



Sustainable digitization

Denis Roegel

► To cite this version:

Denis Roegel. Sustainable digitization. [Research Report] LORIA - Université de Lorraine. 2012. hal-02340532

HAL Id: hal-02340532

<https://inria.hal.science/hal-02340532v1>

Submitted on 30 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sustainable digitization

(*La numérisation durable**)

Denis Roegel

Université de Lorraine & LORIA

21 December 2013

* (a French version is also available on <http://locomat.loria.fr/other/num.html>)

Preamble

Numerous questions

- do we have to digitize?
- what should we digitize?
- why?
- how?
- who should digitize?
- what is sustainable digitization ?

The purpose of this presentation is to provide **elements of answers** to these questions.

Sustainable digitization: a new concept?

- In the 21st century, **sustainability** has become the leading word.
- But, oddly, it has only very little been used for digitization.
Why?
- This concept does not seem well understood and recently, a professor from ENSSIB (Library school in France) told me: “To my knowledge, there is no such thing as sustainable digitization.”
- Is that really the case?

Our context

- in the past 20 years, we have become acutely aware of two fundamental aspects of the quality of documents: that of **mathematics** and that of **graphical objects**;
- more recently, the frequentation and intensive use of digitized documents has led us to have a **critical approach**;
- the **needs of the users**, the **needs of the documents**, and the **dynamics of digitization** and of its actors have been analyzed;
- exchanges with libraries made it possible to reach a better understanding of the **objectives** and **a prioris** of each one involved;
- our analysis has in particular been stimulated by our work of **reconstruction of historical documents**, the LOCOMAT project.

A word on the LOCOMAT project

LOCOMAT (<http://locomat.loria.fr>) is a site dedicated to mathematical and astronomical tables, with currently two highly developed sides:

- ① a **reconstruction of historical tables** (100000 pages)
 - ② a **census of all digitized tables**, from all times, on all supports and in all languages (almost 2000 books)
- the LOCOMAT census provides a (probably) representative overview of the digitization of mathematical and astronomical tables
 - the repositories, interfaces and digitization qualities can be compared.

The problem of quality

A large part of this presentation concerns the notion of **quality**.

- a prerequisite for the sequel of the presentation is to aim for a certain quality;
- **this presentation is therefore not meant for those who would have only commercial or economic interests;**
- the quality is related to the sustainability: we do not want to do again the work which has already been made;
- if quality and sustainability are of concern to you, read on; otherwise, do not read further.

The context of Archives (1)

Before we start, it is necessary to say a word on digitization in Archives. There is a fundamental difference between *archives* and *libraries*:

- libraries keep books which are naturally aimed at being distributed;
- archives keep documents which are usually unique, and were not aimed at being distributed.

The context of Archives (2)

This difference entails that archivists do not feel concerned by what takes place in the libraries. For them, digitization satisfies two needs:

- the need to protect the original documents
- the need to have an access to a few particular types of documents (parish registers, vital records, notary records)

In addition, it should be noticed that:

- digitization makes it possible to avoid the congestion of Archives (the original documents cannot be taken out);
- digitization primarily concerns manuscripts;
- there is no long-term vision of digitization andx the notion of sustainable digitization does not seem to be meaningful to archivists.

The context of the Archives (3)

But, as it is the case in libraries, the digitization process of archives will develop:

- repositories for digitization will continue to be selected among those not yet digitized and, little by little, we will see appear the target of total digitization;
- even if the archive documents have not been planned for distribution, it will happen because the needs will appear.

At the level of books, the digitization policy seems clearer, but it is the only difference between books and archive documentsx.

The context of the Archives (4)

- the technical problems are in fact common, even though some archivists consider that digitizing books is a mundane matter;
- if books were really so simple to digitize, one wonders why it is almost always badly done;
- in any case, the archivists must be familiar with the problem of digitization of documents which are possibly simpler than theirs.

The context of the Archives (5)

- eventually, once the target of priorities is passed, sustainable digitization becomes a reality, even if it is one very distant in time (and in the same way as it has been *artificially* put on a short-term schedule for libraries);
- almost our entire analyses are relevant to archives, but in a context of sustainability.

Important notice:

This document is not a digitization manual
and only some practical and technical aspects are mentioned.

We can now start.

End of preamble

Outline

- 1 The quality of the digitizations
- 2 The quality of interfaces
- 3 The sources of the technical problems
- 4 The problems of digitization policies
- 5 Solutions

The quality of the digitizations

- in the following pages, we will give an overview of a number of representative digitizations;
- these digitizations will be commented and this will make it possible to extract needs and traps to avoid.

Examples of bad digitizations

Andoyer: « La théorie de la lune » (1902)

Gallica (gallica.bnf.fr)
(using a microfilm
probably almost as bad)

11

THÉORIE DE LA LUNE

lité à longue période de la longitude, on a, par l'équation (15),
d'après les propriétés de la fonction B :

$$\begin{aligned} & (1-m-g_0)b_{41} + 2(1-m)b_1a_{17}^2 + 2(1-g_0)b_{39}\left(\frac{3}{8}m^2\frac{2}{2(1-m)}\right) \\ & + (1-g_0)b_{17}\left[2a_1a_{17} + \frac{3}{8}m^2(-2a_{17}-2b_{17})\frac{2}{1-2m-g_0} + a_{17}\frac{3}{8}m^2\frac{2}{2(1-m)}\right] \\ & - 3a_1a_{17}^2 - 6a_0a_{17}a_{19} + 44(m-g_0)^2a_{17}a_{19} \\ & + 10m^2a_{17}a_{19} + \frac{3}{4}m^2[-7a_{39}-8b_{39}+10a_{17}^2+14a_{17}b_{17}+8b_{17}^2] = 0. \end{aligned}$$

The formula typeset with a
modern tool.

$$\begin{aligned} & 4(m-g_0)b_{41} + 2(1-m)b_1a_{17}^2 + 2(1-g_0)b_{39}\left(\frac{3}{8}m^2\frac{2}{2(1-m)}\right) \\ & + (1-g_0)b_{17}\left[2a_1a_{17} + \frac{3}{8}m^2(-2a_{17}-2b_{17})\frac{2}{1-2m-g_0} + a_{17}\frac{3}{8}m^2\frac{2}{2(1-m)}\right] \\ & - 3a_1a_{17}^2 - 6a_0a_{17}a_{19} + 44(m-g_0)^2a_{17}a_{19} \\ & + 10m^2a_{17}a_{19} + \frac{3}{4}m^2[-7a_{39}-8b_{39}+10a_{17}^2+14a_{17}b_{17}+8b_{17}^2] = 0. \end{aligned}$$

Examples of bad digitizations

Andoyer: « La théorie de la lune » (1902)

The same (useless) excerpt
of Gallica

11

THÉORIE DE LA LUNE

lité à longue période de la longitude, on a, par l'équation (1),
d'après les propriétés de la fonction B :

$$\begin{aligned} & (1-m-g_0) b_{11} + 2(1-m) b_1 a_{17}^2 + 2(1-g_0) b_{39} \left(\frac{3}{8} m^2 \frac{2}{1-m} \right) \\ & + (1-g_0) b_{17} \left[2a_1 a_{17} + \frac{3}{8} m^2 (-2a_{17} - 2b_{17}) \frac{2}{1-2m-g_0} \right. \\ & \quad \left. + a_{17} \frac{3}{8} m^2 \frac{2}{2(1-m)} \right] \\ & - 3a_1 a_{17}^2 - 6a_0 a_{17} a_{19} + 44(m-g_0)^2 a_{17} a_{19} \\ & + 10m^3 a_{17} a_{19} + \frac{3}{4} m^2 [-7a_{39} - 8b_{39} + 10a_{17}^2 + 14a_{17} b_{17} + 8b_{17}^2] = 0. \end{aligned}$$

The same excerpt obtained
by scanning a photocopy
of the original:

$$\begin{aligned} & 4(m-g_0) b_{11} + 2(1-m) b_1 a_{17}^2 + 2(1-g_0) b_{39} \left(\frac{3}{8} m^2 \frac{2}{1-m} \right) \\ & + (1-g_0) b_{17} \left[2a_1 a_{17} + \frac{3}{8} m^2 (-2a_{17} - 2b_{17}) \frac{2}{1-2m-g_0} \right. \\ & \quad \left. + a_{17} \frac{3}{8} m^2 \frac{2}{2(1-m)} \right] \\ & - 3a_1 a_{17}^2 - 6a_0 a_{17} a_{19} + 44(m-g_0)^2 a_{17} a_{19} \\ & + 10m^3 a_{17} a_{19} + \frac{3}{4} m^2 [-7a_{39} - 8b_{39} + 10a_{17}^2 + 14a_{17} b_{17} + 8b_{17}^2] = 0. \end{aligned}$$

Examples of bad digitizations (cont'ed)

- In 2009 or 2010, the city of Turin digitized an important book by Giovanni Plana on the theory of the Moon, a total of about 2000 pages.
- Plana was from Turin and it was a valorization of its patrimony
- <http://www.accademiadelle scienze.it/TecaRicerca>
- but:
 - the site is difficult to use, because it requires a configuration which is not universal
 - the resolution is not sufficient
 - downloading is impossible
- this digitization was therefore a mere **waste**
- in the meantime, the book was (slightly) better digitized on Internet Archive.

Examples of bad digitizations (cont'ed)

Plana: « Théorie du mouvement de la lune » (1832), vol. 2, p. 216
(Turin version) **exponents are difficult to read**

Biblioteca Digitale Accademia delle Scienze Torino

- Torna al Modulo Ricerca -

Visualizzatore immagini by (c) Mx Ver. 2.1.5 Zoom: 251%

Versione: 3.2.15 - 30 gennaio 2009

Non riesci a visualizzare la pagina?

Examples of bad digitizations (cont'ed)

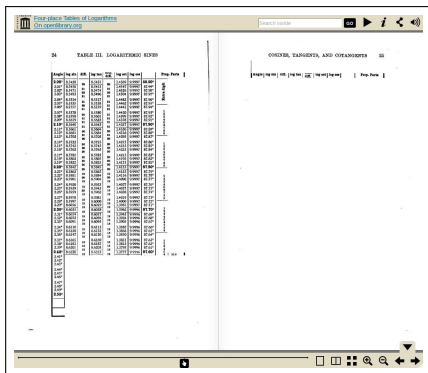
Plana: « Théorie du mouvement de la lune » (1832), vol. 2, p. 216
(Archive.org version) **exponents are difficult to read**

$$\begin{aligned}
 (4) \dots\dots\dots - R_1 \cdot \frac{ds_1}{dv} = \\
 g\nu \quad e\gamma \left\{ \frac{3}{2} - \frac{3}{2} m + \frac{9}{8} m^2 + \frac{9}{8} e^2 - \frac{15}{4} \varepsilon'^2 - \left(\frac{3}{8} + \frac{3}{8} = \right. \right. \\
 g\nu \quad e\gamma \left\{ \frac{3}{2} + \frac{3}{2} m + \frac{9}{8} m^2 + \frac{9}{8} e^2 - \frac{15}{4} \varepsilon'^2 - \left(\frac{3}{8} + \frac{3}{8} = \right. \right. \\
 -g\nu \quad e^2\gamma \left\{ - \frac{15}{8} + \frac{57}{16} m \right\} \\
 \left(\frac{15}{8} + \frac{57}{16} m \dots \right) / 2991 \quad 45 \quad 2271 \backslash \dots \quad 15 \dots
 \end{aligned}$$

(Archive and Google apply thresholds which reduce the quality.)

Examples of bad digitizations (cont'ed)

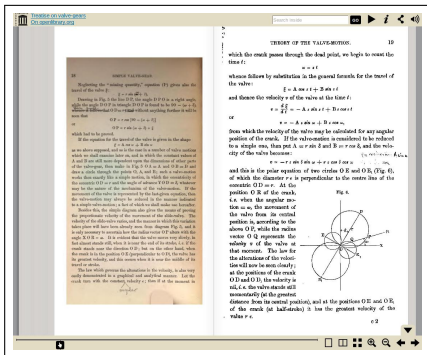
Since Archive includes Google's digitizations, one should not be surprised to find some pages totally messed up:



- Granville: *Four-place tables of logarithms*, 1908
- id = fourplacetales00grangoog

Examples of bad digitizations (cont'ed)

Archive displays inconsistent digitizations:



- Zeuner, *Treatise on valve-gears*, 1869
- id = treatiseonvalve01zeungoo
- left and right pages differ

Felkel: « Tafel aller einfachen Faktoren » (1776)
(<http://resolver.sub.uni-goettingen.de/purl?PPN620754826>)

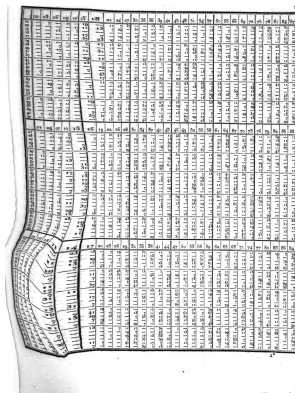
		Factoren von 3600 bis 42000.																												
		t	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
3600	t	1	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z		
	3600	1	a	b	c	d	e	f	g																					

Examples of bad digitizations (cont'ed)

Google Books

(<http://books.google.com/books?id=XmpbAAAAQAAJ>)

Book digitized in Oxford (Bodleian Library), 26 February 2009



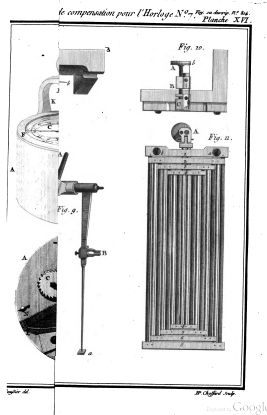
Digitized by Google

Examples of bad digitizations (cont'ed)

Google Books

(<http://books.google.com/books?id=RVtDAAAACAAJ>)

Book digitized at Ghent university (Belgium), July 8, 2010



Examples of bad digitizations (cont'ed)

Here, we have an example of digitization by layers: certain areas are sharper than others. This is probably a digitization based on the JPEG2000 format.

