Mass digitization and digitization projects at National library of Florence

Giovanni Bergamin
Biblioteca Nazionale
Centrale Firenze

Some definitions …

Mass digitization of books ***(MDB)*** =
conversion of materials (books) *on an industrial scale* (not just on a large-scale); *conversion* of whole libraries *without making a selection* of individual materials

source: Karen Coyle

Two main MDB projects:

Google Books

Google books

Internet Archive (Open Content Alliance)

INTERNET ARCHIVE

OPEN Content Alliance

Building a digital archive of global content
for universal access

Some notes on Google Books_1

Started in 2004

Planned end of the project 2020

The Google Books aim is the Google aim:

> “organize the world's information and make it universally accessible and usable” --- so the ***content of all published books*** has to be searchable together with the ***content of all web*** pages

Some notes on Google Books_2

“Just how many books are out there?”

Books of the world, stand up and be counted! All 129,864,880 of you.

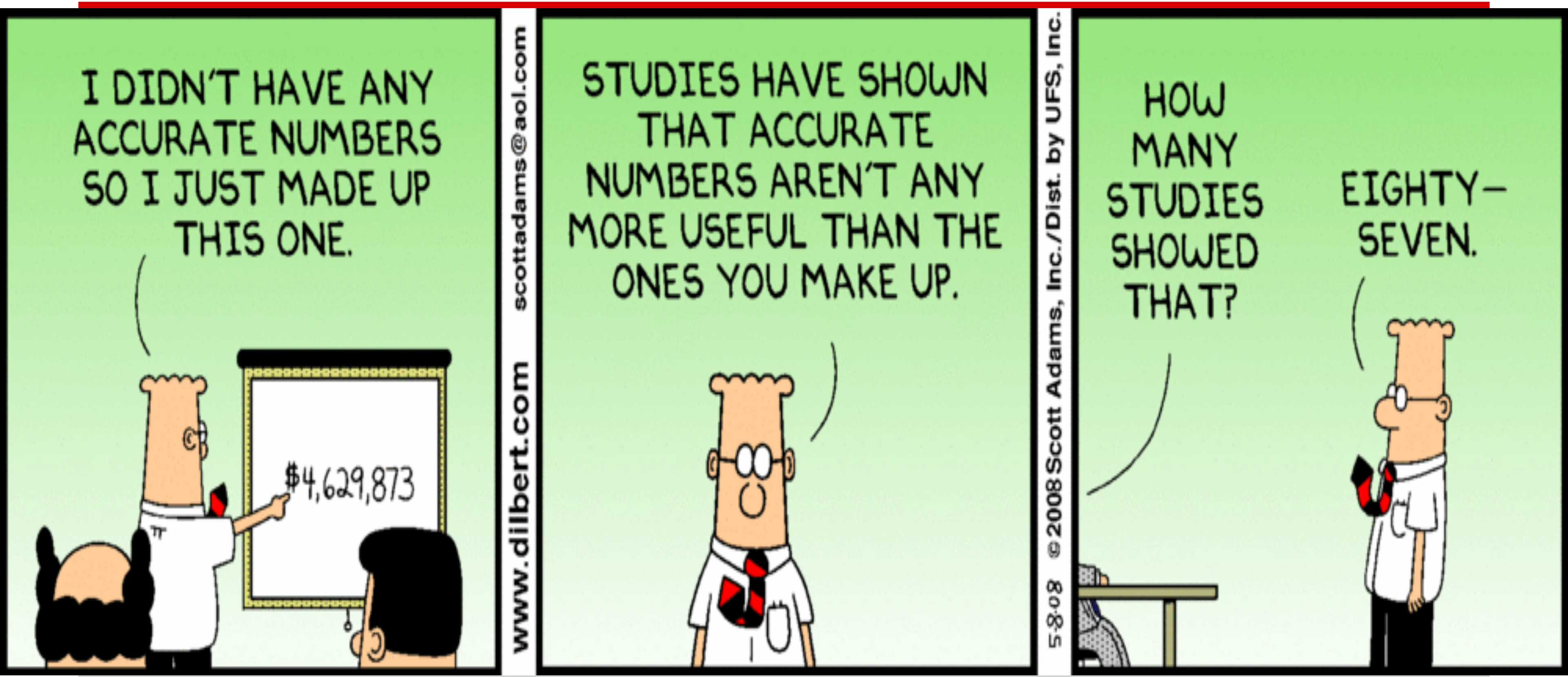
Thursday, August 05, 2010 at 8:26 AM

Posted by Leonid Taycher, software engineer

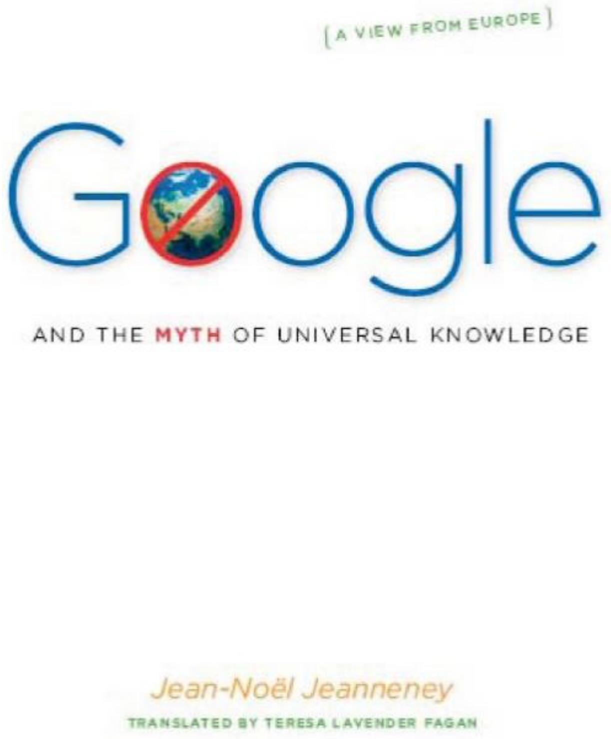
How many books have already been digitized by Google Books?

25-30M (non ci sono statistiche ufficiali)

Numbers ...


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
scottadams@aol.com
