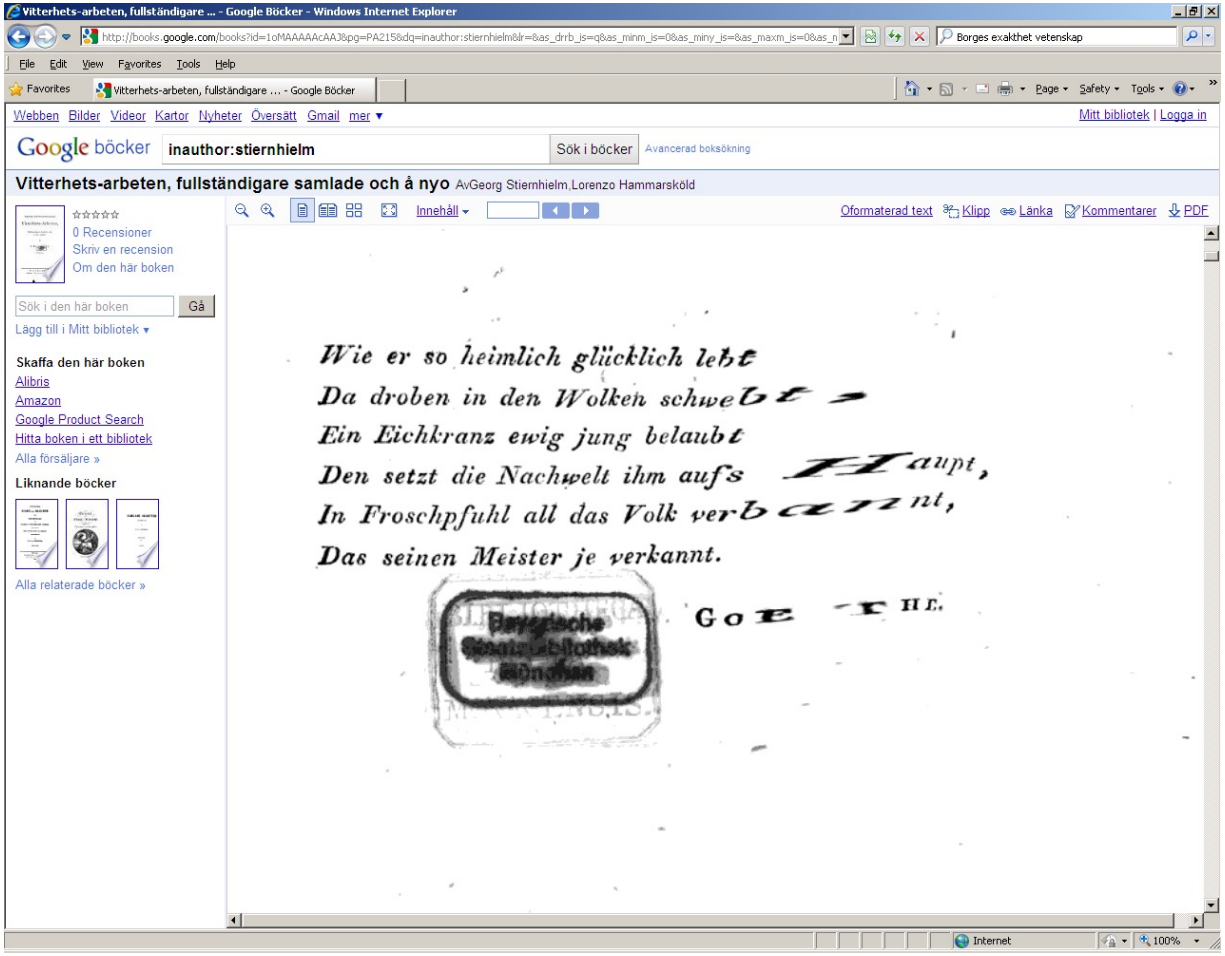
Google Books, Open Content Alliance (via Internet Archive), HATHI Trust, Europeana, Gallica