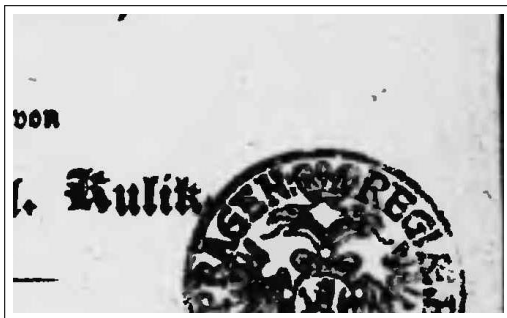
<http://kramerius.nkp.cz/kramerius/MShowMonograph.do?id=25654>



- view of the first page of the book
- the book is displayed as DJVU images, with the possibility to produce partial PDF files
- here, the layered digitization is a total failure

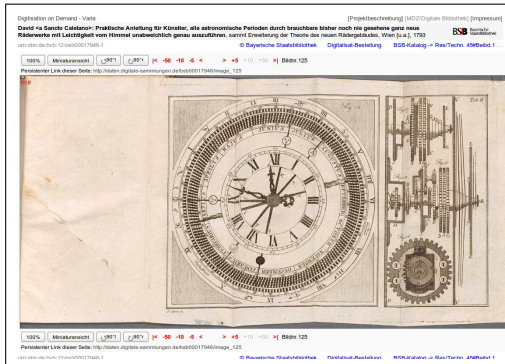
Examples of bad digitizations (cont'ed)

The JPEG2000 format is often incorrectly used (in particular by Archive.org):



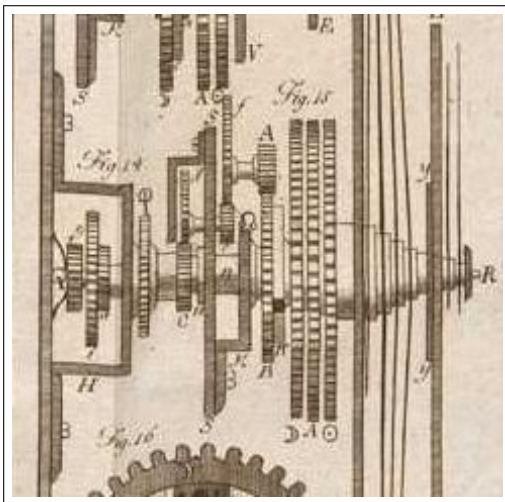
- detail of the previous view
- JPEG2000 should only be used for a good reason, and here, there is a bad usage!
- the selective fuzziness is not what the users are looking for, except perhaps those who have progressive lenses;
- the JPEG2000 format is used here in a (very) lossy way.

Examples of bad digitizations (cont'ed)



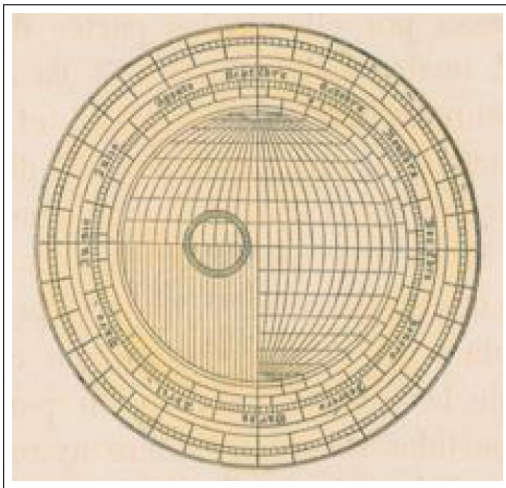
- a figure in a book from Munich
- maximal resolution

Examples of bad digitizations (cont'ed)



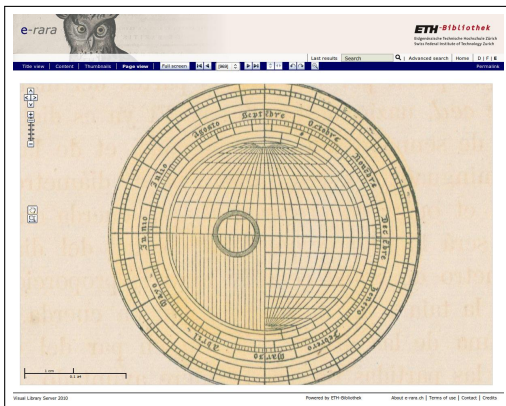
- detail of the previous view
- the resolution displayed is clearly too low

Examples of bad digitizations (cont'ed)



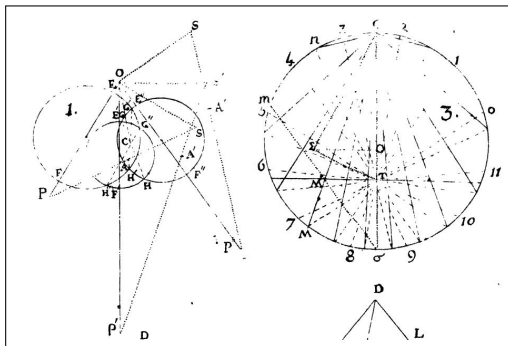
- e-rara project (Zurich)
- a figure of the book *Libros del Saber*, volume 3, p. 143
- inscriptions totally unreadable in the downloaded PDF

Examples of bad digitizations (cont'ed)



- online, however, it is possible to obtain a better quality

Examples of bad digitizations (cont'ed)

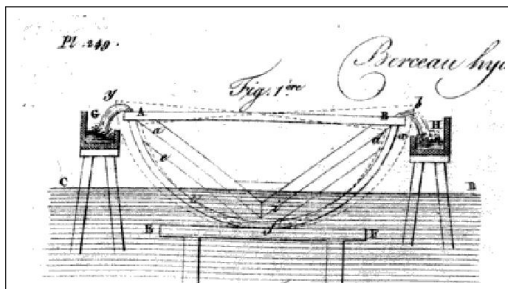


- NUMDAM project (<http://www.numdam.org>, digitization of mathematics journals)
- a geometric figure (Gergonne's annals, volume 17)
- a large part of the lines have vanished because the resolution was not sufficient

Examples of bad digitizations (cont'ed)

Conservatoire numérique des Arts et Métiers
(CNUM, <http://cnum.cnam.fr>)

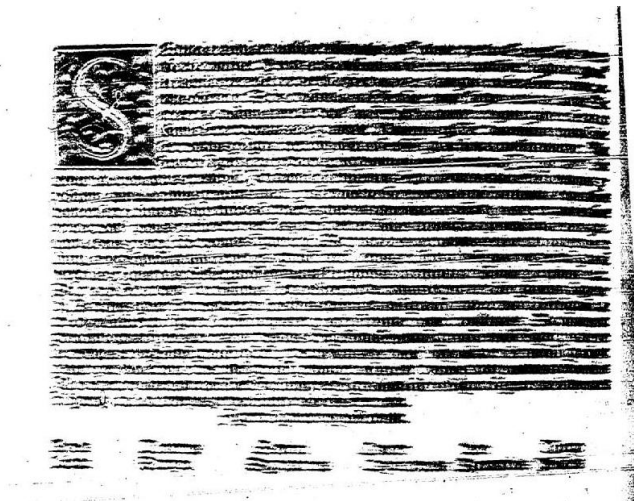
Here, excerpt of the *Bulletin de la Société d'Encouragement*, 1823.



- the figure was digitized with an insufficient resolution
- the plate number (249) cannot be read with certainty
- the plate was not aligned properly
- a new digitization is absolutely necessary

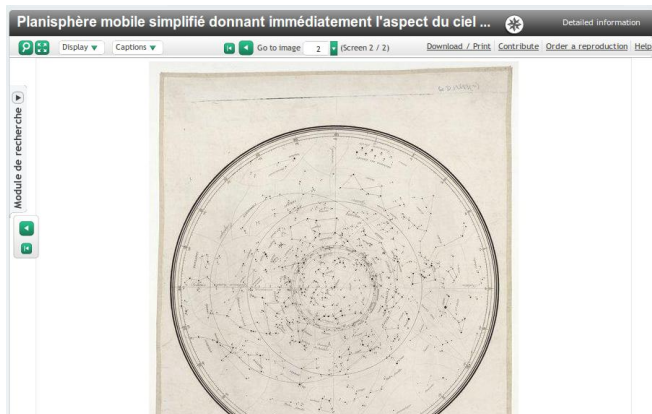
Examples of bad digitizations (cont'ed)

Gallica: *Astrolabium planum in tabulis ascendens*, 1494



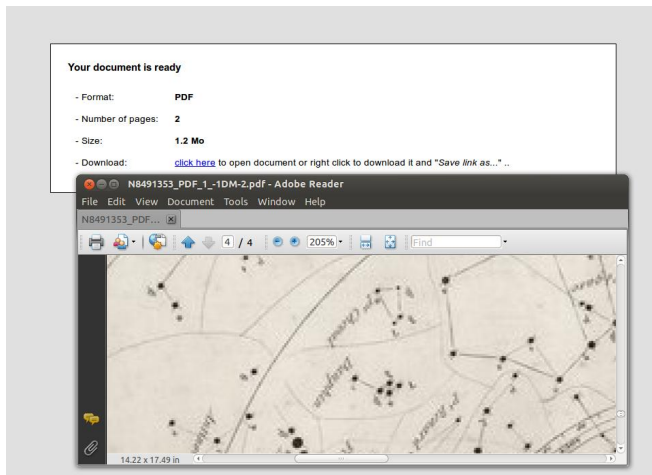
Examples of bad digitizations (cont'ed)

Gallica : *Planisphère mobile simplifié*, 1839



Examples of bad digitizations (cont'ed)

Gallica : *Planisphère mobile simplifié*, 1839



Examples of bad digitizations (cont'ed)

Archive.org : 數理精蘊 (Shuli Jingyun, 1723)

(<http://www.archive.org/details/06076323.cn>)

五〇三〇一	五〇二八一	五〇二六一	五〇二四一	五〇二二一	五〇二〇一	五〇一八一	五〇一六一	欽 定 四 庫 全 書
五〇三〇二	五〇二八二	五〇二六二	五〇二四二	五〇二二二	五〇二〇二	五〇一八二	五〇一六二	
五〇三〇三	五〇二八三	五〇二六三	五〇二四三	五〇二二三	五〇二〇三	五〇一八三	五〇一六三	
五〇三〇四	五〇二八四	五〇二六四	五〇二四四	五〇二二四	五〇二〇四	五〇一八四	五〇一六四	
五〇三〇五	五〇二八五	五〇二六五	五〇二四五	五〇二二五	五〇二〇五	五〇一八五	五〇一六五	
五〇三〇六	五〇二八六	五〇二六六	五〇二四六	五〇二二六	五〇二〇六	五〇一八六	五〇一六六	
五〇三〇七	五〇二八七	五〇二六七	五〇二四七	五〇二二七	五〇二〇七	五〇一八七	五〇一六七	
五〇三〇八	五〇二八八	五〇二六八	五〇二四八	五〇二二八	五〇二〇八	五〇一八八	五〇一六八	

(probably from a microfilm)

Examples of bad digitizations (cont'ed)

Archive.org : 數理精蘊 (Shuli Jingyun, 1723)

The same, but better:

五〇三〇一	五〇二八一	五〇二六一	五〇二四一	五〇二二一	五〇二〇一	五〇一八一	五〇一六一
二四三 二〇七	六五三 七七		一六七四七 三		二九五三 一七	三八九 一二九	四八七 一〇三
五〇三〇二	五〇二八二	五〇二六二	五〇二四二	五〇二二二	五〇二〇二	五〇一八二	五〇一六二
三五九三 一四	八一 六二	八三七七 六	二五一二 二	二五一二 二	二七八九 一八	二二八 二二	三五八三 一四
五〇三〇三	五〇二八三	五〇二六三	五〇二四三	五〇二二三	五〇二〇三	五〇一八三	五〇一六三
二六九 一八七	三三三 一五一		一〇六九 四七	一六七四 一三	八二三 六一	四六九 一〇七	七二七 六九
五〇三〇四	五〇二八四	五〇二六四	五〇二四四	五〇二二四	五〇二〇四	五〇一八四	五〇一六四
二六二 二九二	九六七 五二	二四四 二〇六	二三七 二二	二九二 二七二	三〇八 一六三	二四六 二〇四	一二五四 一四
五〇三〇五	五〇二八五	五〇二六五	五〇二四五	五〇二二五	五〇二〇五	五〇一八五	五〇一六五
一〇〇六 一五	四四五 一一三	一一一七 四四	七七三 六五	二四五 二〇五	三三四七 一五	一〇〇三 七五	三九五 一二七
五〇三〇六	五〇二八六	五〇二六六	五〇二四六	五〇二二六	五〇二〇六	五〇一八六	五〇一六六
二五一五 三二	二八九 一七四	六三一 八二	二五九 一九四	七六一 六六	一九三 二六	一〇九一 四六	九二九 五四
五〇三〇七	五〇二八七	五〇二六七	五〇二四七	五〇二二七	五〇二〇七	五〇一八七	五〇一六七
四〇九 一二三		三〇一 一六七	一八六 二七			一六七二 九三	二二七 二二

Are these digitizations sufficient?

- if the purpose is to have an overview of a document, these digitizations are sufficient;
- for the last example, we can recognize that it is in Chinese, and most people will be content with it;
- but in order to really read what is written, it is not enough;
- moreover, it is easy to see that **the problem is not a display problem, but a problem at the source**;
- **with some digitizations, reading becomes archeology!**
- should reading a digital document be painful?

Quality level and digitizer

The average quality is clearly correlated with the institution which is responsible of digitizing.

- Google books and Internet Archive:
 - average quality, and sometimes very low quality
 - never very good
- German libraries
 - average quality very good
 - sometimes less good on specific aspects
- some institutions produce digitizations and interfaces with a very low or suspicious quality:
 - Digital Library of India
 - Universal Digital Library (Carnegie Mellon, digitizations subcontracted in India and China)
- for many sites, **the quality increases with time**

The case of Google books (1)

- only driven economically;
- the purpose is to make digitized books available on Google's e-book platform;
- Google is not concerned by patrimony, but by the **control of written information**;
- low quality of digitization (but improving, for instance through **color digitization since mid 2009**):
- a large number of libraries have been trapped by Google books;
- access limits even for out of copyright books, and varying from one country to another;

The case of Google books (2)

Books digitized by Google have numerous flaws:

- some pages may be blurred, fingers can be visible, etc.
- folded plates of older books are never unfolded;
- some of these defects are not accidental, but are the result of instructions;
- those who digitize for Google have instructions not to unfold the plates; if that were not the case, we would sometimes see unfolded plates;
- Google doesn't mind to sell books which have been incompletely digitized.

The case of Google books (3)

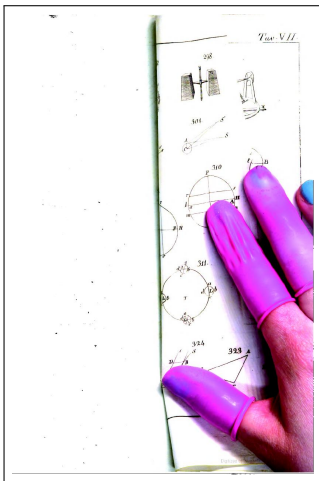
The fingers of an operator



Ozanam: *Récréations mathématiques*, 1696,
Google at the Munich library

The case of Google books (4)

Unfolded and crushed plate

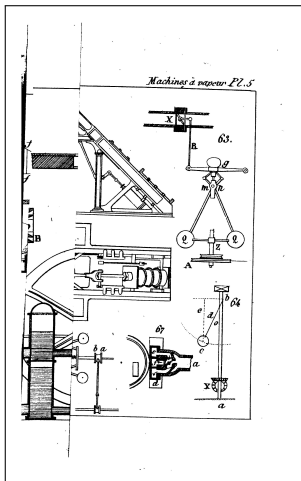


Baumgartner: *La fisica congiunta alle matematiche*, volume 3, 1828

Google at the Austrian National Library, 2012

The case of Google books (5)

Unfolded plate

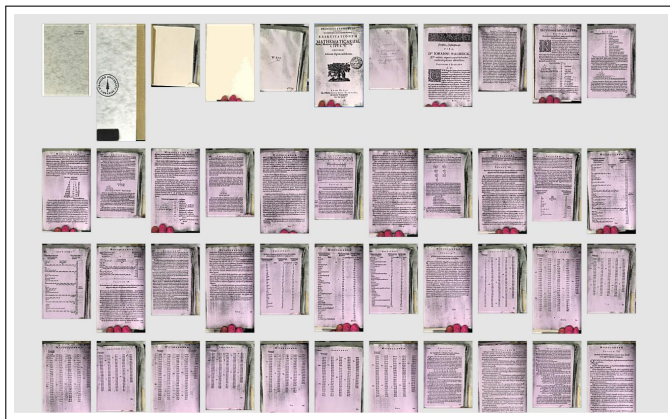


*Traité complet des machines à vapeur
appliquées à l'industrie, 1838*

Google at the Bavarian State Library, 2012

The case of Google books (6)

More fingers



Schooten, *Exercitationum mathematicarum*, 1657,
Google at Stanford

The case of Google books (4')

Even more fingers



The case of Google books (7)

There is a rather general ignorance of Google's technical problems, even in libraries:

- in their books on Google, neither Jean-Noël Jeanneney (former head of the BNF, French National Library, critical towards Google), nor Bruno Racine (current head of the BNF, rather open to a dialog with Google), do mention Google's quality problems, and probably do not know them;
- in general, independently from digitization errors, the matter of insufficient resolution is never raised.

