



Koninklijke Bibliotheek  
National library of the Netherlands

# Library challenges for mass digitisation

Astrid Verheusen  
Head Digitisation Department  
Research & Development Division  
National library of the Netherlands

IMPACT conference 2009

The Hague  
6 April 2009

**Improving access to text is not the most  
urgent challenge...**

## Outline presentation

- Past, current and future digitisation activities
- Issues concerning mass digitisation
- Possible solutions
- Expectations from IMPACT

## Digitisation in the past

- Experience with digitisation since 1994
- Webexpositions / highlights of collections
- Small-scale digitisation projects
- Mainly visually attractive images
- Emphasis on techniques / trial and error
- Exploration of possibilities
- Co-operation on a small scale



## Digitisation 2000-2005

### Shift in emphasis:

- From highlights to larger collections
- Project based
- (Inter)national co-operation
- Established methods and techniques
- Awareness of digital preservation
- More text material



Drawings from the camps

# Drawings from the camps

in the occupied Dutch East Indies

**Search the Memory of the Netherlands**

372576 objects from 67 collections of 62 institutions [advanced search](#)

[>> Themes](#) [>> Exhibitions](#)

## Digitisation - present & future - 1

- Strategic plan 2006-2009: "Development of a national programme for the mass digitisation of sources for research in the humanities"
- Target audience
  - Scientific research
  - Public at large
- Particular attention for digital preservation
- Preservation imaging
- No commercial partners for funding

## Digitisation - present & future - 2

### Text digitisation

- Until recently: on a small scale
- Printed and typed sources (not handwritten)
- Issues differ from images
  - Structure / navigation
  - Conversion to full text (OCR)
  - Scanning from microfilm
  - Search & Retrieval

## Digitisation - present & future - 3

### Digital Library Programme (2009-2013)

- 10 % of our collection available in digital form in 2013

### Far future

- 100% of all books published in- and about the Netherlands available in digital form and have control of quality and availability

## Projects 2009-2011

Project	Number of pages	Budget
Dutch parliamentary papers 1814-1995	2.300.000	M€ 10.5
Dutch daily newspapers 1618-1995	8.000.000	M€ 12.5
Special collections – books before 1800	1.300.000	M€ 3.0
Magazines	1.500.000	M€ 1.8
Books 1850-1950	12.000.000	M€ 1.0
Radio news bulletins	1.500.000	M€ 0.5
Metamorfoze - preservation imaging	4.500.000	M€ 18.0
Memory of the Netherlands	350.000	M€ 3.5
<b>Total</b>	<b>31.450.000</b>	<b>M€ 50.8</b>

## Issues

- Costs
  - Digitisation: € 1.5 per page
  - Exploitation: millions per year from 2011 onwards
- Technical infrastructure
  - Storage (1 PB needed)
  - Processing (2 million files per month)
- Search & retrieval is not effective enough
  - Quality of OCR
- Organisational infrastructure is not efficient
- The process is too slow, we want to digitise faster and more...

## What happens around us?

Example: *Vaderlandsche Letteroefeningen*

- A cultural periodical offering reviews on a wide variety of subjects
- 1761- 1876
- Selected for Dutch Prints Online: digitisation of books before 1800
- 1780-1800: 13.000 pages

## 1. Google books



The screenshot shows a Windows Internet Explorer browser window displaying a Google Books search result. The browser's address bar shows the URL: <http://books.google.nl/books?id=exVKAAAAMAAJ&dq=Vaderlandsche+Letteroefeningen&printsec=frontcover&source=bl&ot>. The search bar contains the text "Vaderlandsche Letteroefeningen".

The main content area displays the title page of the book "Vaderlandsche letteroefeningen". The text on the page is as follows:

**VADERLANDSCHE  
LETTEROEFENINGEN,**  
OF  
T I J D S C H R I F T  
VAN  
KUNSTEN EN WETENSCHAPPEN,  
WAARIN DE  
BOEKEN EN SCHRIFTEN,  
DIE DAGELIJKS IN ONS VADERLAND EN  
ELDERS UITKOMEN, OORDEELKUNDIG  
TEVENS EN VRIJMOEDIG VERHAN-  
DELD WORDEN.  
BENEVENS  
**M E N G E L W E R K,**  
*tot Fraaije Letteren, Kunsten en Wetenschappen,*

On the right side of the page, there are several interactive options:

- [Downloaden PDF - 27.4M](#)
- [Platte tekst weergeven](#)
- [Recensie schrijven](#)
- [Aan mijn bibliotheek toevoegen](#)
- Bestel via een boekhandel**
- [Lokale boekwinkels zoeken](#)
- [Standaard HTML-modus](#)
- [Markeer deze pagina als onleesbaar](#)

## 2. Hathi Trust Digital Library



The screenshot shows a web browser window displaying the Hathi Trust Digital Library interface. The browser's address bar shows the URL: <http://babel.hathitrust.org/cgi/pt?id=mdp.39015065369624>. The page title is "Vaderlandsche letteroefeningen. 1863 pt.2".

