



University of Stuttgart

Department *Digital Humanities*¹
*Stuttgart Research Center for Text Studies*²
Germany

Best practice for digitising small-scale Digital Humanities projects

Peggy
Bockwinkel¹,
Dilan Cakir²



Governmental digitising plans:

BOOKS 12/11/2013 05:34 pm ET

Norway Is Digitizing All Its Books

 By Sara Gates



JUPITERIMAGES VIA GETTY IMAGES

Norway is going digital.

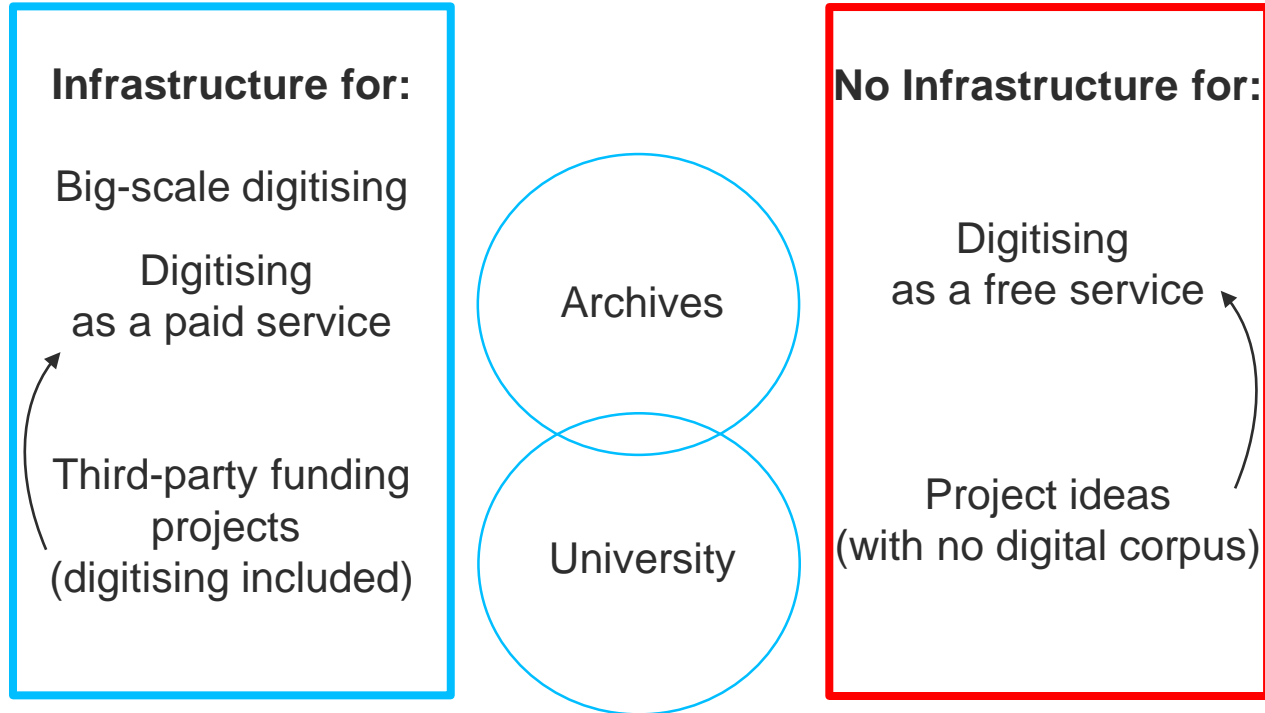
https://www.huffingtonpost.com/2013/12/11/norway-digitizing-all-books-national-library_n_4427164.html

Aim of the presentation

This presentation is intended to

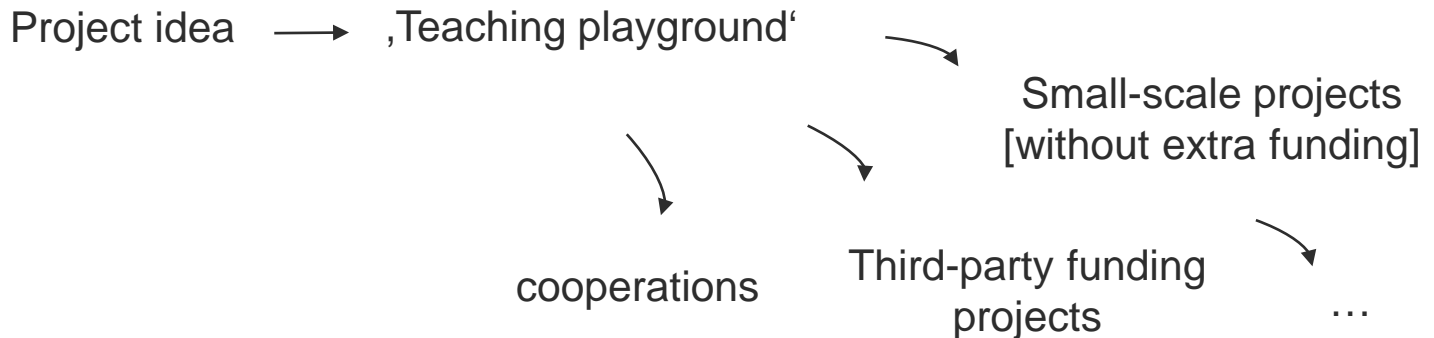
1. draw attention to the fact that there is no infrastructure that provides humanities scholars with machine-readable texts.
2. point out that third-party funding projects in the DH often start with an idea, that is tested in teaching.
3. remind, that DH-teaching can be seen as a playground, where ,things are tried out'. The support of DH-related teaching also supports DH research.
4. suggest some parameters that are important for digitising to make it easier for non-digitising experts to get their machine-readable text.