The case of Google books (8)

- in his book on Google books (*Google Livres et le futur des bibliothèques numériques*, 2010), Alain Jacquesson (former head of the library of Geneva) devotes two pages to the control of quality; for the author, Google's errors are all accidental and he believes Google's promise to correct them; he seems to believe that books with foldouts do not enter the digitization chain; he also seems to ignore that badly digitized books are still put on sale;
- Jacquesson mentions the "manipulation errors when digitizing" and adds that "Google gives precise instructions to its digitization operators."

The case of Google books (9)

Google has introduced a number of **technical innovations**, but they are no wonders:

- Google has filed a patent for correcting curvature effects resulting from thick non-plane books;
- on certain views, we can observe obvious image corrections; these corrections produce such deformations that **the result is almost worse than without corrections**;
- the least that can be said is that Google's patent is not satisfactorily, or not well applied!

It appears that libraries and other book professionals are very little aware of Google's technical problems.

The case of Google books (10)

Concerning foldouts:

- according to Alain Jacquesson, the books which contain foldouts are excluded from digitization
- the fact that such books are digitized is then either an error of Google which takes them nevertheless, or an error of the providing library;
- for consistency, Google should destroy a digitization whenever it contains foldouts that would lead to an incomplete digitization;

Google therefore prefers to digitize badly (and on purpose), than not at all.

The problem of pseudo-publishers (book-recyclers) (1)

- probable contracts between Google and “fake” publishers such as Bibliobazaar = BiblioLife = Nabu Press, Kessinger, Pranava, Woods Press, etc.;
- the digitizations are those of Google, but without the Google logo; these uses are therefore authorized;
- the printing is on demand, but the catalogues are filled of these non-existing books;
- Google is possibly behind the above names;
- the recycled versions are *shadowing* the original versions;
- the recycled versions are put in front of searches, in contradiction with Google’s will to fight “content farms” and low-value pages, but in accordance with the probable collusion;

The problem of pseudo-publishers (2)

- the above “fake” publishers do not seem to overlap, and they seem to divide the market; a given book is not published by more than one of these publishers;
- durability problem: when an old version is reprinted by one of these pseudo-publishers, the original version may vanish from Google books (for an example, see the Wikipedia page of Kessinger Publishing);

Numéro de notice : [077743806](#)

Titre : [Aesthetic measure](#) [Texte Imprimé] / George D. Birkhoff

Alphabet du titre : latin

Auteur(s) : [Birkhoff, George David \(1884-1944\)](#). Auteur

Date(s) : [s.d.]

Langue(s) : anglais

Pays : Etats-Unis

Editeur(s) : [[S.L.](#)] : [Kessinger](#), [s.d.]

Description : xii, [1] p., 2 l., 3-225, [1] p. : illus. (incl. music) XXIII pl. (part col.) diagrs. ; 30 cm

Notes : Reprod. en fac-similé de l'ed. Harvard University Press de 1933

Sujets : [Proportions \(art\)](#)
[Musique -- Philosophie](#)
BH201. .B5

Origine de la notice : AKR

- a SUDOC record for a book recycled by Kessinger
- little by little, the real books will be replaced by bad quality recycled copies (invasion of the body-snatchers?)

The problem of pseudo-publishers (3)

An example of a search on AddALL with 52 answers, 22 of which POD (print-on-demand) books from Woods Press:

Save the Info	Sort Asc TITLE Sort Desc	Asc AUTHOR Desc	Asc PRICE USD Desc	Asc SITE Desc	Asc DEALER Desc	DESCRIPTION
	1 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Claudius Saunier	24.41	AmazonCA	AmazonCA	NEW, Woods Press 2010-11 Paperback 144650672X 24.65 CAD to USD is calculated base on 1 CAD = 0.99018 USD
	2 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Claudius Saunier	39.80	AmazonUK	AmazonUK	NEW, Unknown 2010-11 Paperback 144650672X 25.46 GBP to USD is calculated base on 1 GBP = 1.56336 USD
	3 TraitÃ© d'horlogerie moderne thÃ©orique et pratique (French Edition) Buy it!	Claudius Saunier	40.95	Amazon	Amazon	NEW, Woods Press 2010-11-01 Paperback 144650672X
	4 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Claudius Saunier	41.97	AmazonFR	AmazonFR	NEW, Woods Press 2010-11 BrochÃ© 144650672X 31.31 EUR to USD is calculated base on 1 EUR = 1.34055 USD
	5 Traite D'Horlogerie Moderne Theorique Et Pratique Buy it!	Saunier, Claudius	45.33	Abebooks	Paperbackshop-US	[publisher: WOODS PR 01/01/2012] New print on demand book. Shipped from US. This item is printed on demand. [Aurora, IL, U.S.A.]
	6 Traite D'Horlogerie Moderne Theorique Et	Claudius Saunier	45.79	AmazonFR	AmazonFR	USED, Woods Press 2010-11 BrochÃ© 144650672X

Bad paper digitizations

The problem of bad digitizations does not concern only online books:

- certain new books are of insufficient quality, even from respectable publishers;
- certain reprints display a lack of digitization skills.

The case of Cambridge University Press

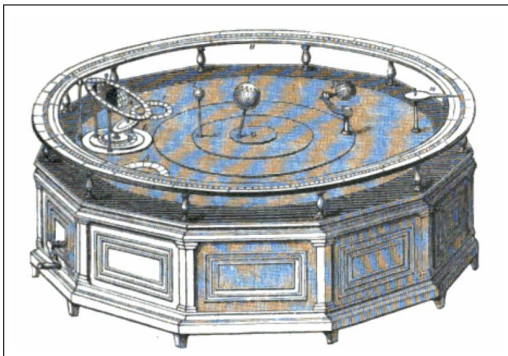
In 2010, Cambridge University Press reprinted Henderson's book, *Life of James Ferguson*, 1867.

There are two problems here:

- it is not clear why CUP chose to reprint the 1867 edition, since a corrected and extended edition was published in 1870;
- a large number of drawings come out blurred, as if their treatment had used an insufficient resolution; oddly, this is not the case of all figures.

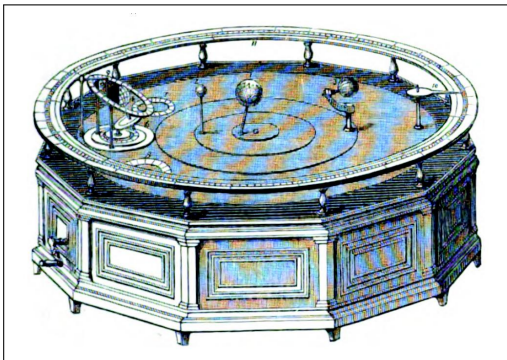
The case of Cambridge University Press (2)

We can compare the figures from 1870 and from the 2010 reprint on Google Books, and also compare the 1870 figures online and in PDF, as these are not exactly the same.



- figure page 73
- 1870 online version

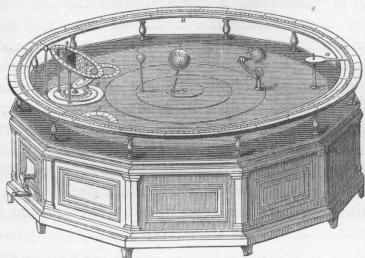
The case of Cambridge University Press (3)



- 1870 PDF version
- more contrasted than the online version

Le cas de Cambridge University Press (4)

preserves its parallelism during its annual course ; thus exhibit-



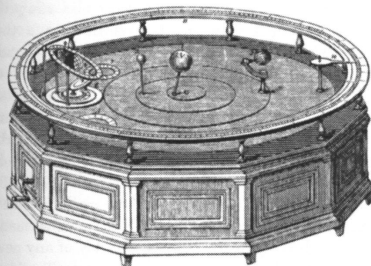
Ferguson's Orrery.

ing her different length of days and nights; her change of sea-

- 1867 version (digitized from an original copy)
- same sharpness than the 1870 version

The case of Cambridge University Press (5)

preserves its parallelism during its annual course ; thus exhibit-



Ferguson's Orrery.

ing her different length of days and nights; her change of sea-

- 2010 online version (digitized from the reprint)
- clearly blurrier than the previous version: the figure has become a halftone
- this problem is also visible on Google books' version of the reprint

The Tessier report

Rapport sur la numérisation du patrimoine écrit

(available on www.culture.gouv.fr)

- report ordered by the French ministry of culture following discussions between the BNF (French National Library) and Google;
- the report was written by Marc Tessier and completed on January 12, 2010
- gives the current state of digitization
- proposes solutions

In the context of the “mediocrity of the digitizations made by Google,” the report mentions (p. 51) that the BNF has now switched from 300 to 400dpi. It is better, but it is not yet sustainable!

Resolution problems

There are essentially two kinds of resolution problems:

- the most common problem is that of a resolution which is too low;
- but the opposite problem, although less known, also occurs.

Overdigitization problems (1)

Example of a plate in a book digitized at Strasbourg (call number H349), digitized with 400 or 500dpi.



Overdigitization problems (2)

Close-up of the previous image:



Overdigitization problems (3)

Close-up of the previous image:



- to the left of the characters: some green; to the right: some yellow;
- chromatic aberration phenomenon.
- the green color is always on the same side of the characters, and this excludes a JPEG compression artefact.

Overdigitization problems (4)

- as demonstrated by the previous example, certain libraries digitize with a resolution which is too large (that is, beyond their possibilities)
- the digitizations then introduce defects, for instance due to the lenses (chromatic aberrations, etc.)
- the material imposes a limit
- the digitizer must therefore have some physical notions, and this is seldom the case

overdigitization = false quality

Evaluating the quality of digitization

How can the quality of digitization be evaluated?

Idea: **ask the users**.

- there are different types of users, which mainly split in two groups:
 - those who browse, and
 - those who really use a book;
- however, the digital libraries are seen as extensions of the web and the way of consulting them is the same: it is the reign of superficiality;
- the users who browse therefore represent the majority and are in general satisfied; does that mean that we should rely on them for choices of quality?

Evaluating the quality of digitization

- the satisfaction of a user should be modulated by his/her needs;
- the needs can be known or not;
- some needs only appear during usage, in particular because these needs are seen as granted;
- for instance, a need for speed only becomes explicit when the service provided is slow; **it is not because a user does not express a certain need that he/she does not have it;**
- knowing the needs of the users and their evolutions is therefore not always simple.

Evaluating the quality of digitization

Asking the user is therefore dangerous:

- it is tempting to evaluate its needs as being those of the majority;
- the users which are critical are *de facto* a minority, as there is a majority of users who browse (how many are we to contact webmasters when the layout or the content of a page is not good?);
- the critical users are then marginalised: they are being told that their opinion is not representative.

Evaluating the quality of digitization

The quality of digitization can also be evaluated using the **reading comfort**. The (visual) reading comfort is mostly independent of the support: parchment, papyrus, clay tablets, paper, etc.

We always look with the same eyes!

In the case of digitization:

- the comfort is more or less close to the comfort of the original;
- if some characters have to be guessed, it is not sufficient;
- each character should be identified without ambiguity;
- the document must allow for extraction of all informations;
- the images and drawings must show all the details and be sufficient for a republication

Storage quality and visible quality (1)

- it may be necessary to distinguish the storage quality from the quality shown to the user;
- it is for instance possible to digitize at 600dpi, and to put online only 300dpi (or lower) images; this is possibly what Google does, but I doubt that the possibly higher-resolution images are exempt of the errors visible online;
- later on, when the accesses become faster, it might be possible to switch to a greater quality;

Storage quality and visible quality (2)

- there should be a certain transparency;
- the user should be informed of the storage quality, even if he/she only occasionally has access to it;

Storage quality and visible quality (3)

- When the quality of an online image is too low, we should not say “it is normal, the online version is downgraded.” Because then, one could ask “why was it put online?” So that the users pay for a better access?
- The real question is: why have those who put the image online not understood 1) that a better online version was needed, and 2) that it is possible?

The answer is that those who digitize and put online do not understand the needs of the users. Moreover, putting online a better version is not necessarily more constraining for the user, when the user is only online. The quality should only have a cost when downloading.

Storage quality and visible quality (4)

The various levels of quality naturally entail a structuration of the documents:

- it is possible to make an entire document accessible with a good quality, but with a high cost (volume, download time);
- (in addition) it is also possible to segment the document in order to allow for partial (predefined or on demand) downloads;
- finally, it is possible to prepare variable qualities, depending on the parts of the document; plates might for instance be digitized with a better quality than text.

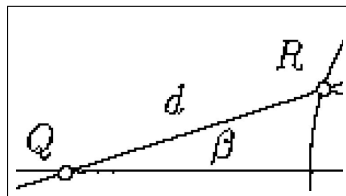
In any case, libraries can not be content with putting online downgraded versions without having appraised the needs of the users.

Storage quality and visible quality (5)

We can observe that some sites are increasing their image quality with time, sometimes using the same internal digitizations. An interesting example is JSTOR.

Here are excerpts of Arthur Baragar's article published in 2002 in the journal *American Mathematical Monthly*. The PDF was generated in 2007, at the time the file was downloaded (<http://www.vcharkarn.com/uploads/2/2557.pdf>).

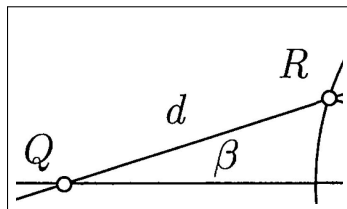
1. INTRODU
 maticians have
 provocative. Tc
 applications to



Storage quality and visible quality (6)

In 2011, JSTOR produces files with a better quality:

1. INTROD
maticians ha
provocative.
applications



Storage quality and visible quality (7)

The examination of the PDF file information reveals that they were produced differently, perhaps from the same source:

in 2007

Created: 02/05/2007 05:08:04 PM
Modified: 02/05/2007 05:08:04 PM
Application:
Advanced
PDF Producer: iText 1.3 by lowagie.com (based on itext-paulo-153)
PDF Version: 1.4 (Acrobat 5.x)

(produced when downloaded)

in 2011

Created: 03/06/2008 09:01:08 AM
Modified: 03/07/2008 11:54:20 AM
Application: PDFplus
Advanced
PDF Producer: Atypion Systems, Inc.
PDF Version: 1.4 (Acrobat 5.x)

(produced three years before)

Moreover, we observe that the two PDFs have similar sizes.