www.dilbert.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5-8-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

A famous debate on GB in 2005_1



(A VIEW FROM EUROPE)

Google

AND THE MYTH OF UNIVERSAL KNOWLEDGE

Jean-Noël Jeanneney

TRANSLATED BY TERESA LAVENDER FAGAN

Jean-Noël Jeanneney

Quand Google défie l'Europe

Plaidoyer pour un sursaut

Essai

MILLE ET UNE NUITS

A famous debate on GB in 2005_2

Jean-Noël Jeanneney, historian and former President of National Library of France wrote in 2005 that:

> “*The promise of Google is enchanting [...]: everyone with access to the Internet can soon view the recorded memory of the ages in the palm of their hand and search this universe in a fraction of a second*” … *however* ...

A famous debate on GB in 2005_3

“We are faced with several possible dangers with respect to:

works of various cultural heritages that have fallen into the public domain, the ***list of priorities will likely weigh in favor of Anglo-Saxon culture****;*

works still under copyright, of which only excerpts, or "snippets," will be offered for the time being, the ***weight of American publishers may be overwhelming****;*

journals and books disseminating ongoing research, the ***dominance of work from the United States*** *may become even greater than it is today”*

11 years later ...

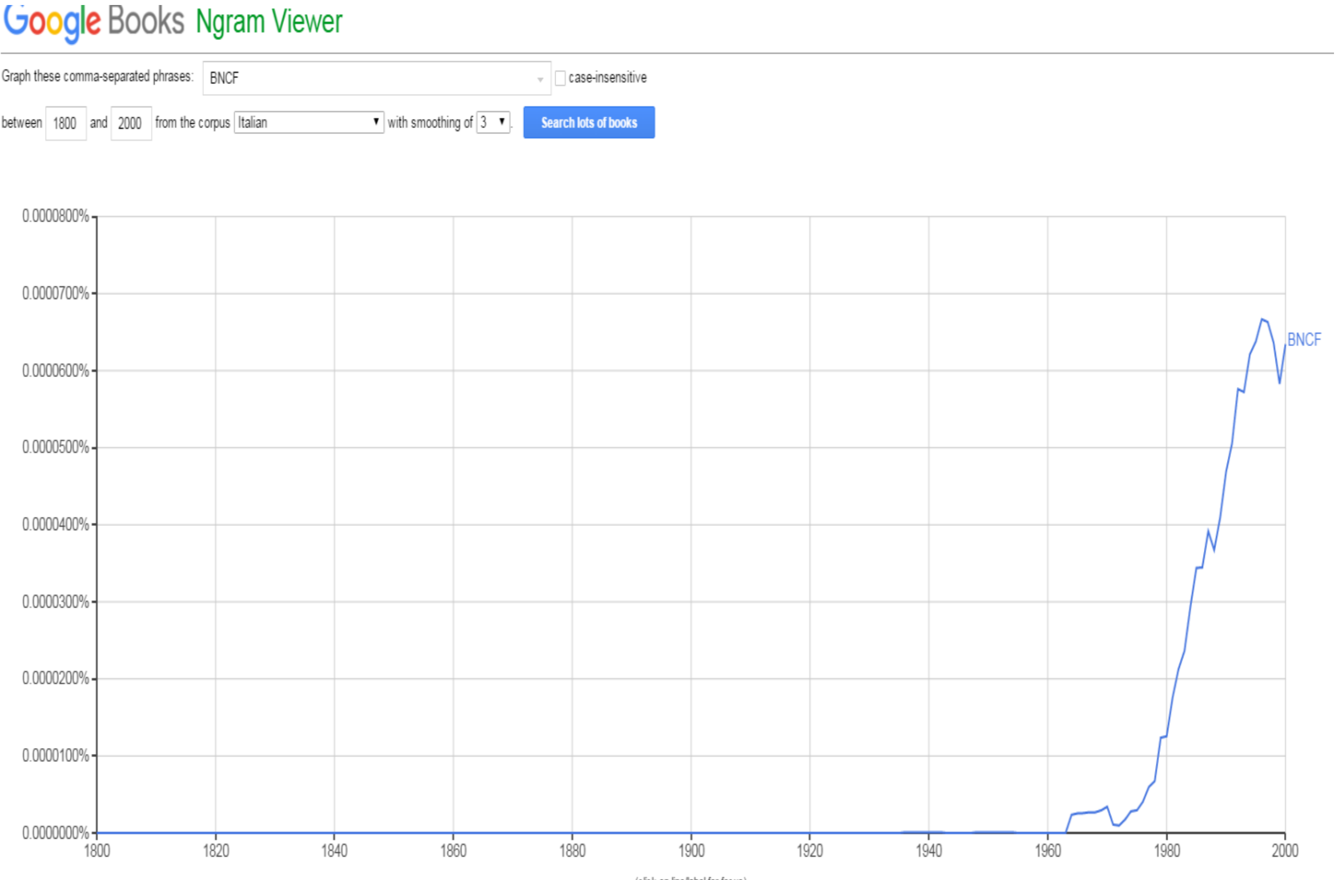
according to “reliable sources” the highest percentage of the digitized books is in English (close to 50% out of 450 languages of books in GB)