Mass (or at least large-scale) digitisation projects

Mass digitization: criticism

- Poor scanning quality
- Transcription poor or even useless
- Often only simple, raw free text search
- Few and often erroneous metadata
- Rudimentary bibliographical information
- Sometimes difficult or even impossible to ascertain the provenance of the digital representation – what do we actually see on the screen?
- Non-transparent contracts and terms conc. copyright, re-use, long-term availability and maintenance of the digital files

An early Stiernhielm in Google Books



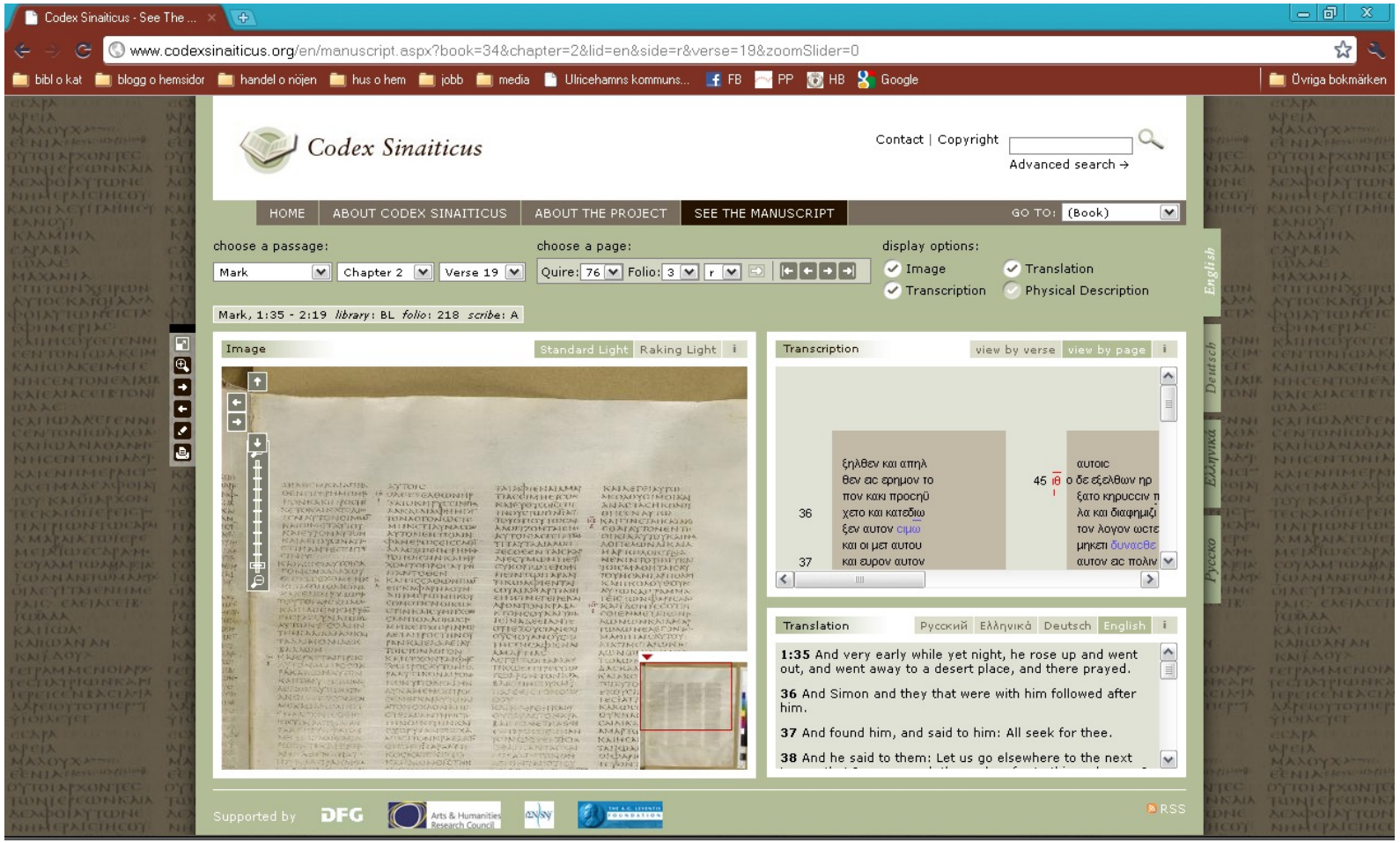
Scanned book on perception, images and form