Required resolution

How can the required resolution be determined?

- the digitizer must understand the relationship between several entities, and above all between the reader and the book;
- the characters have a certain smoothness;
- the user reads with his eyes;
- a text printed at 300dpi displays irregularities which are visible with the naked eye at 10 or 20cm;
- in the same conditions, a text printed at 600dpi appears smooth;
- this is compatible with the angular resolution of the eye, which is about $1'$; at 50 cm, this corresponds to about 300dpi;
- this justifies a **sustainable resolution of 600dpi for text**;

Required resolution and information content

The required resolution mostly depends on the information content of the document:

- a text is aimed to be read at 50 cm, 600dpi should therefore suffice;
- if the text must be increased and if other details are important, a greater resolution may be necessary;

Required resolution and information content

- a book photography can usually be digitized at 600dpi, even if it is a halftone; however, in that case, a smaller resolution will not be a problem, because the user will not see the impact of the low digitization quality on an original which is already of low quality; the question at stake is the conservation of the quality of the original images, even if it is bad;
- certain photographs contain a lot more details than 600dpi can capture; certain old contact prints may require 1000dpi or more, and slides may require 3000dpi;
- the required resolution also depends on the use; a photography is not necessarily aimed at being seen in real size at a distance of 50 cm from the eyes;
- but for prints, 600dpi should almost always be sufficient.

Is there a need limit?

A former digitization manager recently reminded me of the evolution of storage formats:

- 1 B&W, and then color, silver processes
- 2 xeroxes
- 3 microfiches
- 4 B&W, and then color, microfilms
- 5 videodiscs
- 6 low-resolution TIFF
- 7 etc.

Many managers seem to view the evolution as an eternal replacement of technologies, with increasing lifespans and capacities.

Is there a need limit?

What the previous list suggests is that the needs increase.

- Libraries and digitization managers therefore became accustomed of thinking that non-sustainability was the norm.
- For some digitization managers, the concept of sustainability does consequently not have a very precise meaning.

But, as a doctor recently reminded me: the capacities of computers increase, but it is not the case, or very little so, with our brains. And our eyes also remain the same.

The eyes and the writing are fixing the limiting needs.

There is a limit of needs! Limit = equivalence

- (general case) if the purpose is to read a text:
When the digitization of a text reaches a resolution such that this text, when examined in normal reading conditions, does no longer appear rugged, then it is no longer necessary to go beyond for the reading comfort.
This limiting resolution only depends on the angular resolution of the eye.
- if the purpose is to extract another useful information:
It is necessary to define what is this useful information and to deduce the required resolution (or any other parameter).

These are the considerations which make it possible to define the minimum criteria of a sustainable digitization.

Do we need reading comfort?

We can do a parallel with screen technologies:

- we are evolving towards HD television, larger computer screens, thinner laptop screens, blu-ray movies, etc.
- these technologies will more and more marginalize low quality digitizations;
- today's and tomorrow's readers will more and more want a better image quality.

There is therefore no doubt that if comfort is not taken into account, sustainability will not follow.

Back to the storage and displayed qualities (1)

- we do now know that there is a limit resolution for common usages, that is for books read with the eyes (and not a lens); it is therefore not necessary to go beyond this resolution (for this usage);
- this does therefore define the sustainable storage quality; below this quality, the digitization will one day have to be redone;
- the quality provided to the user depends on the comfort that we can (or are willing) to provide;
- some contents render it necessary that the provided quality is closer to the storage quality than it is with other contents; this is in particular the case with the examples seen above.

Back to the storage and displayed qualities (2)

The user often does not know the storage quality, but if what the libraries provide is a *useless* quality (as in the examples seen above), one can ask

- if the libraries understand what they do and if they understand the needs of the users;
- if the internal storage (in case it is different from the displayed one) is really of sufficient quality.

In order to avoid these doubts, the libraries owe to the users a total transparency on the means and choices, both technical and political.

Color or black & white?

When a document is only printed in black and white, we can ask if a B&W digitization would not be sufficient.

- this is true, but the paper being seldom white, the appearance of the digitized page does not correspond to that of the actual document; there is therefore an impact on the feeling of identity;
- if only certain pages are digitized in color, this creates a consistency problem; these digitizations will not be viable in the long term;
- Google puts the color digitizations online since 2009 (but downloading remains B&W).

Most of the large German libraries are also digitizing in color, and downloading is also in color.

Quantity or quality?

As often is the case, quality is opposed to quantity:

- those who digitize little tend to digitize better;
- those who digitize a lot tend to digitize less well.

This raises the question: is it really more interesting for the user to have a lot of books digitized with an average quality, or less books digitized with a better quality?

I think that **one tries to make us believe that we have a need for quantity.**

Information explosion

- Within a few years, a large amount of documents have become accessible.
- In 10 years, we have gone from practically 0 book to 10 millions or more.
- But the needs are not yet there.
- And we do not have the time to exploit these resources!
- Who needs 10 millions books?
- And moreover, not sustainable?

We are been sold quantity (and bad quality), when we actually would be satisfied with a lot less.

Quantity is an economical need (1)

We have heard statements such as:

Google can digitize in 5 years what would have taken us 50 years

But is it **necessary** to go that fast?

We (almost) all have a cellphone:

but do we **really** need it?

Quantity is an economical need (2)

- to go fast (and bad) is an economic necessity
- it is not a user's need
- but if an actor goes fast (and bad) and another slowly (and well), the user does not perceive the qualitative advantage;
- Google therefore forces the other projects to go fast and to have a commercial attitude, although this is not justified from a scientific point of view.

Quality and sustainability

- Let us focus on quality and sustainability.
- And let us get down to the digitization criteria.

Minimum criteria for a sustainable digitization

Our recommendations are the following:

- at least 600 dpi
- color
- no photo manipulations
- TIFF is not necessary
- compression with invisible losses
- no visible aberrations (optical or other)
- plates unfolded
- originals aligned
- certain photographic supports (in particular older contact prints) require more than 600 dpi

The quality of interfaces

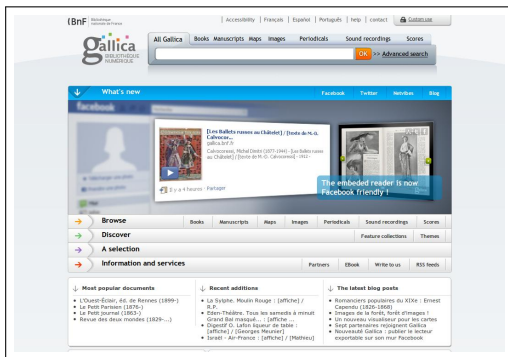
- we will now give a little overview of the interfaces;
- and take note of positive and negative features.

Analysis of several interfaces

The problem of quality

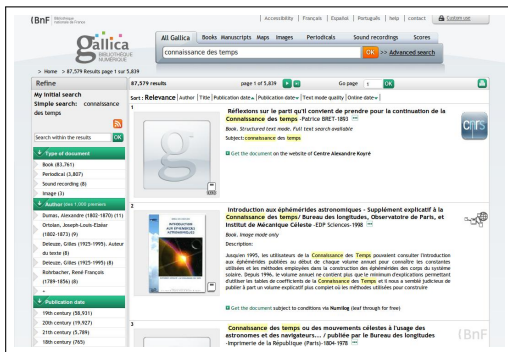
- short critical overview
- Gallica, Strasbourg, Berlin, Goettingen, Dresden, HAB, Munich, ETH, Swiss manuscripts, etc.
- specialized interfaces, for instance for a particular manuscript (Madame Bovary by Flaubert)
- there are a certain number of works on evaluation criteria, but they will not be mentioned here

Gallica: start page



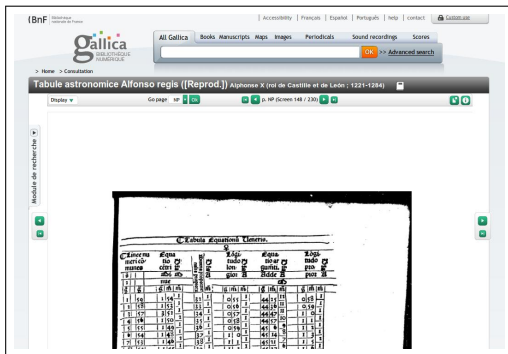
- too many things
- unnecessary informations for 99% of people
- the average user comes for a specific request
- 90% of this page is a distraction
- the distraction might be accessible from a link “Gallica’s buzz” and that would be enough

Gallica: search result



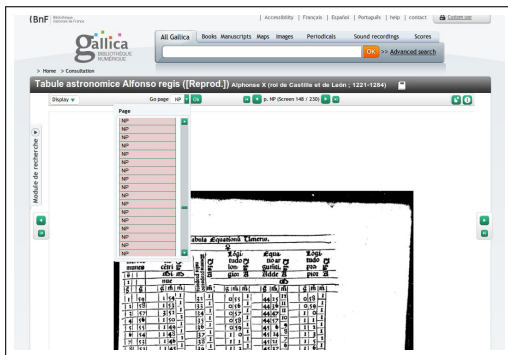
- on the left, informations with little usefulness
- too few thumbnails and neutral thumbnails
- too much text
- not very efficient
- search probably incomplete

Gallica: navigation in a document



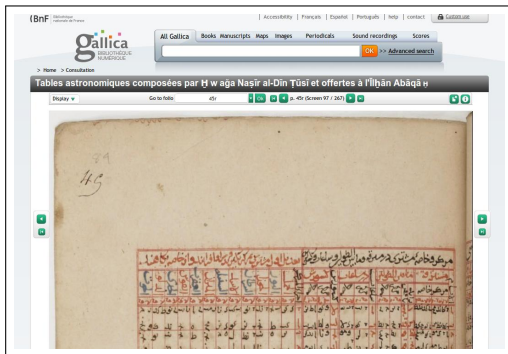
- by default, a page is seen only partially
- the left and right green arrows represent distractions, because they do not appear at constant locations (this is a programming error)
- microfilm digitization (hence B&W)
- some space could be saved at the top

Gallica: navigation (cont'ed)



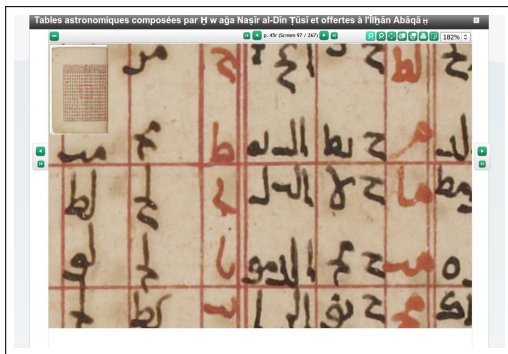
- the list of pages does not clearly indicate which one is current (they are all called NP = not paginated)
- thumbnails can be displayed, but indirectly using “display”

Gallica: navigation (cont'ed)



- case of a manuscript
- it is only possible to zoom indirectly by entering into the “display” menu

Gallica: navigation (cont'ed)

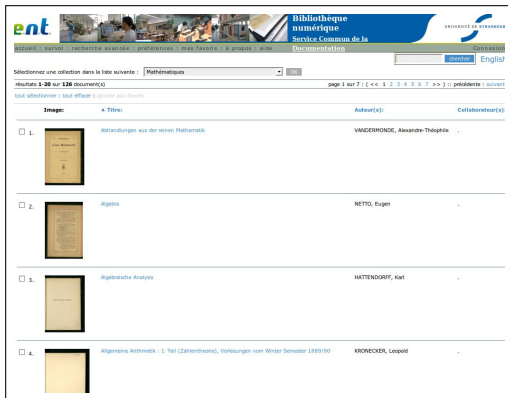


- zooming in the manuscript via an applet

Contentdm

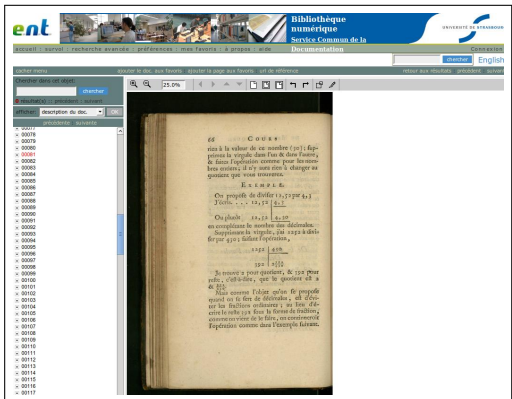
- platform developed by OCLC, an organization providing computer support to libraries
- used by Strasbourg, Linda Hall, Warwick, etc.
- <http://docnum.u-strasbg.fr>
- <http://lhldigital.lindahall.org>
- <http://contentdm.warwick.ac.uk/index2.php>

Contentdm: displaying a collection



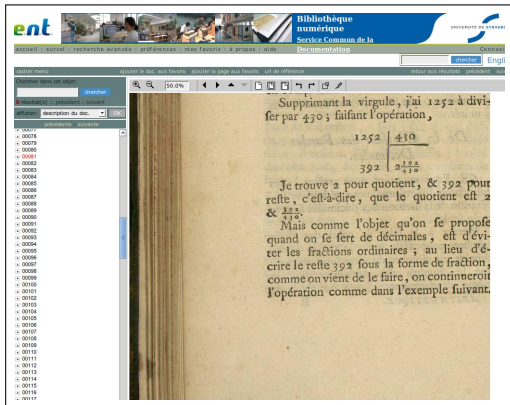
- displaying a subcollection
- collections should be replaced by a search on the subject
- a lot of empty space
- the authors are too far from the titles
- the page change at the upper right is often outside of the window which is not dynamic enough

Contentdm: view of a document



- view of a document
- it is possible to search within the OCR of a document, but the OCR has not been proofread
- the number of clicks to perform is too big (we often have to click on OK)
- the page can be downloaded in low resolution, but not easily in zoom mode;
- there is no downloading of the whole document in PDF

Contentdm: the zoom mode (1)



- here, the “next” button is only partly visible
- moving over the entire page is no longer possible by acting merely on the image! this problem is almost specific to Contentdm.
- it is possible to go around the zoom problem with the filled black arrows, but this is not intuitive

Contentdm: the zoom mode (2)



- it is not intuitive because we already have a partial (and natural) motion of the image obtained by pulling the mouse;
- the user needs to be able to act directly at the level of the text he/she sees, not elsewhere;
- moreover, moving using the arrows changes the position too much, and is therefore not smooth, and hence not comfortable;
- we have therefore here evidently a problem of interface design;

Contentdm: the zoom mode (3)

- Some of the advertised functionalities do not work at all, or are dependent on certain platforms.
- For instance, the button at the upper right of the interface is for “cutting the image in a new window.”
- But in that window, it is supposedly possible to select a part, then to select an operation with the right button of the mouse. However, this does not work everywhere!
- In general, the zoom can not be positioned with sufficient accuracy, even with the thumbnail visualizing the zooming frame (only certain positions are allowed).