1. Digitising infrastructure in Germany



2. + 3. Importance of project ideas and teaching

Availability of machine-readable texts



If you boost project ideas with DH teaching, you boost DH research

Examples from the University of Stuttgart:

<http://www.ts.uni-stuttgart.de/kaetehamburger/>

http://www.germanliteratureglobal.com/index.php/Karten_und_Konjunkturgraphen_zu_%C3%9Cbbersetzungen_deutschsprachiger_Literatur

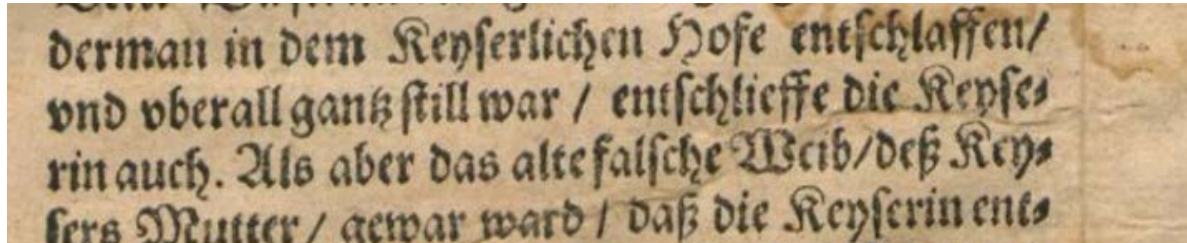
<https://quadrama.github.io/index.en>

4. Digitising (Scanning)

1st parameter: the book itself

Sample with good results (colour/400dpi):


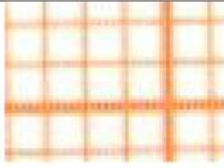


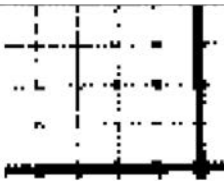


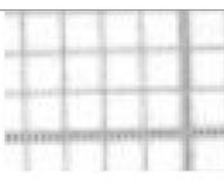
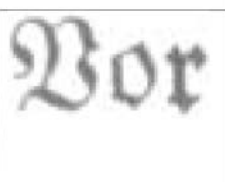
Sample with unacceptable results (settings: ?)



- Bad results due to
 - the bad quality of the book (stains, folds, etc.)
 - the use of initials as part of the typography

4. Digitising (Scanning + OCR)

2nd parameter: settings (colour + dpi)

colour settings	dpi	„sun picture“	millimetre paper	detail of gothic types	number of mistakes in one double page, gothic types
colour	300				7
black/white	300				17
grey scale	300				9

→ Best setting: colour, 300dpi

4. Digitising (Scanning + OCR)

3rd parameter: types of errors

	e →c	u →n	st →si	st →il	ss →st	st →ss	st →ff	b →d	m →n	l →i	s →j	s →f	b →h	f →f	i →l	c →:	e →o	e →h	L →T
colour	300		1							1	1	1	2	1					
black/ white	300	1	6						3						1	1	1	1	1
grey scale	300	1	5	1	1			1											

→ Best setting: grey scale, 300dpi

→ if you train on „st“: 5 errors less (per double page)

4. Digitising (OCR)

Postscript

- For texts in Antiqua:

Also consider to OCR with Adobe professional

→ It might be faster with better results

- Other OCR software, e.g.:

PoCoTo (Post Correction Tool):

<https://github.com/cisocrgroup/PoCoTo>

<https://thorstenv.github.io/PoCoTo/>

- A list with more parameters on digitising:

Holley, R. (2009). How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine* 15 (3/4).



University of Stuttgart
Germany

Thank you! Questions? Comments?



Peggy Bockwinkel

github: <https://github.com/Bockwinkel/OCR>

e-mail peggy.bockwinkel@ilw.uni-stuttgart.de

phone +49 (0) 711 685-82279

fax +49 (0) 711 685-

University of Stuttgart
Department *Digital Humanities*
Herdweg 51, D-70173 Stuttgart