The Hathi Trust Digital Library logo is visible at the top left, with navigation links for "Public Collections", "My Collections", and "Getting Started". A search bar is present with the text "Search in this text" and a "go" button. A "bookmark" icon is also visible.

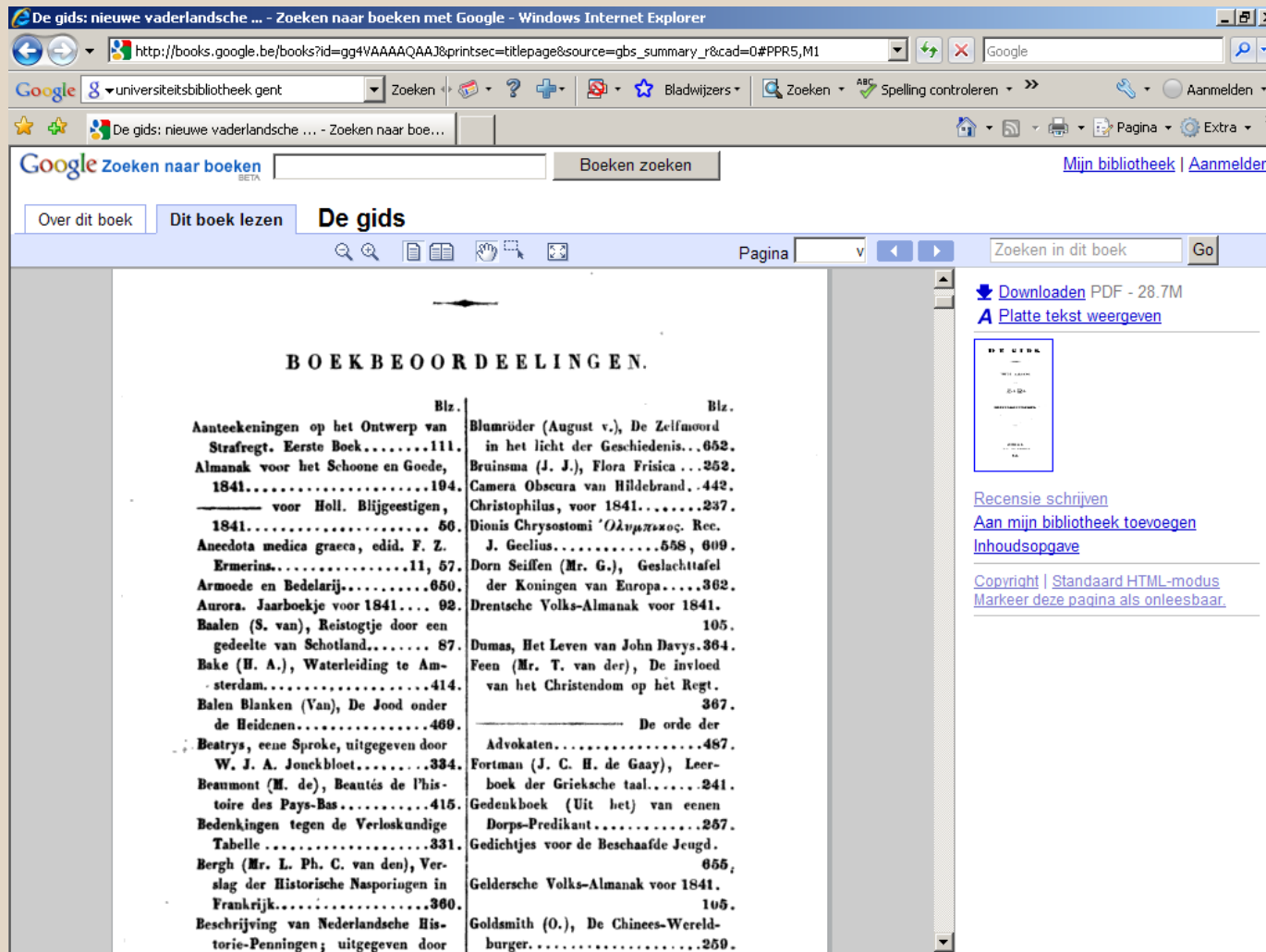
The main content area displays a scanned page with the following text:

**VADERLANDSCHE**  
**LETTEROEFENINGEN,**  
 OF  
**TIJDSCHRIFT**  
 VAN  
**Kunsten en Wetenschappen,**  
 WAARIN DE  
**BOEKEN EN SCHRIFTEN,**

On the left side of the interface, there are several utility sections:

- Login** to make your personal collections permanent.
- Add to your collection:** A dropdown menu for "Select Collection" and an "Add" button.
- Mirlyn Record** with navigation arrows.
- go to #** with an input field.
- size** set to 100% with a dropdown menu.
- rotate** with circular arrows.
- view page as** with options for "image", "text", and "PDF".