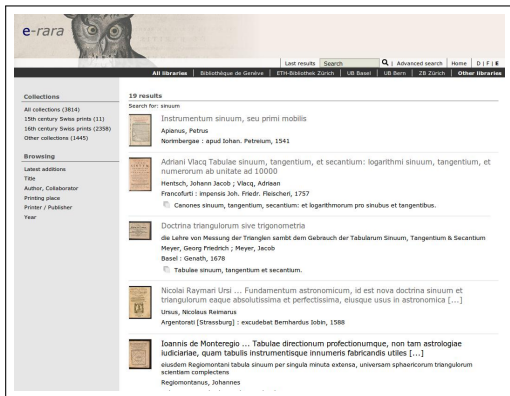
Contentdm

- it is easy to find yet other problems, given the bad design of this interface;
- the problems observed in Strasbourg are found also on the sites of Linda Hall and Warwick
- improvements are perhaps possible, but they are not made
- the (few) users have probably become accustomed to the disability which is imposed on them
- an interface such as Contentdm is deterring

Visual Library

- developed in Germany
- used in the Swiss project e-rara
- used in Düsseldorf

e-rara (Visual Library): search result



e-rara

Last results Search Q Advanced search Home D I F E

All libraries Bibliothèque de Genève ETH-Bibliothek Zürich UB Basel UB Bern ZB Zürich Other libraries

Collections

All collections (3814)

15th century Swiss prints (11)

16th century Swiss prints (2358)

Other collections (1445)

Browsing

Latest additions

Title

Author, Collaborator

Printing place

Printer / Publisher

Year

19 results

Search for: sinuum

 Instrumentum sinuum, seu primi mobilis
Aplanus, Petrus
Norimbegae : apud Johan. Petreium, 1541

 Adriani Vlacq Tabulae sinuum, tangentium, et secantium: logarithmi sinuum, tangentium, et numerorum ab unitate ad 10000
Hentsch, Johann Jacob ; Vlacq, Adrian
Francofurti : impensis Joh. Friedr. Fleischeri, 1757

 Canones sinuum, tangentium, secantium: et logarithmorum pro sinibus et tangentibus.

 Doctrina triangulorum sive trigonometria
die Lehre von Messung der Triangeln sambt dem Gebrauch der Tabularum Sinuum, Tangentium & Secantium
Meyer, Georg Friedrich ; Meyer, Jacob
Basel : Genath, 1678

 Tabulae sinuum, tangentium et secantium.

 Nicolai Raymari Ursi ... Fundamentum astronomicum, id est nova doctrina sinuum et triangulorum eaque absolutissima et perfectissima, eiusque usus in astronomica [...]
Ursus, Nicolaus Reimarus
Argentorati [Strassburg] : excudebat Bernhardus Jobin, 1588

 Ioannis de Montereigio ... Tabulae directionum projectionumque, non tam astrologiae iudicariae, quam tabulis instrumentisque innumeris fabricandis utiles [...]
eiusdem Regionum tabula sinuum per singula minuta extensa, universam sphaericorum triangulorum scientiam complectens
Regionianus, Johannes

- the interface is simple
- search page on an expression (here “sinuum”)
- five thumbnails and a short description (where, however, the **author is not always clearly given**)
- thumbnails are neither too big, nor too small
- clicking on the thumbnail or the title brings up the book

e-rara (Visual Library): main page of a document

e-rara **ETH-Bibliothek**
Höhere Technische Hochschule Zürich
Swiss Federal Institute of Technology Zürich

Last results Search Q Advanced search Home D I F E

Title view | Content | Thumbnails | Page view | **Libraries**

Bibliographic Metadata

Title	Nicolai Copernici Torinensis de revolutionibus orbium coelestium, Libri VI
Author, Contributor	Copernicus, Nicolaus ⁷⁶
Imprint	Nürnberg : apud Joh. Petresum, 1543
Description	[6]•196 Bl., [1] Bl. 18. (Holzschnitte im Text) 29 cm
Bibliographical reference	VD 16 K 2099
Language	Latin
Owner of original copy	ETH-Bibliothek Zürich, BA1.1370 ⁸
Persistent Identifier	10.3931/e-rara-420

Links

Entire Book as PDF	[^] Nicolai Copernici Torinensis de revolutionibus orbium coelestium Libri VI [109.07 mb]
--------------------	---

Collection

Other collections > Astronomy (ETH-Bibliothek)

Content

[Table of Contents](#)

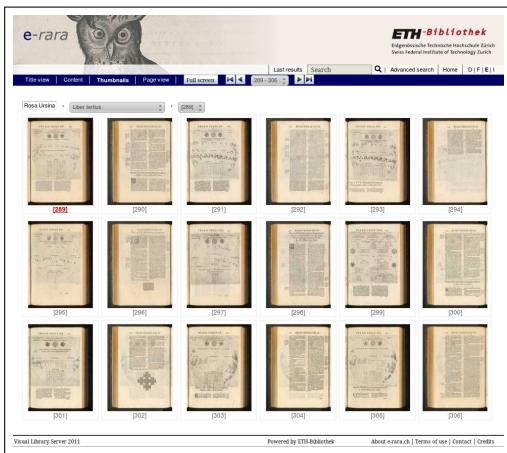
- access to metadata, including the location of the original
- access to the “Page” or “Thumbnails” views
- immediate download (but downgraded)
- other possibilities like going from chapter to chapter are also possible

e-rara (Visual Library): page view



- it is possible to advance in the document, to choose any page, or to zoom (without being restricted in an area) using the buttons at the top
- buttons are too small or too close to each other
- two zooming modes in early 2011 (here), but reduced to one in 2012 (making it now difficult to download a good image)
- rotations are possible
- hierarchical navigation

e-rara (Visual Library): thumbnails



- the thumbnail mode displays a number of pages from the document;
- it is possible to go from one set to the next;
- it is also possible to reach a given thumbnail interval through a menu;
- but a fast traversal (Google-style) is not possible

e-rara (Visual Library): quality of text in PDF

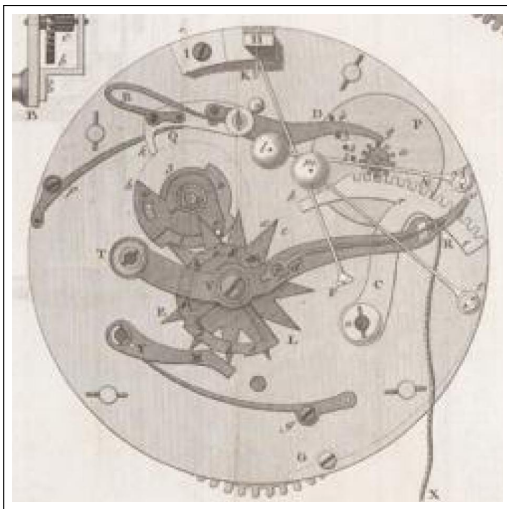
16 HISTOIRE DE LA MESURE DU TEMPS.

une description^a de ces différens instrumens. Nous allons en rapporter quelques détails, en distinguant ce qui semble dû à cette École, de ce qui appartient à des temps antérieurs. »

Les premières horloges d'eau ont été simples et même grossières. On aura d'abord voulu mesurer le temps par l'eau écoulée d'un vase : mais on n'aura pas tardé à s'apercevoir que les quantités d'eau n'étoient pas proportionnelles au temps ; et, après avoir reconnu que l'erreur naissoit de la chute inégale de l'eau, on aura ensuite cherché à y remédier en employant, au contraire, le temps de l'immersion des corps dans l'eau. Le petit bateau des Indiens, percé d'un trou, qui surnage d'abord et s'enfonce au bout d'un certain temps fixé par l'expérience, a peut-être été, dans ce genre, le premier degré de perfection des clepsydras.

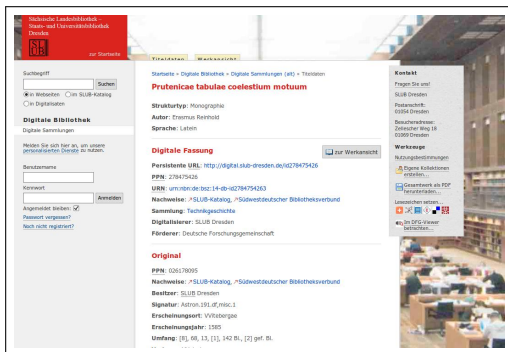
- Berthoud: *Histoire de la mesure du temps par les horloges* (1802)
- the PDF format for text is acceptable

e-rara (Visual Library): quality of figures in PDF



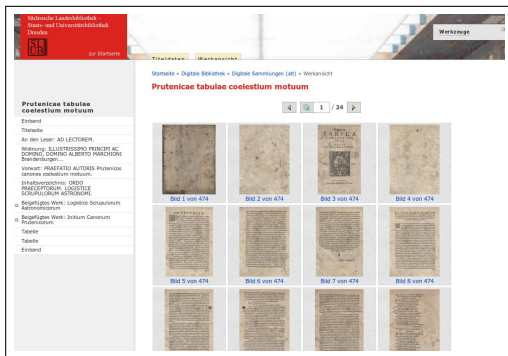
- Berthoud: *Histoire de la mesure du temps par les horloges* (1802)
- the PDF format for the figure is extremely bad
- moreover, the images can no longer be downloaded directly with a good quality
- the internal format is very detailed, but is not downloadable, except in mosaic

Dresden: main page of a document



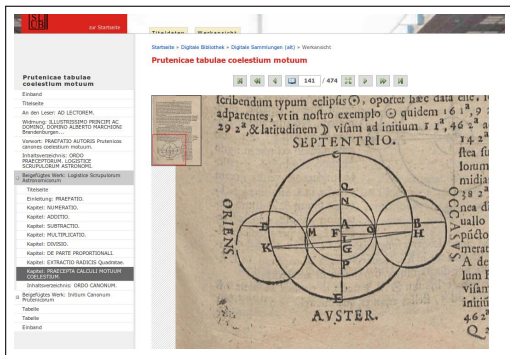
- no document thumbnail
- many metadata
- only some of these metadata are essential

Dresden: view of the document



- by default, the document is displayed in “thumbnail” view
- it is possible to go from one group of thumbnails to the next one
- downloading a PDF is possible

Dresden: zooming on a page



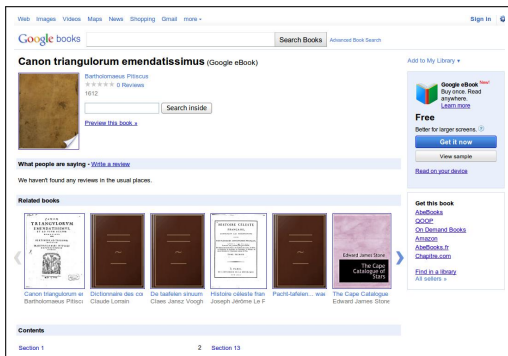
- detail of a page
- it is possible to navigate in a natural way, and, moreover,
- the scrollwheel can be used for zooming, but since it is one of the only sites which provides this feature, it causes problems when switching to another site

[http://www.slub-dresden.de/index.php?id=5363&tx_dlf\[id\]=18717](http://www.slub-dresden.de/index.php?id=5363&tx_dlf[id]=18717)

[illegible]

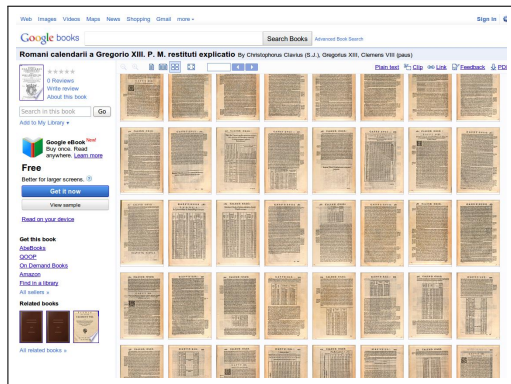
- new version of the interface
- in this version, it was no longer possible to download a PDF in March 2011, but it does apparently work again

Google books



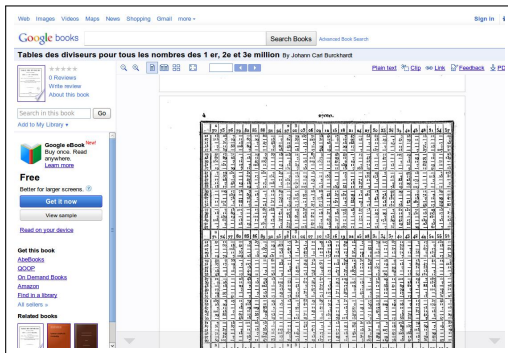
- the first page of the book is not visible
- the fact that the book can be downloaded is not clear right away
- one even has the feeling that the book can only be bought

Google books: thumbnails



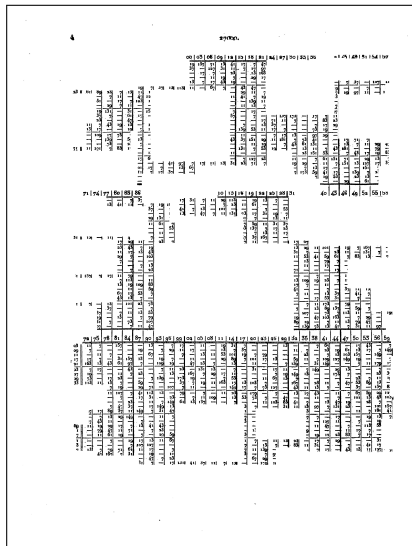
- thumbnails
- selected by a button
- moving through pages using the scrollbar, which is convenient and not available with the other interfaces

Google books: detail of a page



- the page appears to be normal
- the interface is rather simple
- it favors the navigation with the scrollbar or the scrollwheel
- part of the left window is of a commercial nature
- the year of the book is not given here

Google books: internal representation of a page



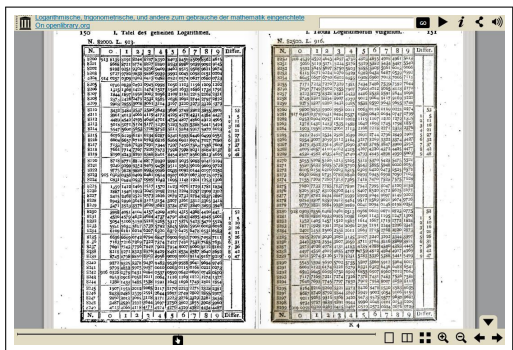
- actually, the page contains holes
- in the PDF, it is stored in the form of layers (JPEG2000 format)
- the division in layers is often clearly visible in the original
- this reduces the quality of the image
- it is safe to think that Google's display interface is not perfect

Archive.org

The screenshot shows the Internet Archive website interface. At the top, there's a navigation bar with links like 'Web', 'Moving Images', 'Texts', 'Audio', 'Software', 'Patron Info', 'About IA', and 'Projects'. Below this is a search bar and a list of collections. The 'American Libraries' section is highlighted. A list of books is shown, with the selected book being 'Logarithmische, trigonometrische, und andere zum gebrauch der mathematik eingerichtete tafeln und formeln (1783)' by Vega, Georg. The page provides details about the book, including its author, subject, and a list of available formats (PDF, HTML, etc.).