The interest of Google for non English languages is growing

ex. g. Ngram Viewer

Ngram service now available also for texts in: German,French, Italian, Spanish, Russian, Hebrew, Chinese

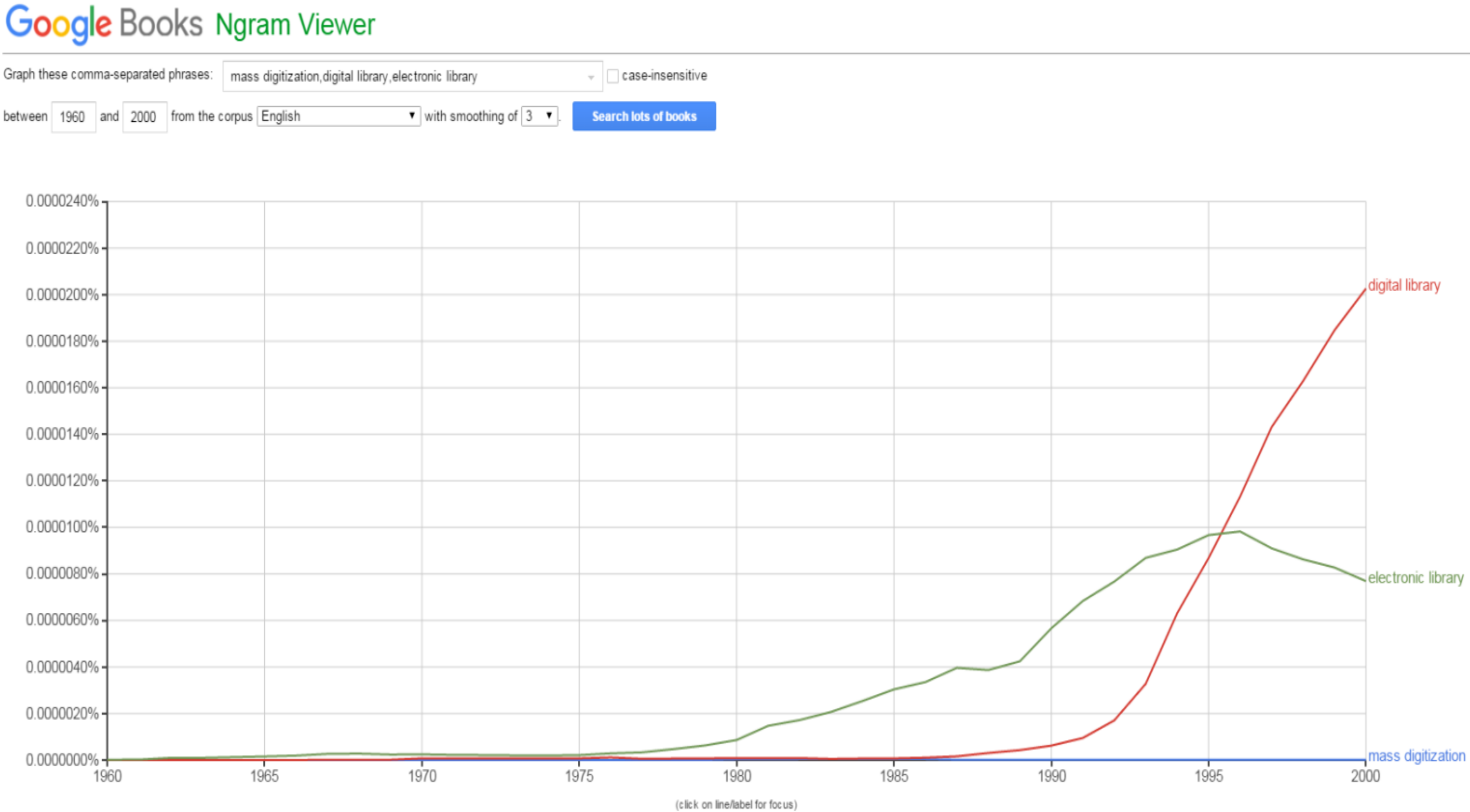
From 2009 year after year they are adding new languages