### 3. University Library Gent



De gids: nieuwe vaderlandsche ... - Zoeken naar boeken met Google - Windows Internet Explorer

http://books.google.be/books?id=gg4VAAAAQAAJ&printsec=titlepage&source=gbs\_summary\_r&cad=0#PPR5,M1

Google universiteitsbibliotheek gent Zoeken Bladwijzers Spelling controleren Aanmelden

Google Zoeken naar boeken Boeken zoeken Mijn bibliotheek | Aanmelden

Over dit boek Dit boek lezen **De gids**

Pagina Zoeken in dit boek Go

**BOEKBEORDEELINGEN.**

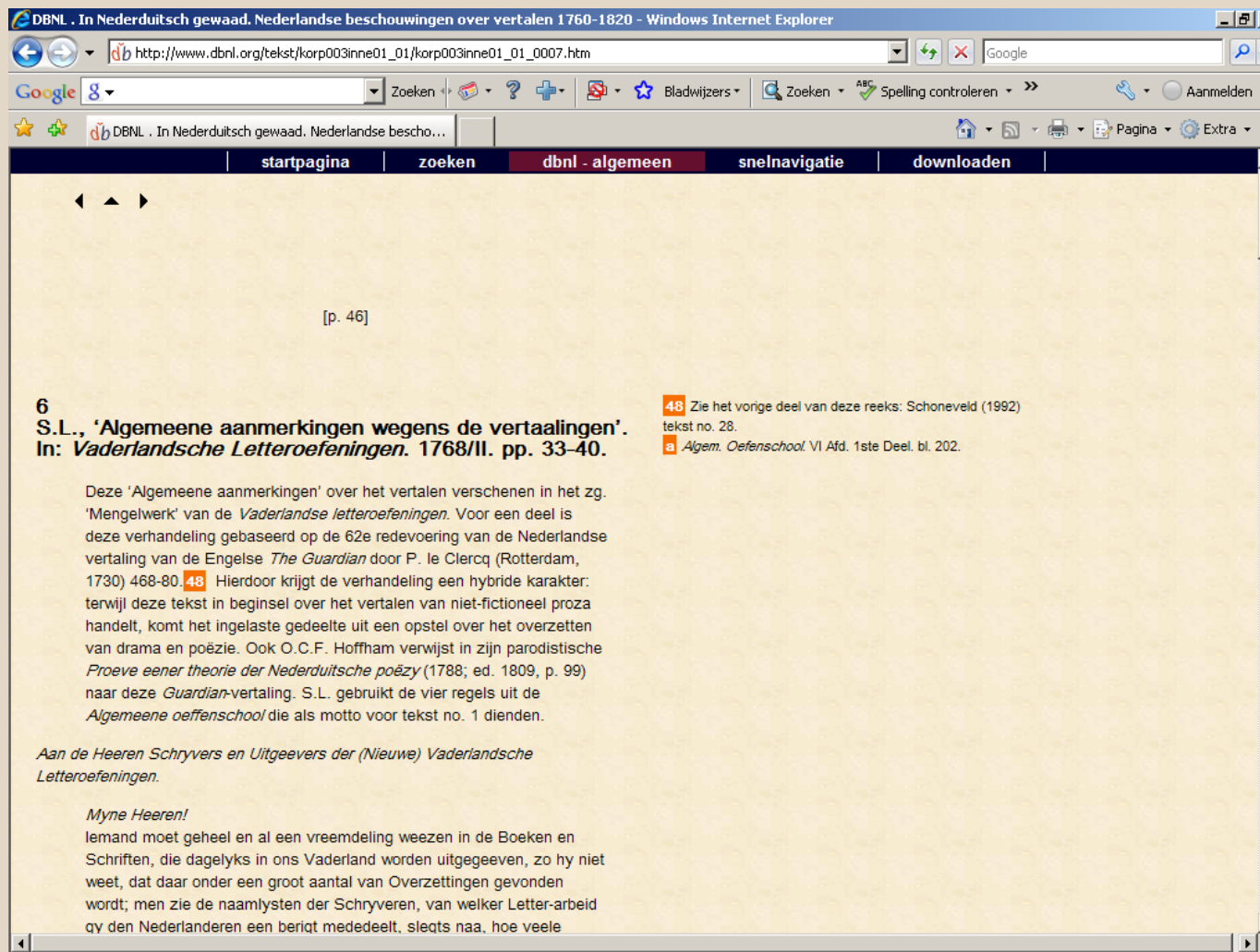
	Blz.		Blz.
Aanteekeningen op het Ontwerp van Strafrecht. Eerste Boek.....	111.	Blumrüder (August v.), De Zelfmoord in het licht der Geschiedenis...	652.
Almanak voor het Schoone en Goede, 1841.....	194.	Bruinsma (J. J.), Flora Frisica .....	352.
— voor Holl. Blijgeestigen, 1841.....	50.	Camera Obscura van Hildebrand.....	442.
Anekdota medica graeca, edid. F. Z. Ermerins.....	11, 57.	Christophilus, voor 1841.....	237.
Armoede en Bedelarij.....	650.	Dionis Chrysostomi 'Ολυμπιακος. Rec. J. Geelius.....	558, 609.
Aurora. Jaarboekje voor 1841.....	92.	Dorn Seiffen (Mr. G.), Geslachttafel der Koningen van Europa.....	362.
Baalen (S. van), Reistogtje door een gedeelte van Schotland.....	87.	Drentsche Volks-Almanak voor 1841.	105.
Bake (H. A.), Waterleiding te Amsterdam.....	414.	Dumas, Het Leven van John Davys.....	364.
Balen Blanken (Van), De Jood onder de Heidenen.....	469.	Feen (Mr. T. van der), De invloed van het Christendom op het Regt.....	367.
Beatrys, eene Sproke, uitgegeven door W. J. A. Jonckbloet.....	334.	— De orde der Advokaten.....	487.
Beaumont (M. de), Beautés de l'histoire des Pays-Bas.....	415.	Fortman (J. C. H. de Gaay), Leerboek der Grieksche taal.....	241.
Bedenkingen tegen de Verloeskundige Tabelle.....	331.	Gedenkboek (Uit het) van eenen Dorps-Predikant.....	257.
Bergh (Mr. L. Ph. C. van den), Verslag der Historische Nasporingen in Frankrijk.....	300.	Gedichtjes voor de Beschaafde Jeugd.	655.
Beschrijving van Nederlandsche Historie-Penningen; uitgegeven door		Geldersche Volks-Almanak voor 1841.	105.
		Goldsmith (O.), De Chinees-Wereldburger.....	259.