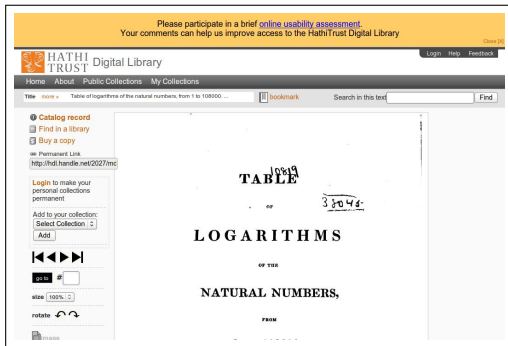
- part of archive.org's books originate from Google books, but the PDF format is not the same
- archive also allows for a 'virtual page' browsing
- the quality is similar to that of Google
- link towards the "Open Library" project
- sometimes, a book can be downloaded on Archives, but not on Google

Archive.org: detail of a page



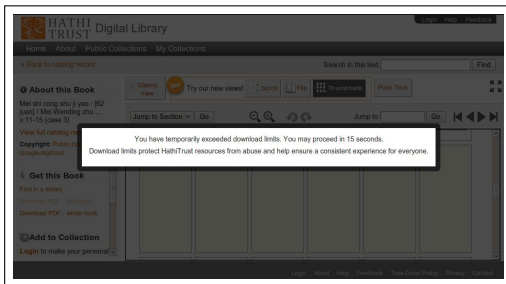
- quick navigation is possible using the bar at the bottom
- colored artefacts
- some parts seem sharper than other, because of the layer structure of the pages

Hathi Trust (1)



- another newcomer
- some books indirectly originate from Google (actually, deposits of Google digitizations)
- restrictions may be different

Hathi Trust (2)



Unless one is a member of a privileged institution, some documents (here from 1761) have such access restrictions that we were not even able to navigate the thumbnails without regularly reaching a screen whose purpose was to slow down the search.

Hathi Trust (3)

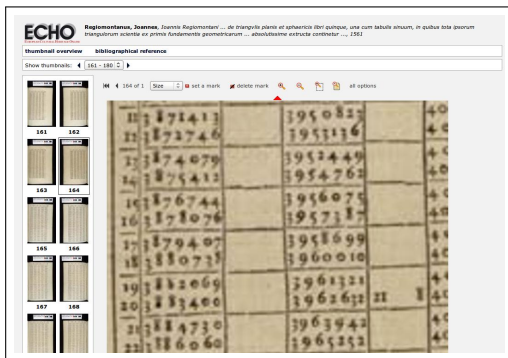
- Hathi's interface is lacking conviviality so much that it is deterring;
- consulting documents for research is very difficult, because of the restrictions;
- the interface apparently keeps a register of page accesses, but weighs thumbnails and full pages identically;
- however, in order to quickly locate a part (for instance an image), thumbnails are essential, and these thumbnails then use all the credit;
- the normal view does not allow a quick navigation;
- Hathi seems to be an example of possibly well furnished site, but probably almost unused, because unusable.

Berlin



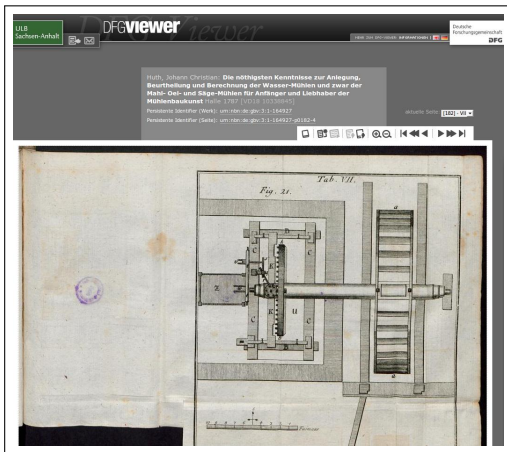
- downloading seems impossible
- the document originates from HAB (Wolfenbüttel library)

Berlin: zooming



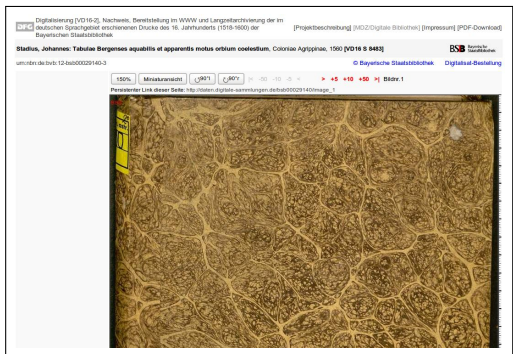
- the resolution is highly insufficient
- the figures should be sharper
- is that the famous Wolffenbüttel library which wasn't able to digitize adequately?
- the zooming interface provides a **navigation mode which is very inconvenient**, worse than the one given by Contentdm

ULB Sachsen-Anhalt



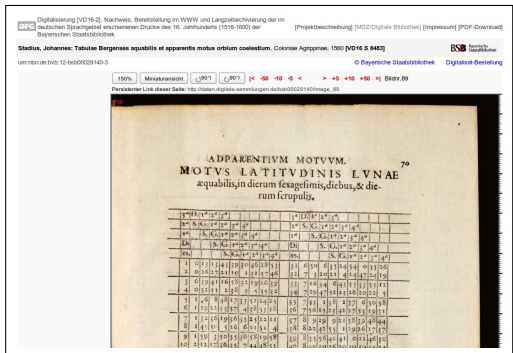
- example where the plates have been digitized (folded *and* unfolded)
- but not well aligned ...

Munich: main page of a document



- view of the cover first (not very useful)
- downloading is possible

Munich: detail of the content of a document



Munich: interface for Google digitizations

Munich also puts Google digitizations online, but with a different interface:

BSB Bayerische Staatsbibliothek digital

Katalog (OPAC) | Impressum | English

Dittmann, C.: Coordinaten- und Tangenten-Tafeln (1859) [BibTeX](#)

Seite 11 von 11

◀ -3 -10 -50 +50 +10 ▶▶

10050417

73

Grad.

5	6	7	8	9	Minuten
Abcisse	Ordinate	Abcisse	Ordinate	Abcisse	Ordinate
409 5700	344 1458	491 4912	401 5034	373 4064	458 3611
4926	2888	3911	6702	2896	459 0517
4091	4317	2909	8370	1727	2423
3266	5746	1907	402 0037	0538	4328
2420	7175	0904	1704	572 9288	0235
1584	8603	490 9901	3370	8218	8137
0748	345 0031	8898	5036	7047	460 0041
408 9912	1459	7894	6692	5816	1945
9075	2586	6590	8667	4705	3848
8238	4314	5885	403 0033	3533	5752
7400	5740	4880	1697	2360	7654
6502	7167	3875	3862	1187	9556
5724	8508	2969	5026	0014	461 1438
4886	346 0020	1803	6630	571 8840	3360
4047	1445	0856	8553	7655	5260
3:07	2871	489 9849	404 0016	6490	7163
2268	4296	8841	1678	5315	5081
1528	5721	7853	3341	4139	0981
0683	7145	6825	5003	2963	2860
407 9847	8570	5817	6665	1786	4760
9006	9993	4507	8326	0609	6658
8165	317 1417	3798	9287	570 9431	8556
7323	3840	2788	103 1647	5252	463 0454
6481	4764	1777	3308	7074	2352
5639	5687	0766	4968	5894	4249
4796	7109	488 9755	6628	4715	6146
2953	8531	8745	8287	3554	8042
3109	9953	7281	9946	2358	9938
					6975
					9939
					1597
					33

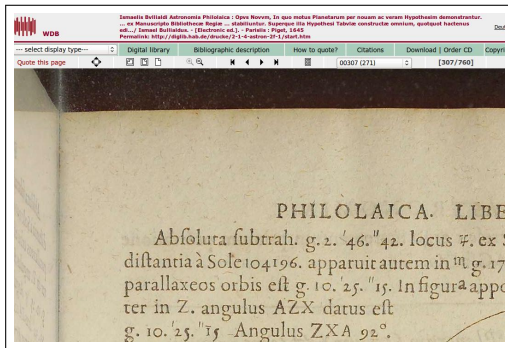
- the fast forward buttons are kept, but the interface remains rather limited

Goettingen



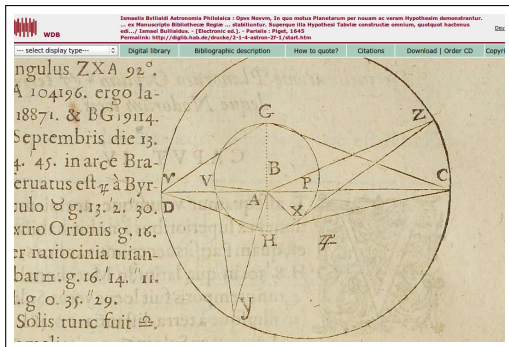
- here, as on other interfaces, it would be useful to have fast forward buttons

HAB (Wolfenbüttel)



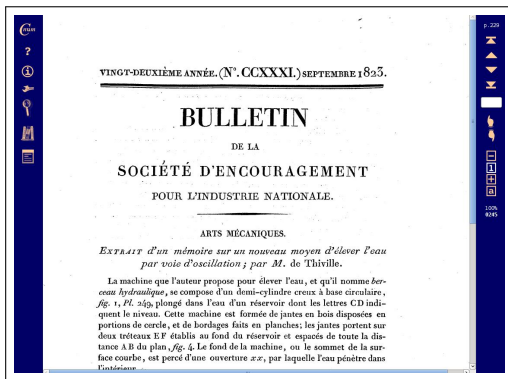
- zoom view from the top of the page
- here, the resolution is excellent
- in the document seen before, there is in fact only the low resolution version

HAB (Wolfenbüttel)



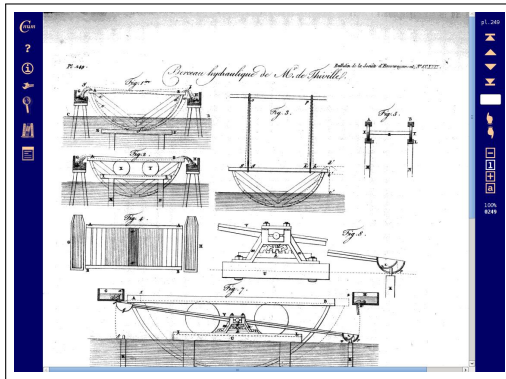
- zoom view
- when scrolling down, we do no longer have access to the navigation buttons
- this problem is common
- Gallica tried to solved it, but the solution employed is not perfect

CNUM (CNAM)



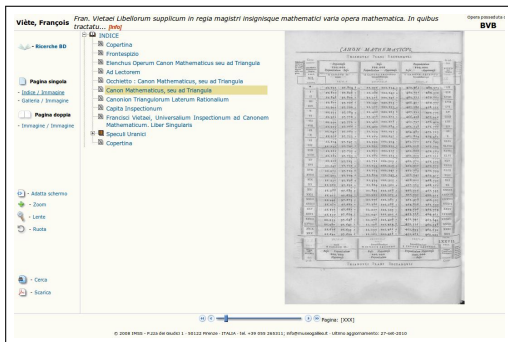
- pages of a document digitized by the CNUM
<http://cnum.cnam.fr>
- very limited interface
- the “hands” allow for a separate navigation through plates
- no thumbnails page, nor a quick scrollbar navigation
- a complete PDF can be downloaded on the main page of the document

CNUM (CNAM)



- view of a plate
- it is not aligned and its resolution is not sufficient

IMSS (Florence)



- downloading is possible, but only in low resolution (not sufficient)
- B&W pages (also tattooed)
- zooming opens a separate window
- the zooming and other buttons are at the left, too far from the image
- navigation through the book using the bar at the bottom
- too big scattering of accesses

Sevilla

UNIVERSIDAD DE SEVILLA

Fondo Antiguo

Home > Fondo home > A 627/04321

General search

Advanced search >

Book Views

- Detail View
- Wide View
- Read
- Book Index

Index of the book

- Portals
- Texts

1. Books

- 1.1. Book Groups
- 1.2. Latest Books
- 1.3. Jewels of the Library
- 1.4. Request Digital Resource

2. Illustrations

- 2.1. Illustration groups
- 2.2. Latest Illustrations

3. Historical newspapers

Last News II

Subscription at the newsletter I/G

Links 4

Contact us

Monstrar páginas

Canon trigonometricus : continens logarithmos, sinuum, et tangentium, ad singula scrupula, totius semicirculi

1 de 48

Go to

Download book

Full Record

Download METS

Download MARC XML

Authors

- Zarate, José (S.I.). 1627-1678

Others

- Villadiego, Bernardo de editor

Data from the play

- Publication date: 1672
- Publication place: Madrid
- Signature: A 627/04321

Book's groups

- Filosofía natural y otras de carácter científico
- Libros del siglo XVII

CANON TRIGONOMETRICVS.

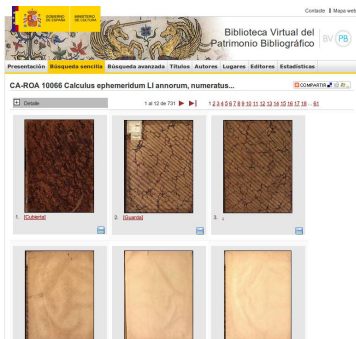
CONTINENS LOGARITHMOS, SINVM, ET TANGENTVM, ad fingula scrupula notos Smaicrulli.

Matriti.

Apud Bernardum à Villa-Diego.

- B&W digitizations

Biblioteca Virtual (Espagne)



- the entry page of a document gives the thumbnails

Biblioteca Virtual (Espagne) : détail d'un document



- detail of a 1559 eclipse
- the scrollwheel can be used for zooming

Heidelberg

The screenshot shows the website of the Universitätsbibliothek Heidelberg. The header includes the university name and navigation links like 'Sitemap', 'Kontakt', 'Layout anpassen', and 'English'. A left sidebar contains various navigation options such as 'Schnellsuche', 'HEIDI', 'Web-Seiten', 'Literatursuche und -bestellung', 'Elektronische Medien', 'Nutzung und Service', 'Fachbezogene Informationen', 'Schulungen', 'Bibliotheken der Universität', 'A bis Z', 'Aktuelles', 'UB für Einzeliger', 'Konto', and 'Studiengebühren'. The main content area displays details for 'Heid. Ms. 3394' by Heinrich Rüdiger, titled 'Planetenbuch'. It includes the date 'Heidelberg (Weinheim), 1551-1564', a 'Wissenschaftliche Beschreibung Sammlung', a 'Persistente URL', and a 'URN'. There are buttons for 'Bestellung' and 'Download (PDF, 35 MB)'. A thumbnail image of a manuscript page is shown. Below the main content, there is a 'Sprung zur Seite' field and an 'Inhalt' section listing the manuscript's structure: Einband, 1er Titelblatt, Widmung, 1r Wappen und Devise Ottheinrich, 1v - 56v Deutscher astronomischer Kalender, 57r - 75v Johann Mercurius: Geburtsprognostik, 76r - 93v Petrus Apian: Usus Almanach, 94r - 131v Lucas Gauricus: Astrologischer Traktat, and Einband hinten.