Critical vs mass digitization

Primarily manual	Primarily automated
Depth	Breadth
Maximized editing and metadata	Minimized editing and metadata
Interpretation	Clone
Critical selection	Random and totality
Document	Text
Qualitative	Quantitative

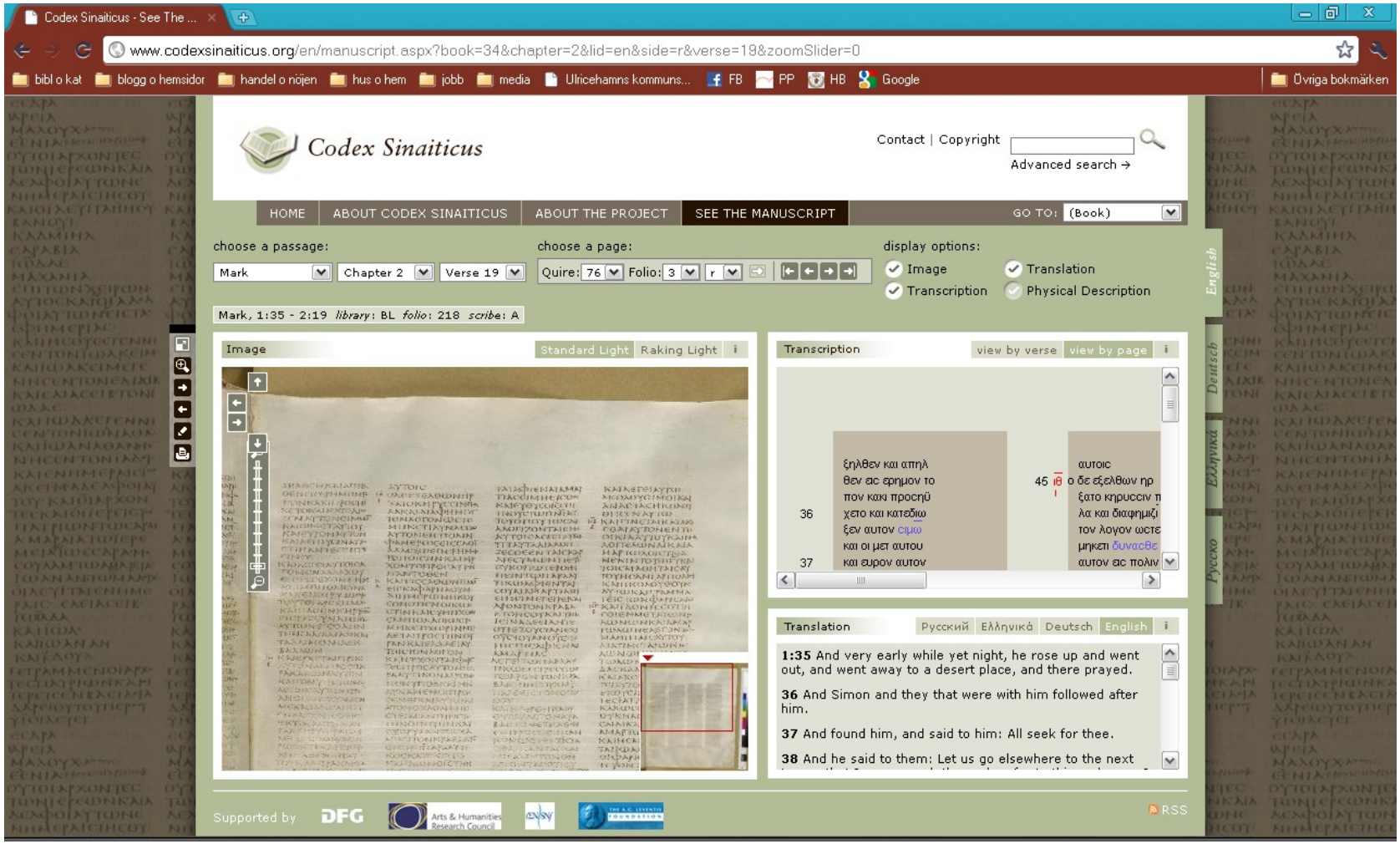
Codex Sinaiticus

”standard light”