Downloaden PDF - 28.7M  
 Platte tekst weergeven

Recensie schrijven  
 Aan mijn bibliotheek toevoegen  
 Inhoudsopgave

Copyright | Standaard HTML-modus  
 Markeer deze pagina als onleesbaar

## 4. Digital Library of Dutch Literature



DBNL . In Nederlandsch gewaad. Nederlandse beschouwingen over vertalen 1760-1820 - Windows Internet Explorer

http://www.dbnl.org/tekst/korp003inne01\_01/korp003inne01\_01\_0007.htm

startpagina zoeken dbnl - algemeen snelnavigatie downloaden

[p. 46]

**6**  
**S.L., 'Algemeene aanmerkingen wegens de vertaalingen'.**  
**In: *Vaderlandsche Letteroefeningen*. 1768/II. pp. 33-40.**

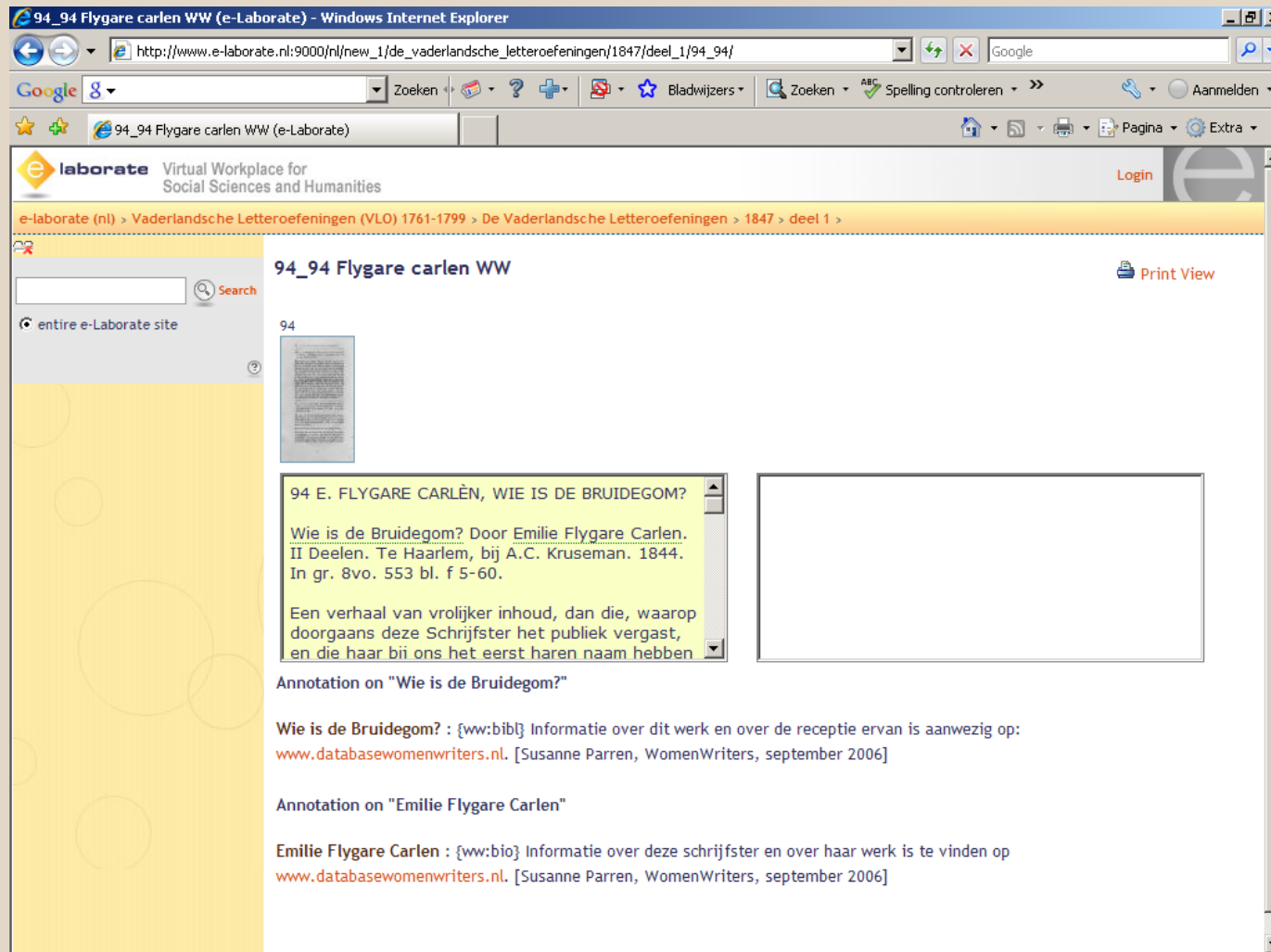
Deze 'Algemeene aanmerkingen' over het vertalen verschenen in het zg. 'Mengelwerk' van de *Vaderlandsche letteroefeningen*. Voor een deel is deze verhandeling gebaseerd op de 62e redevoering van de Nederlandse vertaling van de Engelse *The Guardian* door P. le Clercq (Rotterdam, 1730) 468-80. **48** Hierdoor krijgt de verhandeling een hybride karakter: terwijl deze tekst in beginsel over het vertalen van niet-fictioneel proza handelt, komt het ingelaste gedeelte uit een opstel over het overzetten van drama en poëzie. Ook O.C.F. Hoffham verwijst in zijn parodistische *Proeve eener theorie der Nederduitsche poëzy* (1788; ed. 1809, p. 99) naar deze *Guardian*-vertaling. S.L. gebruikt de vier regels uit de *Algemeene oeffenschool* die als motto voor tekst no. 1 dienden.