Google Books Ngram Viewer

Graph these comma-separated phrases: BNCF □ case-insensitive

between 1800 and 2000 from the corpus Italian with smoothing of 3



Graph these comma-separated phrases: mass digitization,digital library,electronic library ☐ case-insensitive

between 1960 and 2000 from the corpus English with smoothing of 3.

Somes notes on Internet Archive (OCA)_1

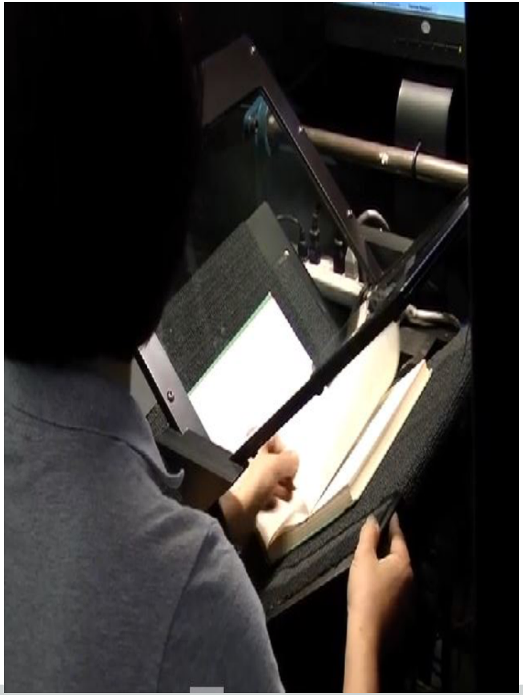
The Open Content Alliance (OCA) is a consortium of organizations contributing to a permanent, publicly accessible archive of digitized texts.

Its creation was announced in October 2005 by Yahoo!, the Internet Archive, the University of California, the University of Toronto and others.

Scanning for the OCA is administered by the Internet Archive, which also provides permanent storage and access through its website

Somes notes on Internet Archive (OCA)_2

More than 8,7 million of texts available up to now



Some differences

Google books

- private owned (1 company)
  - huge amount of resources available for
    - digitization
    - research and development

INTERNET ARCHIVE

- consortium between companies and nonprofit institutions
- depends on donations and self-financing (limited resources available)

Digitization projects at BNCF (DPB)

started in the early 1990s when the size of HD was 40 Megabytes (and there was no WWW)

WHEN I STARTED PROGRAMMING, WE DIDN'T HAVE ANY OF THESE SISSY "ICONS" AND "WINDOWS."
© 1992 United Feature Syndicate, Inc.
S.Adams
ALL WE HAD WERE ZEROS AND ONES -- AND SOMETIMES WE DIDN'T EVEN HAVE ONES.
9-8
I WROTE AN ENTIRE DATABASE PROGRAM USING ONLY ZEROS.
YOU HAD ZEROS? WE HAD TO USE THE LETTER "O."

DPB: faithful copy or searchable text? - 1

Early projects aim:

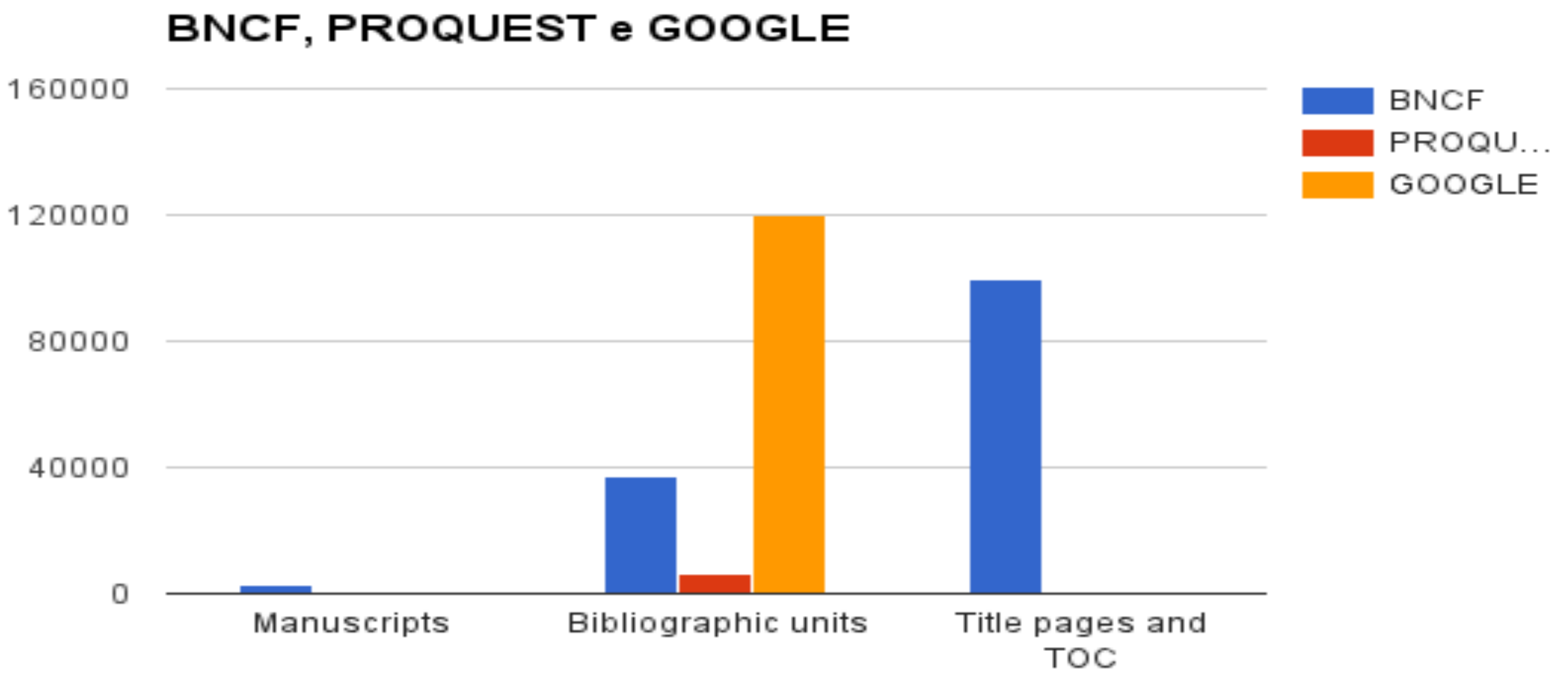
- enrichment of bibliographic records through the digitization of title pages, table of contents etc (OCR)

DPB: faithful copy or searchable text? - 2

Following projects aim:

- faithful copy for manuscripts, ancient books,
  - maps etc

DPB results


BNCF, PROQUEST e GOOGLE
160000
120000
80000
40000
0
Manuscripts
Bibliographic units
Title pages and TOC
BNCF
PROQU...
GOOGLE

Google Books and Proquest EEB at BNCF - 1

GB	EEB
range: 1701-1875	*range:* -1700
ebooks with “liquid” and searchable text	faithful copy
national project	BNCF project
costs: books circulation (inside lib.)	*costs:* none

Google Books and Proquest EEB at BNCF - 2

GB

- *scanning location:* outside lib. (Italy)
- *outcomes:* GRIN and free worldwide accessibility

EEB

- *scanning location:* inside lib.
- *outcomes:* "master" files and
  - free access from Italian IP
  - access fee outside and royalties for BNCF