- here, we have a manuscript
- the first page gives the hierarchical structure

Heidelberg: navigation detail

Heidelberger historische Bestände - digital

UNIVERSITÄTSBIBLIOTHEK HEIDELBERG

Heid. Hs. 3394
Heinrich Büdinger
Planetenbuch
Heidelberg (Weinheim), 1551-1564
Seite: 88v

Wissenschaftliche Beschreibung
Download (PDF, 35 MB)
Startseite des Bandes
Sammlung

Spring zur Seite (z. B.: 12v, 20r)

Inhaltsverz.

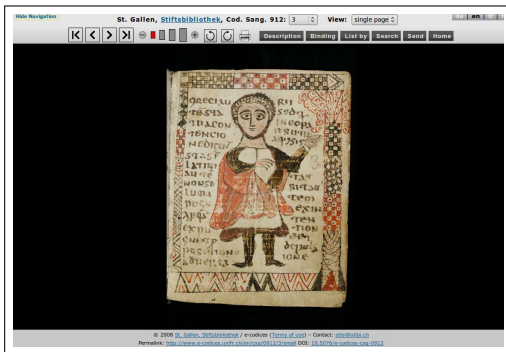
- Einband
- 1r Titelblatt, Widmung
- 1r Wapen und Devise Ottheinrich
- 1v - 56v Deutscher astronomischer Kalender
- 57r - 75v Johann Mercurius: Geburtsprognostik
- 76r - 93v Petrus Apianus: **Uusus Almanach**
- 94r - 131v Lucas Gauricus: Astronomischer Traktat
- Einband hinten

Persistent URL:
<http://digi.ub.uni-heidelberg.de/diglit/heidhs3394/0181>

Seitenansicht Vorschau

Swiss manuscripts: www.e-codices.unifr.ch

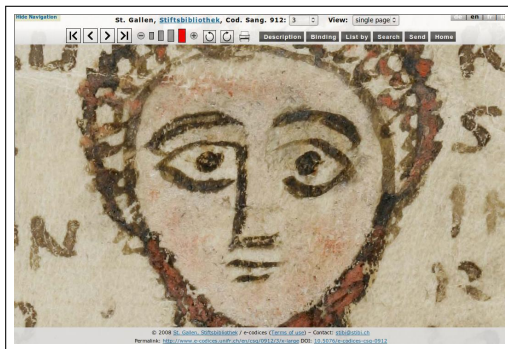
A palimpsest manuscript from the 5th century



- it is apparently only possible to obtain one-page PDFs
- the images can also be downloaded separately

Swiss manuscript: www.e-codices.unifr.ch

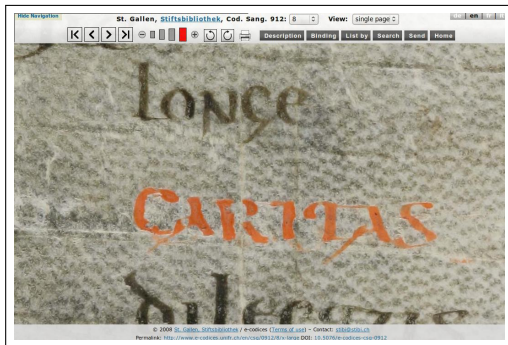
A palimpsest manuscript from the 5th century



- maximal zoom (in red)

Swiss manuscript: www.e-codices.unifr.ch

A palimpsest manuscript from the 5th century



- detail of a word with the same zoom

- specialized interface
- zooming buttons in the upper right corner, too far from the image to which they apply

Quality of the interfaces: summary

- sometimes distressing
- either the users are not taken into account, or the users are not very demanding

A good interface should:

- provide a number of basic functionalities;
- be ergonomic and as much as possible be invisible;

A good interface should be natural and discrete, like a good font.
It is not sufficient for an interface to provide all functionalities.

The matter of downloading

- full download is useful, because not every person can (or wants to) read a book online, or on another screen
- the downloadable quality may differ from the real quality
- the downloadable quality should be sufficient for a comfortable reading, and this is not always the case (**what is the purpose of an insufficient quality?**)
- it is possible that some libraries intentionally chose an interface making downloading difficult, but the lack of transparency of many libraries makes it impossible to be sure.

The matter of downloading (2)

Downloading induces a **fear**:

- libraries are attached to their books;
- they often would like not to *give* the digital versions entirely;
- they would/will like to introduce a traceability: who downloads what?

These fears correspond to an **outdated perception of how libraries work**.

The library's activities will little by little have to be redefined:

- conservation of originals
- creation of digital versions
- enrichment of digital versions

The original books remain the natural property of the libraries. The digital versions should be distributed as widely as possible.

Minimum criteria for a good interface (1)

- simple pages, without useless information
- title, author, year on every page
- common navigation functionalities
- navigation through groups of thumbnails
- navigation on full pages, by steps of 5, 10, 50, 100 (very useful, but seldom available)
- functionalities should be grouped (no scattering)
- less important functionalities → separate page
- avoid white zones (waste)
- internationalized interface, with at least the English option
- quick access

Minimum criteria for a good interface (2)

- zooming in should not restrict the navigation area, nor require special operations (as in Contentdm)
- moving (without restrictions) within a zoom window should be done by pulling the image
- the number of clicks should be minimized
- buttons should be near the images
- thumbnails should be neither too big, nor too small
- it must be easy to navigate with a constant zoom factor
- buttons should not be too close to each other
- image (page) download and PDF download
- resolution downgrading is acceptable in the PDF, but it should be possible to bypass the problem

What are the causes of all these problems?

The job of librarians has changed (1)

- until recently, libraries created little: the books were not made by the libraries;
- a library was an intermediate between the publisher and the reader;
- creation was limited to catalogues (cards, then computerized);
- the book borrowed by a reader was the same as the one borrowed elsewhere.

The job of librarians has changed (2)

- today, libraries create more: they digitize;
- they do it essentially for three reasons:
 - they own the books;
 - the digitization material is relatively simple and rather unexpensive (camera and computer)
 - it is easy to make an acceptable digitization
- if the digitizations were more difficult, they would be much more centralized

The job of librarians has changed (3)

Digitization naturally ended up in the hands of libraries:

- the libraries are responsible for the choices
- the user is almost absent from this chain:
 - mainly because the digitizations do not correspond to a pre-existing need;
 - the users are not demanding.
- the libraries had and still have full powers.
- but ... this leads to problems.

Part of these problems comes from a **breakdown of roles**.

The breakdown of roles in digitization

- those who digitize are not those who use;
- those who digitize do not listen to the users;
- those who digitize are no technicians;
- those who decide do not digitize;
- those who program do not digitize.

Those who digitize are not those who use

Foreseeable and visible consequences:

- no knowledge of the real needs
- the quality of digitization is not taken into account
- no real understanding of the aims of quality and comfort

Those who digitize do not listen to the users

- Traditionally, the libraries have only consulted the users little, or not at all, for their choices.
- This habit has been kept.
- But it is no longer adapted to the digital world.

Moreover, listening to the users is biased:

- for some libraries, the absence of feedback legitimates the choices
- for some libraries, a little bit of feedback also legitimates the choices

Those who digitize are no technicians

- no training on the limit resolution capacity
- no long term vision
- no desire to get rid of the chaos
- the digitizers have no training in physics, and are not computers
- the digitizers use digitization devices, and then use photo manipulation tools and image databases

Those who decide do not digitize

- and they do also not use the digitized books
- consequently, they do not see some of the errors, or do not realize that they are errors

Those who program do not digitize

- they do not read what has been digitized
- and they are too low in the chain to transmit possible observations
- when errors are noticed, it is too late, the book has been shelved, and it is too costly to go back to it.

Interactions between the libraries and Google

The libraries have been abused by Google:

- is it possible that so many prestigious libraries have accepted that Google makes use of such a low level of quality?
- the library heads are actually unaware of these problems!
- a serious patrimony curator should actually refuse that digitizations are that much botched
- Google knows about the errors (and could correct them), but still manages to get accepted
- Google even sells books which are incomplete
- for many libraries, an average quality is compensated by a large quantity at an almost zero cost
- the Lyon municipal library started to be digitized in 2009, and we can anticipate the same catastrophe

Hindsight problems

The librarians:

- for them, the users are satisfied (in fact, the users do not use)
- each technology is replaced by the next one (no notion of sustainability)
- means are limited
- the subcontractors have said that ...
- they trust historians which are not technicians
- they are ignorant of Google's details (since librarians and users are different) and they are even not interested to know

The wrong priorities of the libraries

- some libraries swear too much by metadata
- the first priority should be the faithfulness of the content, the title, the author, the year
- OCR can be applied, but in that case it is advisable to correct it (and OCR can be externalized)
- OCR should not be a precondition for putting a document online (and this, some libraries do)
- one should not digitize for the sole purpose of showing that one has a digitization activity
- by focussing on metadata, the libraries miss the problems of quality of digitization

The wrong priorities of the libraries (2)

The feeling that one gets is that libraries

- view digitization as subcontracting
- consider that their work is about the “coating” (metadata, OCR, interface, etc.)

This way of viewing things creates a distancing between the libraries and the books.

Bad digitization and duplicates

- a same book may have been digitized several times;
- Alain Jacquesson wrote (p. 75 of his book): “We can only regret that Stendhal’s work, *The Red and the Black*, has been digitized, in a same edition, both by the University of California and by the library of the University of the State of Bavaria in Munich.”
- but, should this really be regretted?
- in fact, **Jacquesson’s question is ill-posed**, for it does not take the quality of digitization into account;

Duplicates are only an evil at equivalent or inferior quality. One first should compare the qualities. Not comparing them cripples future digitizations.

But ... duplicates are sometimes useful

- In certain cases, copies which at first sight seem identical are not.
- An heterogeneous repository (for instance the archives of a writer) may contain documents which have already been digitized elsewhere, but that one will still wish to include in the digitized repository, for reasons of coherency.

Duplicates should not be fortuitous.

The heavy legacy of bad digitization

- the bad digitizations block the production of better ones
- in Strasbourg, for instance, a digitization which has been made elsewhere reduces the priority of a new digitization of the same book
- this penalty is all the more important that the digitization managers do not understand the quality requirements

The problems of digitization policies

Aims of digitization

At the beginning, we have assumed that **the aim was the quality**.
This is still our aim.

We have examined the following points:

- what should the quality of the digitizations be?
- how should the digitizations be displayed?
- what are the causes of the encountered problems?

We now have to examine the choice of contents:

what should be digitized?

How should a repository be selected for digitization?

Usually, digitizations are done repository by repository. Various criteria can be considered:

- unique collection, for instance manuscripts;
- very demanded collection (essentially manuscripts);
- etc.

In any case, it is necessary

- to have a knowledge of the collection before choosing;
- to have a long-term policy, and not a local one;
- to implicate the users.

Specifications

The specification prepared by an organization which plans to digitize is beyond the scope of this presentation, but it has to give:

- the conditions by which the collection's integrity will be ensured (for instance the order and layout of the documents)
- the technical conditions of the digitization
- the perennial storage conditions
- the restrictions to user access

It is also advisable that the specifications are not confidential and are freely available to the users, so that they can comment upon them.

Strategies for a small digitization center

A small center does not have the means that the great actors have.
Its choices are consequently affected.

A strategy is then to

- digitize correctly what others will not do correctly;
- digitize correctly what others do not possess, for instance some manuscripts

Are there things that should not be digitized?

This question is often raised, but:

- it is an ill-posed question;
- everything that has been kept in form of paper, parchment, papyrus, etc., manuscript or printed, is entitled to be digitized;
- the amount of documents which do not natively exist in digital form will certainly decrease in libraries and archives, and this will eventually limit the amount to digitize;
- the problem of non-digitization must be brought back to that of non-conservation; the digitizer should not ask him/herself if a certain document should or should not be digitized; if it has been kept, it should be digitized.

The problem of the choices

The choice of collections often rests on misunderstandings:

- historically important $\neq$ urgent to digitize;
- there is almost no digitization emergency;
- digitized books are not read.

The example of Strasbourg University

- the library of the University of Strasbourg has been digitizing books since approximately 2005
- the scientific collection is very rich
- the books are chosen by persons of different domains, often retired professors
- nevertheless, the selections seem to be made without a real knowledge of the collection;
- in 2010, in order to make up for this problem, I have gathered a 250 pages overview of the collection of scientific books of the University;
- this overview was initially meant to help better select the books that had to be digitized, but it was also meant to avoid falling in an abusive subcontracting
- however, this work highlighted the limits of the thematic approach followed at Strasbourg;

The limits of a thematic approach

By digitizing only thematically and with limited means (100 to 200 books a year):

- a rich collection can only be touched upon;
- the sampling is subjective and the choices do not necessarily correspond to needs;
- the thematic approach (quotas for each field) is satisfying for those who are distantly interested (one pleases every one), but the truth is that there is a standstill everywhere;
- by trying to favor the themes, the nature and the needs of the documents are neglected;
- a thematic approach on books which do exist elsewhere is often a waste of public money.

Subcontracting digitization

- some libraries have a digitization activity, but, in addition, they subcontract to others; this is the case of the French National Library which subcontracts some digitizations to the Strasbourg national library
- subcontracting is acceptable if it does not harm the collection of the subcontracting library
- in 2009, it was attempted to use the Strasbourg University libraries as a subcontractor for an institution in Paris; this subcontracting would have ignored the local collections, and was therefore not acceptable;
- such drifts are a consequence of the absence of a strong centralization of digitizations.

Window-digitizations

Digitizations have now become fashionable:

- limited budget $\implies$ limited action
- a specific collection is digitized, for testing purposes
- is the window-digitization sustainable?
- is the purpose of the window-digitization to set a global (national) action in motion?
- one should stop digitizing for the sole purpose of showing that one knows how to digitize, or for the sole purpose of copying others
- digitization should be done with hindsight

Digitization as affirmative action

It is a valorization strategy:

- the purpose is here to highlight a collection, by favoring it for digitization;
- this is therefore what can be called an **affirmative action**.

Such a digitization encourages research in a given collection.

Sustainability in digitization

- digitization must produce a permanent object
- the content may have a much longer lifespan than the interface
- examples of non sustainable digitizations:
 - all the digitizations which have not been made from the original documents
 - some of Gallica's digitizations originating from microfilms
 - digitizations of vital records in Archives (in some places, made from microfilms)
- when the object is a book, we have the option to wait that another library performs a better digitization
- but for manuscripts, it is up to the library which owns them
- sustainability will be reached when the digitization quality makes it possible to reproduce the quality of the original document
- insufficient digitizations waste money

Sustainability in digitization

With a sustainable digitization:

- the digitization is comfortable
- the use of the original document is no longer necessary

On-demand digitizations (1)

- in 2009 or 2010, I suggested to a German library to digitize an old book which had not yet been digitized
- the library answered that it was possible, and told me at which cost
- question: who decides about digitizations if the researchers are not listened to?
- the good solution would have been
 - either to submit this suggestion to those who are responsible for the selections, in order to see if the suggestion is consistent with the digitization policy and to integrate it; if not, then the user may be charged
 - or to plan a slot for on-demand digitizations, without charge
- eventually, I gave up, and decided to wait that another library digitizes the book

On-demand digitizations (2)

- in 2009 or 2010, the Strasbourg National Library digitized for me (and for it!) several books which were charged to me (and which are now available online at other places. . .);
- this was then natural, because the library does not have a digitization policy as developed as the German libraries do

The example of mathematics

- there is a community interested in the digitization of mathematics
- it is estimated that the total number of pages of the mathematical corpus is about 100 millions
- “Digital Mathematics Library” (DML) project
- the last conference on this theme took place in July 2011.