*Aan de Heeren Schryvers en Uitgeevvers der (Nieuwe) Vaderlandsche Letteroefeningen.*

*Myne Heeren!*  
 Iemand moet geheel en al een vreemdling weezen in de Boeken en Schriften, die dagelyks in ons Vaderland worden uitgegeeven, zo hy niet weet, dat daar onder een groot aantal van Overzettingen gevonden wordt; men zie de naamlysten der Schryveren, van welker Letter-arbeid ay den Nederlanderen een bariat mededeelt. sleqts naa, hoe veele

**48** Zie het vorige deel van deze reeks: Schoneveld (1992) tekst no. 28.  
**a** *Algem. Oeffenschool*. VI Afd. 1ste Deel. bl. 202.

## 5. Netherlands Institute for Scientific Information Services



94\_94 Flygare carlen WW

94

94 E. FLYGARE CARLÈN, WIE IS DE BRUIDEGOM?

Wie is de Bruidegom? Door Emilie Flygare Carlen. II Deelen. Te Haarlem, bij A.C. Kruseman. 1844. In gr. 8vo. 553 bl. f 5-60.

Een verhaal van vrolijker inhoud, dan die, waarop doorgaans deze Schrijfster het publiek vergast, en die haar bij ons het eerst haren naam hebben

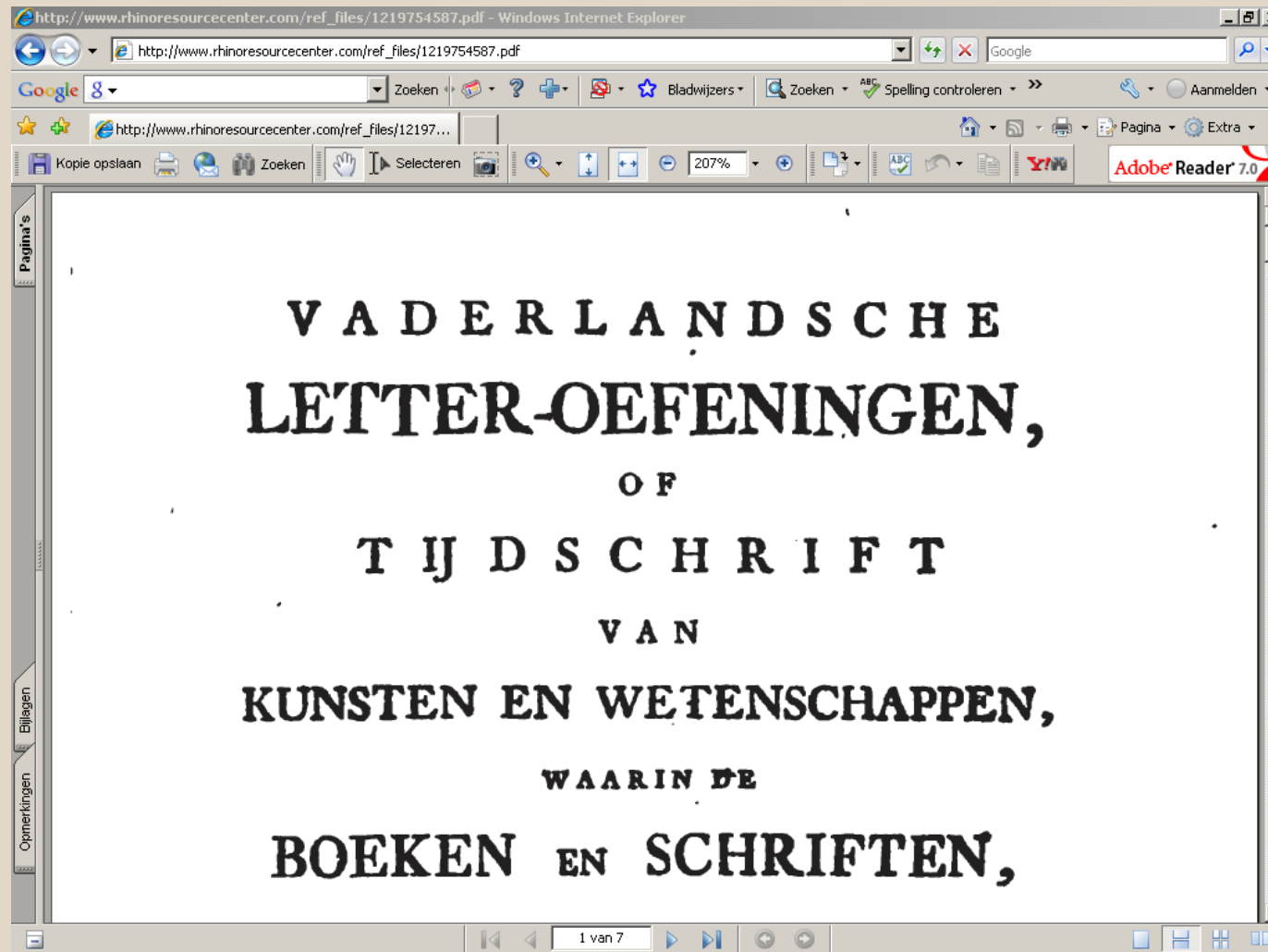
Annotation on "Wie is de Bruidegom?"

Wie is de Bruidegom? : {ww:bib} Informatie over dit werk en over de receptie ervan is aanwezig op: [www.databasewomenwriters.nl](http://www.databasewomenwriters.nl). [Susanne Parren, WomenWriters, september 2006]

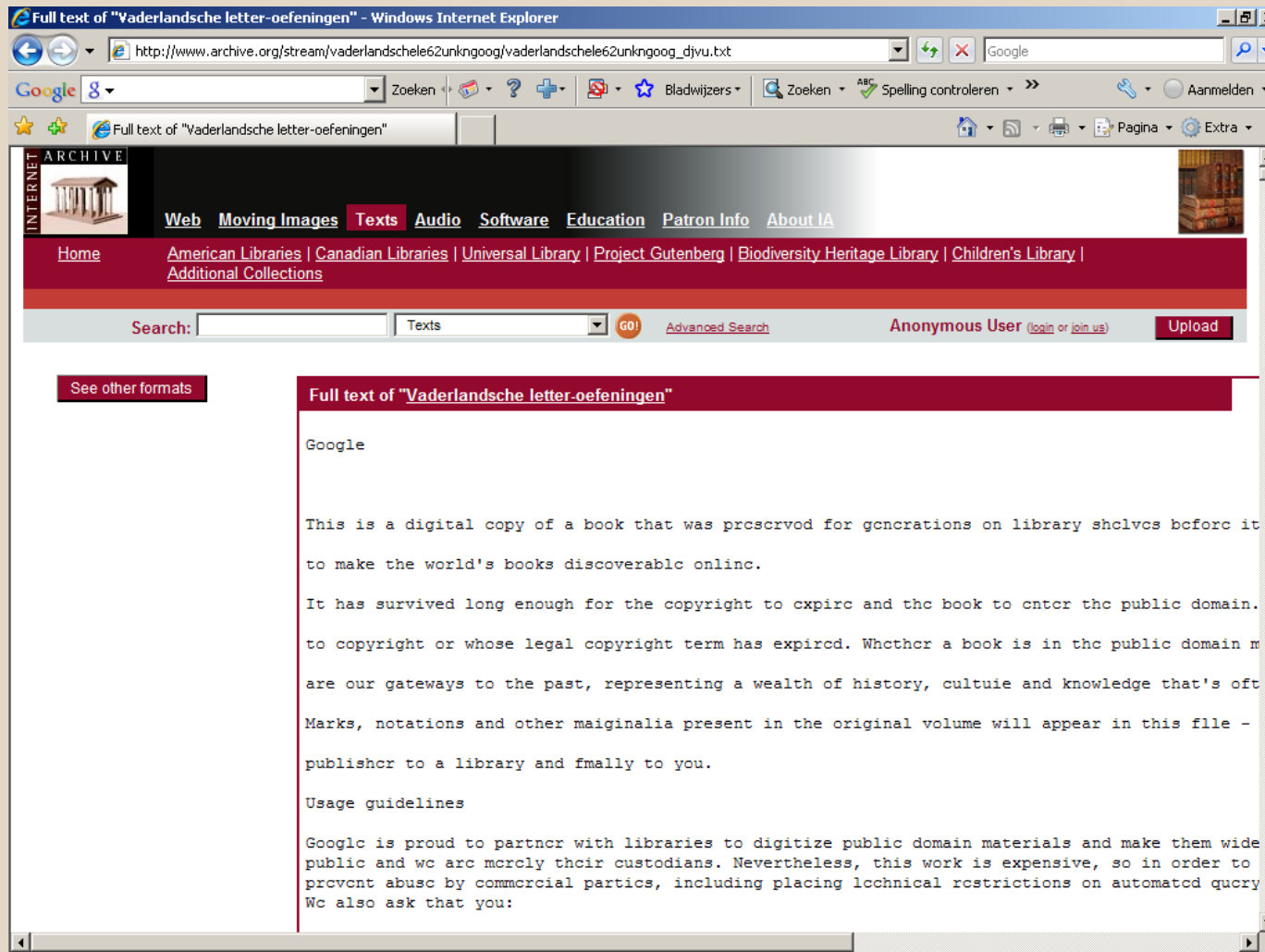
Annotation on "Emilie Flygare Carlen"