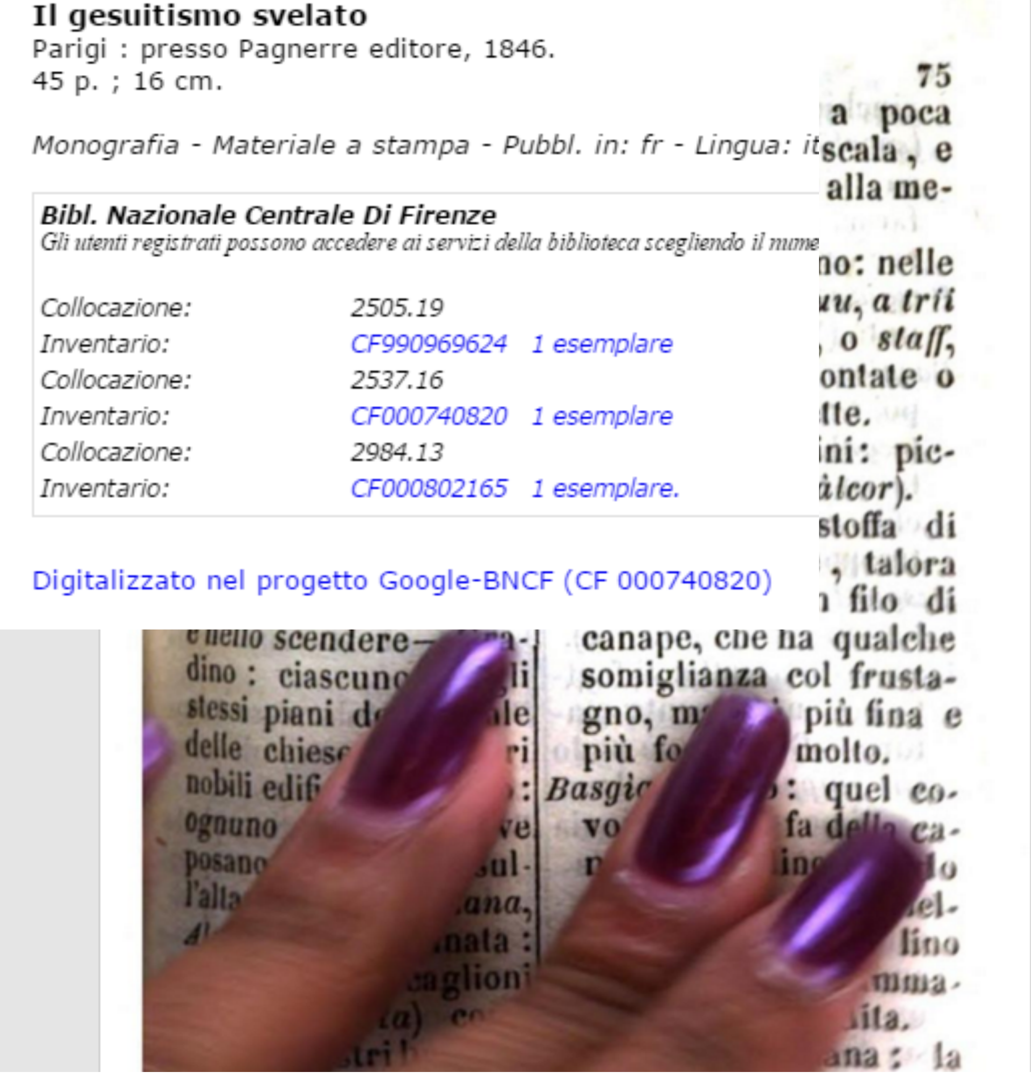
MD problems (Google)

limitations, ex. g.:

- size of books
- foldouts (from 2016 it will be possible)

note:

- MPOB Modified Process for older books pre 1700 (color *and* text)



Il gesuitismo svelato
Parigi : presso Pagnerre editore, 1846.
45 p. ; 16 cm.

Monografia - Materiale a stampa - Pubbl. in: fr - Lingua: it

Bibl. Nazionale Centrale Di Firenze
Gli utenti registrati possono accedere ai servizi della biblioteca scegliendo il nume

Collocazione:	*2505.19*	
Inventario:	*CF990969624*	*1 esemplare*
Collocazione:	*2537.16*	
Inventario:	*CF000740820*	*1 esemplare*
Collocazione:	*2984.13*	
Inventario:	*CF000802165*	*1 esemplare.*

Digitalizzato nel progetto Google-BNCF (CF 000740820)

75
a poca
scala, e
alla me-
no: nelle
uu, a trii
, o staff,
ontate o
tte.
ini: pic-
àlcor).
stoffa di
, talora
filo di

e nello scendere—
dino: ciascuno
stessi piani d
delle chiese
nobili edif
ognuno
posano
l'alta
A
nata:
aglioni
a) co
tri

canape, che ha qualche
somiglianza col frusta-
gno, m più fina e
più fo molto.
Basgi : quel co-
vo fa della ca-
lo
el-
lino
mma-
ita.
ana: la

distanza dalla scab, e che acceniiatio alla medesima. Basellin. Predellino: nelle carrozze — *a duu, a trii pass* o *pastad*, **o** *staff*, a due, a Ire **montate o** ballenti **o** palelle. *Basgerilt*. Calcolini: piecole calcóle *(calcar)*. *Basen*. Basilio: **stoffa** di filo di **colone** , **talora** mista anco con filo di canape, **che** ha qualclie col fruelapiú **fina e molto.** : qui' fa

rinchiuso in
baccello
(sgorbia) erboso
, cra.s so, più
facile a
pulrefarsi che
non a
disseccarsi.

Basyinnoeura.
Fava minuta
(Vicia faba
minori.