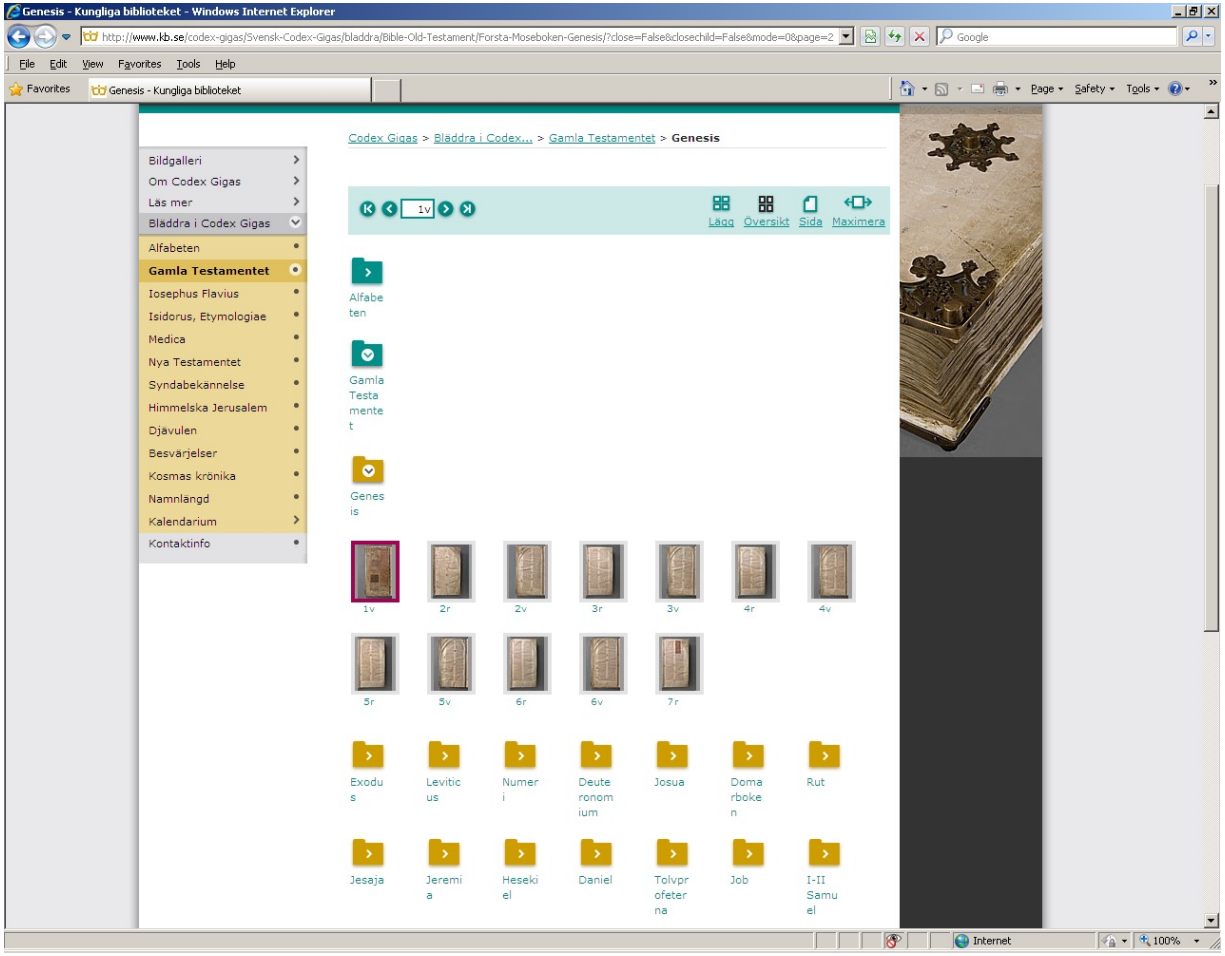


Codex Sinaiticus

”raking light”



Codex Gigas


Genesis - Kungliga biblioteket - Windows Internet Explorer
http://www.kb.se/codex-gigas/Svensk-Codex-Gigas/bladdra/Bible-Old-Testament/Forsta-Moseboken-Genesis/?close=False&closechild=False&mode=0&page=2
File Edit View Favorites Tools Help
Favorites
Genesis - Kungliga biblioteket
Page Safety Tools
Codex Gigas > Bläddra i Codex... > Gamla Testamentet > Genesis
Bildgalleri
Om Codex Gigas
Läs mer
Bläddra i Codex Gigas
Alfabeten
Gamla Testamentet
Iosephus Flavius
Isidorus, Etymologiae
Medica
Nya Testamentet
Syndabekännelse
Himmelska Jerusalem
Djävulen
Besvärjelser
Kosmas krönika
Namnlängd
Kalendarium
Kontaktinfo
1v
Lägg
Översikt
Sida
Maximera
Alfabeten
Gamla Testamentet
Genesis
1v
2r
2v
3r
3v
4r
4v
5r
5v
6r
6v
7r
Exodus
Leviticus
Numeri
Deuteronomium
Josua
Domarboken
Rut
Jesaja
Jeremia
Hesekiel
Daniel
Tolvprofeterna
Job
I-II Samuel
Internet
100%

Critical digitization: criticism

- Costly
- Slow
- Small audiences
- Necessitates skills in e.g. textual, historical or bibliographical scholarship
- Often neglects to produce machine readable transcriptions
- Sometimes lacks in documentation and ”repeatability”