Emilie Flygare Carlen : {ww:bio} Informatie over deze schrijfster en over haar werk is te vinden op [www.databasewomenwriters.nl](http://www.databasewomenwriters.nl). [Susanne Parren, WomenWriters, september 2006]

## 6. Rhino Resource Center



## 7. Internet Archive



The screenshot shows a Windows Internet Explorer browser window displaying the Internet Archive website. The address bar shows the URL: [http://www.archive.org/stream/vaderlandschele62unkngoog/vaderlandschele62unkngoog\\_djvu.txt](http://www.archive.org/stream/vaderlandschele62unkngoog/vaderlandschele62unkngoog_djvu.txt). The page title is "Full text of 'Vaderlandsche letter-oefeningen'".

The website header includes the Internet Archive logo and navigation links: [Web](#), [Moving Images](#), [Texts](#), [Audio](#), [Software](#), [Education](#), [Patron Info](#), and [About IA](#). Below this is a red navigation bar with links: [Home](#), [American Libraries](#), [Canadian Libraries](#), [Universal Library](#), [Project Gutenberg](#), [Biodiversity Heritage Library](#), and [Children's Library](#), followed by [Additional Collections](#).

The search bar contains the text "Search:" and "Texts". To the right, it says "Anonymous User" with links for [login](#) or [join us](#), and an "Upload" button.

The main content area has a red header "Full text of 'Vaderlandsche letter-oefeningen'". Below this, there is a "See other formats" button. The text of the document is displayed in a monospaced font:

Google

This is a digital copy of a book that was preserved for generations on library shelves before it to make the world's books discoverable online.