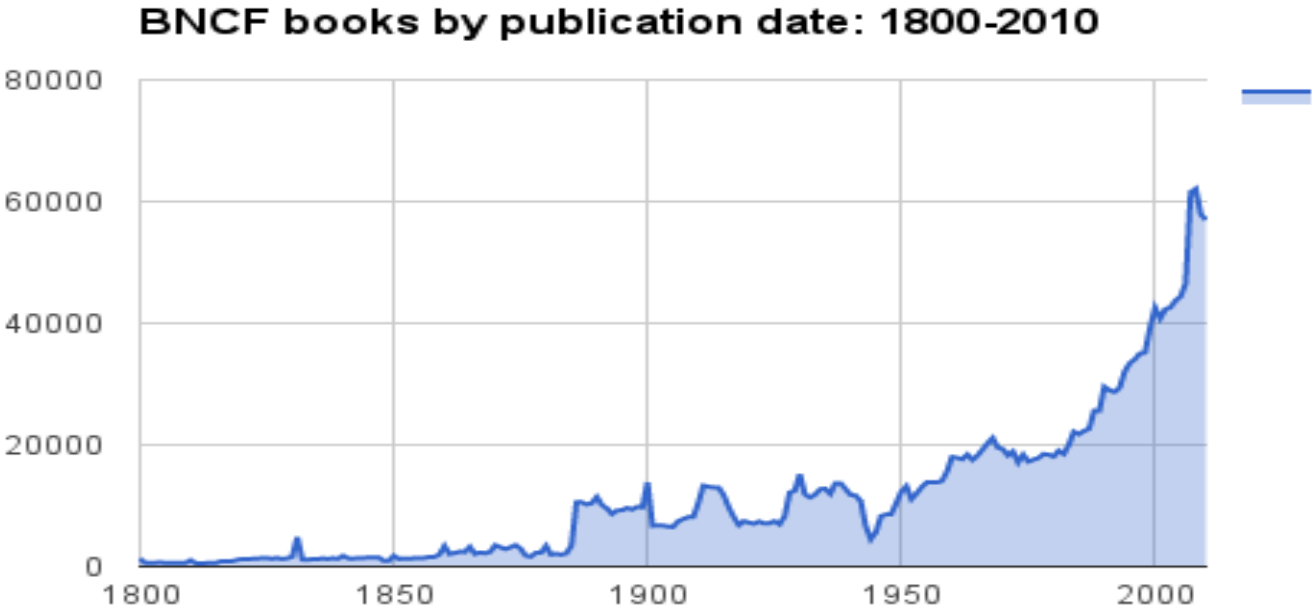
, ove appender i

/» ciappa), riè a

piede,

canape, che ha qualche
somiglianza col frusta-
gno, m più fina e
più fo molto.
Basgi : quel co-
vo fa della ca-
lo
el-
lino
mma-
ita.
ana: la
ato assai
Bazzetta: la
agnellino (be-
n nato, o da po-
Fava (Vicia faba);
ume (lemm) di for-
na bislunga, schiaccia-
ta, col bellico (oggia),
o segno del germe a
una delle estremità, e

MD and copyright (orphan works etc.)

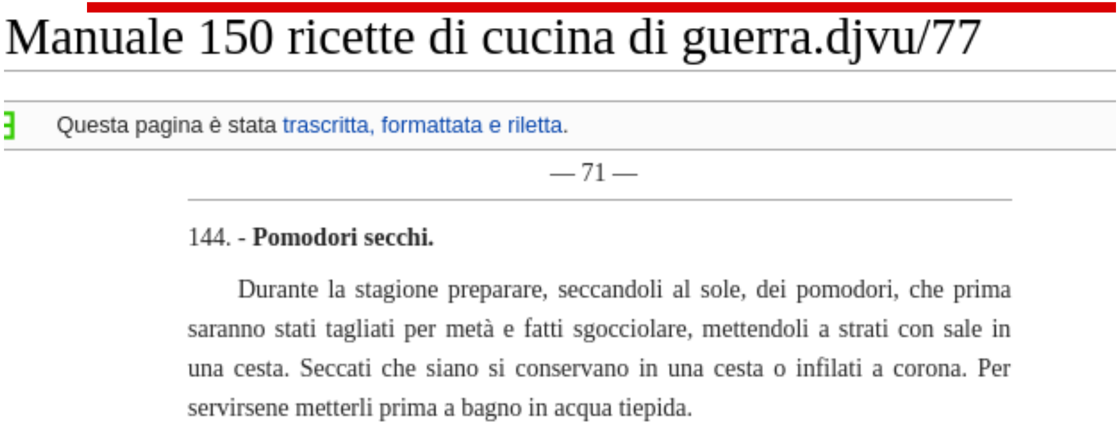


BNCF and Wikisource

2014 Agreement between BNCF and Wikimedia Italia for Wikisource

- starting point: public domain book digitized by BNCF
- aim: improve access to digitized books
- results: crowdsourced text correction in Wikisource (the free library that anyone can improve)