- The material has been explicitly interpreted and narrowly formed for a particular purpose
- A cloning ideal?

Lecture on scholarship and access

- advanced digitisation
- digitisation and scholarship
- access
- use
- re-use

Some references

- Dahlström, M. (2015). Critical Transmission. In: P. Svensson & D. T. Goldberg (eds.), *Between Humanities and the Digital*. Cambridge: Mass., The MIT Press, p. 467-481.
- Dahlström, M. and W. Dillen (2017). Litteraturbanken: the Swedish Literature Bank. *RIDE* 6. doi: 10.18716/ride.a.6.2. https://ride.i-d-e.de/issues/issue-6/litteraturbanken-the-swedishliterature-bank/
- Deegan, Marilyn & Kathryn Sutherland (2009). *Transferred Illusions: Digital Technology and the Forms of Print*. Aldershot: Ashgate.
- Duffy, Christina (2015). *Revealing the secrets of the burnt Magna Carta*. The British Library. https://www.bl.uk/magna-carta/articles/revealing-the-secrets-of-the-burnt-magna-carta
- Pierazzo, Elena (2014). Digital Documentary Editions and the Others. *Scholarly Editing*, 35. http://scholarlyediting.org/2014/essays/essay.pierazzo.html
- Prescott, Andrew & Lorna Hughes (2018). Why Do We Digitize? The Case for Slow Digitization. *Archive Journal*. http://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/