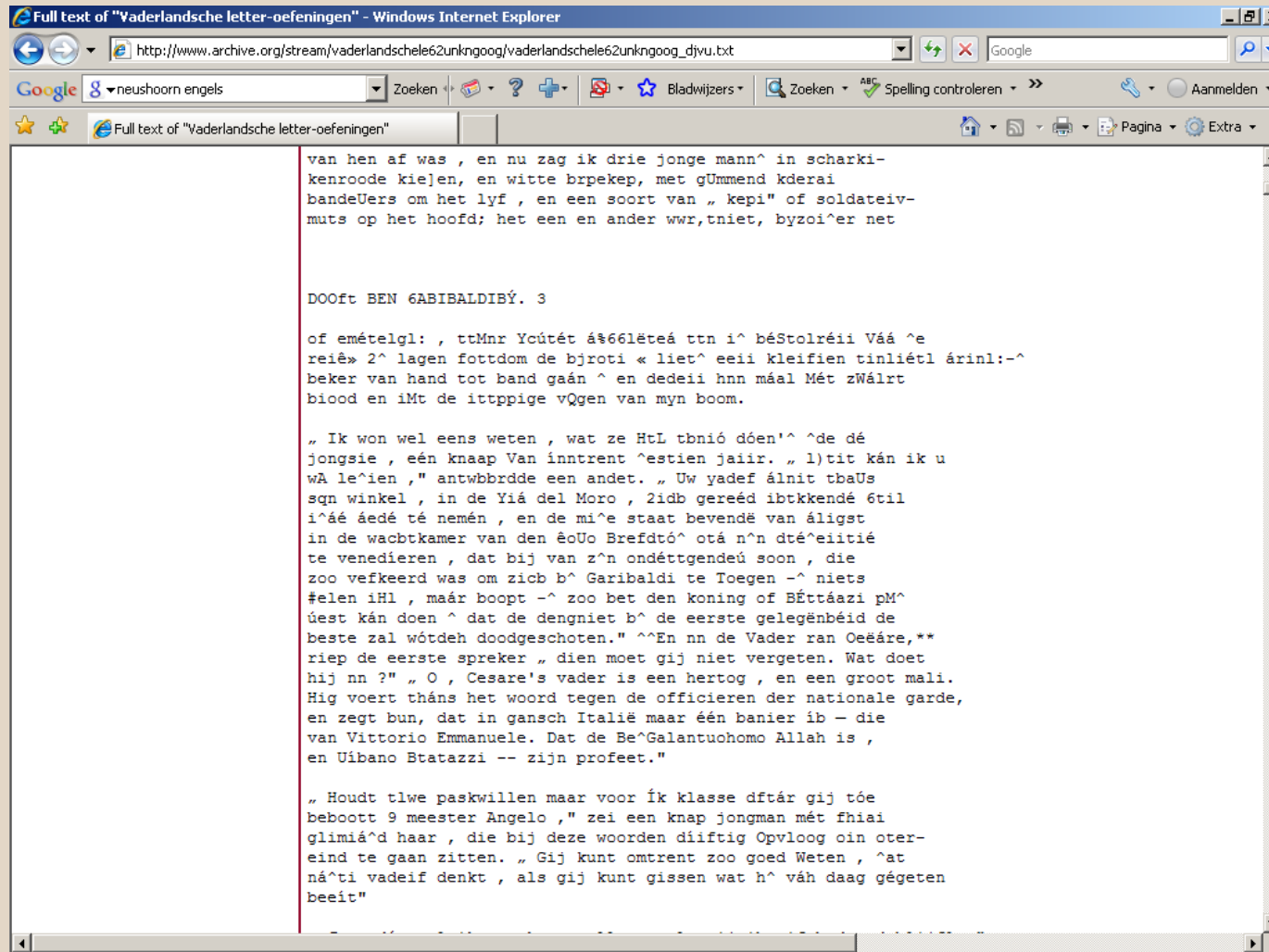
It has survived long enough for the copyright to expire and the book to enter the public domain. to copyright or whose legal copyright term has expired. Whether a book is in the public domain are our gateways to the past, representing a wealth of history, culture and knowledge that's oft

Marks, notations and other marginalia present in the original volume will appear in this file - publisher to a library and finally to you.

Usage guidelines

Google is proud to partner with libraries to digitize public domain materials and make them wide public and we are merely their custodians. Nevertheless, this work is expensive, so in order to prevent abuse by commercial parties, including placing technical restrictions on automated query We also ask that you:

## Quality of OCR...



## Why not leave it to others?

- What about quality?
- What about completeness?
- What about digital preservation?
- What about availability on the long term?
- What about free availability?
- What about copyright?
- What about special collections & more difficult materials?

## Library challenges for mass digitisation

- Make the digitisation process more efficient
  - Automated processes
- Improve quality for mass digitisation
  - Improving Access to Text!
- Transition to a digital library

# Solutions

## Selection

- Avoid scanning twice
  - Central registry of digital masters
  - Bye content
  - Develop portals
- Deal with copyright
- Digitise complete collections

## Digital imaging

- Quality of OCR is most important
- New formats to save storage (JPEG2000)
- Consider .txt as a master...

## Quality assurance

- Not realistic to check quality for all files
- Automatic quality assurance tools
- QA for OCR ignored

## Search & Retrieval

- All text digitisation projects should include OCR
- Find solutions for OCR of historical text
- Find solutions for historical spelling variants

## Storage

- Storage strategy which balances costs, access and preservation
- Alternative file formats to minimize storage costs & increase throughput for delivery and transfer
- Use one file both as master and access file

### Finance

- All costs are now specified
- Division of budget
  - 30 % Staff
  - 10 % Hard- & software
  - 10 % Research & Development
  - 50 % Digitisation, OCR & metadata
- New business models

## Organisation

- All digitisation activities take place in R&D department
  - Involvement of other departments is necessary
- Digitisation activities are all project based
  - Integration with standing organisation is necessary
- Co-orporation is necessary

## Expectations from IMPACT - 1

- Tools to improve the quality of mass digitisation of historical text
  - Image enhancement
  - Segmentation
  - OCR
  - Post correction
  - Language tools

## Expectations from IMPACT - 2

- Share knowledge
  - Quidelines
  - Training
  - Helpdesk
  - Website



**Thank you!**

[Astrid.Verheusen@kb.nl](mailto:Astrid.Verheusen@kb.nl)