How it works



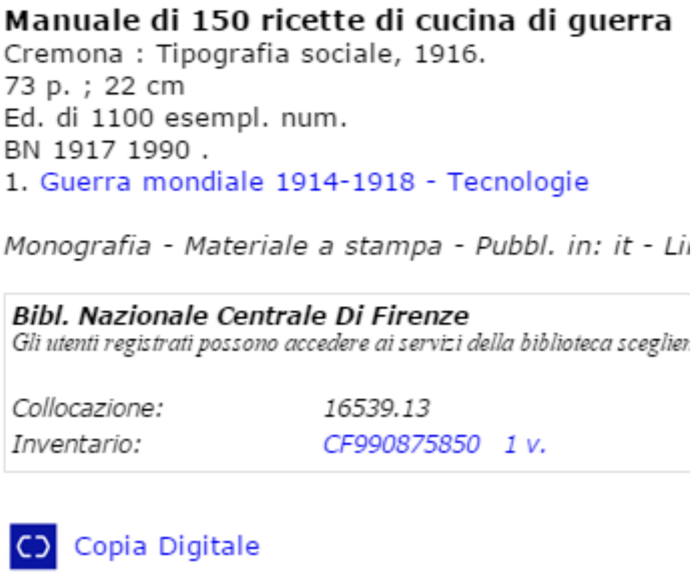
Manuale 150 ricette di cucina di guerra.djvu/77

Questa pagina è stata trascritta, formattata e riletta.

— 71 —

144. - **Pomodori secchi.**

Durante la stagione preparare, seccandoli al sole, dei pomodori, che prima saranno stati tagliati per metà e fatti sgocciolare, mettendoli a strati con sale in una cesta. Seccati che siano si conservano in una cesta o infilati a corona. Per servirsene metterli prima a bagno in acqua tiepida.



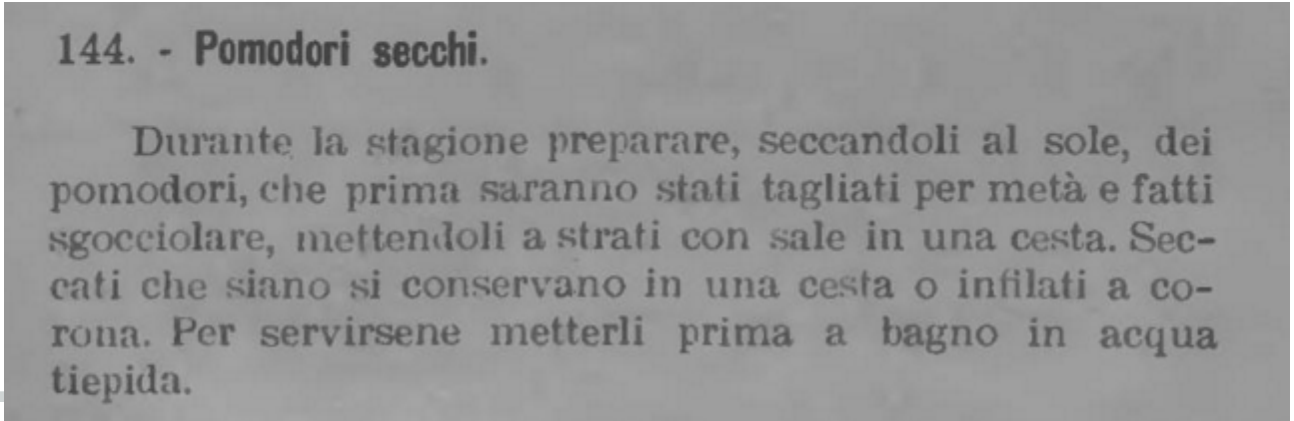
Manuale di 150 ricette di cucina di guerra
Cremona : Tipografia sociale, 1916.
73 p. ; 22 cm
Ed. di 1100 esempl. num.
BN 1917 1990 .
1. Guerra mondiale 1914-1918 - Tecnologie

Monografia - Materiale a stampa - Pubbl. in: it - Li

Bibl. Nazionale Centrale Di Firenze
Gli utenti registrati possono accedere ai servizi della biblioteca scegliei

Collocazione:	*16539.13*
Inventario:	*CF990875850 1 v.*

Copia Digitale



144. - **Pomodori secchi.**

Durante la stagione preparare, seccandoli al sole, dei pomodori, che prima saranno stati tagliati per metà e fatti sgocciolare, mettendoli a strati con sale in una cesta. Seccati che siano si conservano in una cesta o infilati a corona. Per servirsene metterli prima a bagno in acqua tiepida.

Closing remarks and open questions

MDB: Is there an alternative to Google Books?

cooperation with IA and Wikisource

140 years buffer (orphan works and cooperation with publishers)

Thank you for your patience

giovanni.bergamin@gmail.com