Digital Mathematics Library (DML)

Themes of the DML 2011 conference:

- o search, indexing and retrieval of mathematical documents
- o ranking of mathematical papers, similarity of mathematical documents
- o math OCR with MathML/TeX output
- o document conversions from/to MathML, OpenMath, ..., PDF
- o conversions between various mathematical formalisms
- o mathematical document compression, processing of scanned images
- o algorithms for crosslinking of bibliographical items, intext citations search
- o mathematical document classification, MSC 2010
- o mathematical text mining
- o mathematical documents metadata exchange
- o long term archiving, data migration
- o reports and experience from math digitization projects
- o math publishing with long term archival goal
- o software engineering aspects of creating, handling MathML, OMDoc, OpenMath documents, and displaying them in web browsers

The example of mathematics

The previous list shows that

- those who are interested in the digitization of mathematics are mainly the mathematicians
- the aim is to make accessible useful and not yet digitized mathematics
- paradoxically, the conference themes refer only very little to old documents
- in fact, this conference does not reflect very well the interests of the community of historians of mathematics
- consequently, we should fear that the digitization of ancient mathematics will not be that well defended, because these books are practically useless nowadays
- such problems occur in other domains, in particular in astronomy

Questions for a digitization

Several questions are raised in case of digitization:

- who selects the collection to digitize?
- why this collection? is there a demand?
- is this digitization sustainable?
- is it a window-digitization?
- who will digitize? do these persons understand the limits of digitization?
- will the digitization be adapted to the needs of the contents (formulæ, images, figures, etc.)?
- who chooses the interface? how does this interface compare to the other interfaces?
- will there be downloading? if not, why not?
- will the text be given? if yes, will it be indexed on the web?
- will it be possible to put the digitized collection back into a larger collection in an homogeneous way?

Solutions

We have already seen specific solutions for two problems:

- problems of the quality of digitization;
- problems of the quality of interfaces.

There now mainly remains to solve the political problems.

The chaos of digitization

One of the current problems is that there is a certain anarchy, both in what is produced and among those who produce it.

- overdigitization (in quantity)
- differences of quality, interface, choices
- the quality is not taken into account
- problems of managers' training and problems of roles

Who should select the documents to digitize?

In order to prevent the anarchy, digitizations should obey rules:

- libraries should not decide alone which digitizations are conducted
- libraries should only have a consultative role
- domain specialists should also not decide alone what is digitized

In fact, **digitizations should be planned globally.**

The chaos and its solutions: digitize better

- ideally, **we should everywhere stop to digitize and start thinking**
- notion of “layer” in order to integrate the multiple versions of a digitization (integration of quality)
- take the quality into account: doing bad is acceptable, provided it is taken into account in the production chain
- **definition of a quality scale**
- stop doing local actions
- think about the sustainability of the digital object, and favor the quality and the reusability to the quantity
- understand the needs of the users
- be interested in the book

Selection mechanism

Some problems can get solved alone, merely using a natural selection mechanism.

- the evolution can occur by means of national or international dialogue;
- it can also occur by mere natural selection;
- the multiplicity of digitizations will produce a selection;
- but these two possibilities will not produce the same result; liberalism is not the solution!
- the cost is certainly not the same, for there are great losses with chaos;
- digitizations being made for all, do we have the right to do anything that pleases us?

The user has changed

Today's user is no longer the same as in the past:

- in the past, the books of a library were mainly meant for the immediate readers of this library (city, university, etc.)
- today, the libraries produce for a national, and even international, audience

Today, the libraries have much higher duties and responsibilities than before.

The frontiers of a digitization are no longer those of a university, and yet some digitization managers still have this mindset!

Gallica, the German libraries, etc., are all accountable to all.

Different kinds of users

- the average user, who may be demanding, does not want to look for a configuration menu;
- the average user is an occasional one; he/she comes for only one book;
- it is possible to please this average user and many sites are rather satisfying;
- but some of them, in particular those based on Contentdm, are not
- a user may want to study a book at home, without an internet connection; he/she must be able to download or to print a good quality version.

Overcautious libraries?

As we said above, the libraries do now give a feeling of timeworn institutions:

- will to control online documents (tatooing, steganography, etc.), in order to track the usages
- restrictions to downloading
- attachment to the lending model

The redefinition of the role of the libraries

- the libraries have a new role to play: what will the libraries do when *everything* will have been digitized?
- they hold the original documents
- they should not fear giving away the digital versions
- the new role of the libraries might be to enrich the documents (OCR, documentary files, etc.)

Document enrichment

- can be split from digitization, both in time and in space
- drift towards the **penetration into the documents**
- it is the added value provided by the libraries (like the catalogs)

Among the enrichment forms which will develop:

- the entire transcription of books, with typesetting (Gutenberg project and more, LOCOMAT project, etc.)
- the structured representation of books (XML, TEI, MathML, etc.)
- the production of additional files for anchoring a book on the web; this is the role of a documentation center, but such a role will necessarily penetrate that of libraries.

Managing quality

The quality of digitization requires a precise management:

- each library should maintain a list of digitizations which have to be redone
- the digitization quality can be appraised right away, and it can also be appraised outside of the library
- Gallica should for instance plan a new digitization of the book seen above (and many others)
- certain libraries possibly do it, but it is not visible

A simple scale of quality (1)

- We propose to give a letter from A to E for the quality of digitization. It is possible to distinguish the stored quality and the provided quality, but in general the user can only evaluate the provided quality.
- This scale only considers the images, and not the quality of the text or the metadata. This is a different problem.
- It is illusory to believe that everything will be OCR'ed anytime soon, especially mathematical texts and there is therefore a real need to evaluate the quality of the images.

A simple scale of quality (2)

- A: the digitization might replace the original document (resolution equal or greater than 600dpi, in color, well aligned, plates well scanned, no defect);
- B: correct, but not comfortable enough, or does not respect the colors, or other small defects which do not prevent the reading; plates well scanned; copying the digitization does not require understanding it;
- C: only one serious defect (insufficient resolution, unfolded plates, deformations, etc.);
- D: two serious defects;
- E: three serious defects;
- etc.

Normalization of online documents

- today, it is great time to think about normalizing the uploading of digitizations;
- precise rules should be followed by the major digitizers, both for the quality of digitizations as for the quality of the interfaces;
- this is unfortunately still far from being the case, as the examples in this document prove;
- in the future, the norms can of course evolve.

Manual or automatic digitization?

There are digitization machines which turn the pages automatically, but:

- old documents can not be handled without the risk to damage them;
- Google does not use entirely automatic machines;
- a quality management can currently not be obtained automatically, because **in order to digitize well, one must be able to read the content** (for alignment, for adapting the resolution, etc.);

Part of the bad digitizations are due to the fact that those who digitize are almost machines.

A good digitizer should understand the needs of the book.

The death of digitizations

- the digital world will be populated by a growing number of digitizations and there will be **many multiple digitizations** (the same original, or the same book several times) in different qualities
- this **profusion of information** will harm research, in the same way as “content farms” have made internet searches almost useless
- “content farms” drive the users to fall back on a few well-known places, such as Wikipedia
- the same might happen with digitizations, especially if **illegal copies of whole libraries** appear, which is likely
- if the quality is handled, then it is necessary to **plan for the elimination of bad digitizations**

The problem with microfilms

There is a fundamental difference between microfilms and the current digitizations:

- microfilms have never had the purpose to replace the originals, but to protect them or to make them accessible
- microfilms are almost always in B&W and of bad quality
- current digitizations are in color and can reach resolutions which make it almost superfluous to access the original documents, except in some rare cases

This difference of purpose explains that it was historically acceptable to produce bad microfilms.

The problem of cost

- cost determines the means
- the means might determine the quality
- in fact, there should be no compromise on quality
- compromises on quality are the result of false emergency problems
- with less money, less should be digitized
- one should not digitize less well

The physics of digitization and the useful information

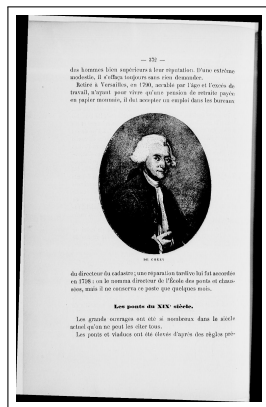
- notion of reachable and useful quality
- basic notions in image processing (example: rotations)
- characters should be smooth, because we are used to see them smooth; this is our **abstract notion of characters**
- smooth characters can contribute to comfort
- for paper, **an excessive resolution makes the fiber apparent**
- the maximal useful information is therefore related to the support
- it can be limited by the digitization conditions (optics, etc.)

A word on image rotations

- many digitized images are not aligned;
- it is possible to rotate the images;
- this rotation almost always produces a loss of quality (see our article on the topic)
- it is therefore essential to minimize this step, either by a better alignment, or by a greater resolution.

Example of rotation done by Gallica

Debauxe: *Les travaux publics et les ingénieurs des ponts et chaussées depuis le XVII^e siècle*, 1893



- the page was not correctly aligned at the beginning
- a subsequent operation rotated it, leaving the black border triangles
- this rotation caused a reduction in the apparent image resolution

Storage formats: with or without losses? (1)

- there is currently often an obsession with the TIFF format
- it is a format which can keep all the data from the original image, but it results in large files
- in the context of a “preservation” approach, it often seems natural for libraries to use this format internally
- but, **it is a wrong vision of the things!**
- if we select the format by reasoning in a binary way (losses or no losses), the digitization chain is not taken into account
- the TIFF is a lossless representation of the result of chain which has perturbed the object (book or else)

Storage formats: with or without losses? (2)

- the subcontractors for digitization will advise the use of TIFF, often because of ignorance of the other parameters
- some users also want TIFF, because their publishers ask for it
- by requesting TIFF, the publishers guard themselves from errors due to compression, but not from errors upstream in the digitization
- in almost all cases, libraries do not need to use this format; it doesn't do harm (except in storage), but it is not necessary
- the real challenge is to determine which losses are acceptable given the limits introduced by the production chain (in particular optical limits) and by the physiological limits (eye resolution)
- the absence of losses is in fact a chimera

A word on Europeana

Europeana is essentially a **portal** towards the European libraries:

- somewhat analogous to SUDOC or KVK (Karlsruhe) for catalogues
- the idea of gathering a **unified catalog of European digitized collections** is very good
- still very incomplete (the collections from the Strasbourg University Libraries are for instance not included)
- the interface is rather simple
- no unified vision of the works
- identical works are not grouped and the digitization quality is not measured

Several big problems

We can currently distinguish several problems:

- it is often difficult to know when a book has been digitized somewhere;
- some books have been digitized in several copies, although this wasn't always necessary;
- digitizations are scattered on many sites and presented in a heterogeneous way

Census of digitizations

In order to have a better grasp on the existence of a digitization, it is essential to have a global census of everything that has been digitized.

- some censuses do exist, but they are not complete (example: <http://digreg.mathguide.de> in Goettingen)
- another example is <http://www.zvdd.de/startseite>
- these censuses are based on lists provided by the libraries involved;
- there certainly is no notion of a unique identifier for a given copy of a book covering all libraries, as well as all uncatalogued collections outside of libraries
- it is however likely that the constitution of such a census will take place in a near future.

List of digitization tasks

The census of digitizations does not prevent the existence of multiple duplicates, because the census is not necessarily used, and that duplicates can be simultaneous.

In order to solve this problem, there should be a list of digitization tasks:

- unique identifier for items to digitize (books or parts of books)
- assignment of tasks
- identification of books with “marginalia” whose annotations require a new digitization, even if the book has already been digitized elsewhere

Central digital server

One of the advantages of Google is that its interface is always the same, no matter where the books are located.

- the libraries which are digitizing should deposit all their digitizations on a central server
- the presentation of all digitizations could then be unified
- Europeana is only a portal and only redirects towards the various libraries
- we can assume that this centralized server will one day see the light of day
- Google has forbidden certain of its libraries under contract to do such deposits using Google's digitizations, but it is certainly not forbidden with the new digitizations

The training of digitization managers

- the digitization managers should have a better knowledge of the actual technical problems
- one should not be content with an approximate knowledge provided by third parties
- in order to be a digitization manager, one should for instance have gone through (at least) the exploitation and edition of a non trivial book, which would then have forced the candidates to use the collections they want to manage
- nowadays, some people become digitization managers without having done any work on digitization;
- comparing interfaces, and doing other market studies, should not be considered sufficient;
- the digitization manager should be a visionary and should not work alone in his office; digitization is a global problem.

Required competencies

A digitization manager should

- be interested by the entire knowledge
- have done historical research through books and manuscripts
- have technical competencies: in computer science, physics, optics (in particular physiological), etc.
- understand the needs of the users and interact with them
- understand the needs of the books
- get involved in global programs
- seek quality
- be a visionary: have a long term vision, be able to anticipate

Required competencies (2)

- it is not forbidden to have graduated from a school for archivists, or for librarians, or to have a PhD in history, or even to know dead languages, but this plays only a minor role in the work of a digitization manager
- on the other hand, if the manager lacks certain competences, he/she must be able to surround him/herself with people having these competences, and not venture alone in such actions
- some of the current insufficiencies find their source in the inadequacies between the profiles of certain managers and the required competencies (problems of legitimacy)

And ... if quality is not aimed?

It is however possible to live with average digitizations, provided that:

- free and fast sustainable digitization is available on demand
- digitization errors can be quickly corrected.

At this point, no library, and certainly not Google, provides such a service.

Should one come to terms with Google?

- a contract with Google makes it possible to have quickly a large number of average quality digitizations;
- it is possible to improve this situation by asking that Google follows more restrictive clauses:
 - immediate detection of errors (almost all errors can be identified immediately);
 - either a book is digitized entirely (with unfolded plates), or not at all, but not partially;
 - Google must commit not to put online (and not to sell) incomplete versions;
- in any case, even with these caveats, the contracting libraries must understand that the **digitization will have to be redone some time in the future** (possibly a very distant one).

Perspectives

Today, a large majority of the digitizations produced in libraries will have to be redone. We are far from sustainability!

- The quality of digitizations will increase when the competencies will be better distributed and when the needs will be well understood by all.
- The quantity and order of digitizations will grow when the digitization will be better organized, nationally, and then at the European and world level, and not only organized by random deposits of uncontrolled digitizations.

An optimistic note for the end?

The fact that today's digitizations are not sustainable does not mean that they should not be done:

- they help to make aware of the problems and they help to reach a better definition of the needs (but do we have to first have non sustainable digitizations?);
- one might almost draw a parallel with the generations of machines, generations of mobile phones (1G, 2G, 3G, etc.), etc., each generation being used to access or prepare the next one; but is it really the same with digitizations?

Thanks to all for your input
and have fun digitizing!