Mass digitization and OCR

Aly Conteh

Collections Digitisation & Strategy Programme Manager
British Library

Polish Digital Libraries 2010

About the British Library

BRITISH LIBRARY

THE BRITISH LIBRARY
Explore the world's knowledge

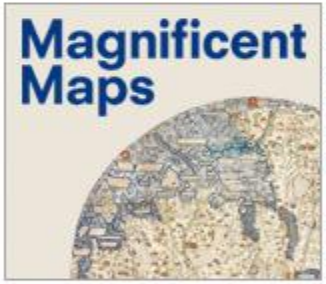
We hold 14 million books, 920,000 journal and newspaper titles, 58 million patents, 3 million sound recordings, and so much more. Start exploring here.

SEARCH [ ] GO

Search tips and advanced searching

- **British Library** 10,000 pages on our main website
- **Online Gallery** 30,000 treasures from our collection
- **Catalogue records** 14 million items in our collections
- **Journal articles** 9 million articles from 20,000 journals

Quick links | What's on | Site highlights | Your library



Exhibition closed

- Opening times, maps
- Reader Registration
- Reading Rooms
- Help for researchers
- Online catalogues
- Information in foreign languages
- For higher education
- For entrepreneurs
- For librarians
- For publishers: legal deposit etc.
- Collection Care
- Press Room
- Contact us

News

17 Sep 2010
2020 Vision launched

16 Sep 2010
New Learning Centre

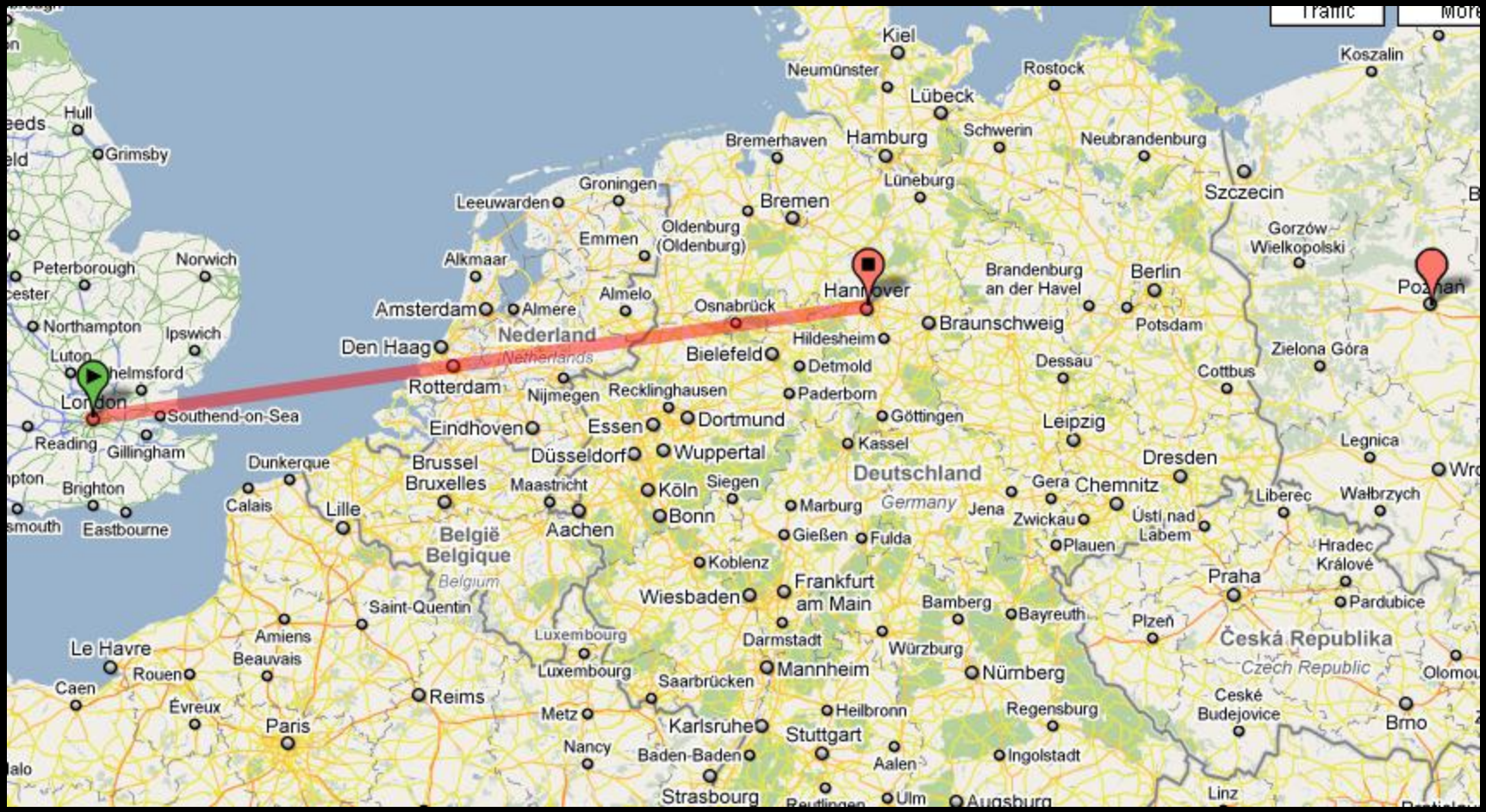
15 Sep 2010
UK SoundMap

15 Sep 2010
Olwyn Hughes Archive

3 Sep 2010
Innovation Season at the British Library

British Library websites [Please choose...]

150 million items



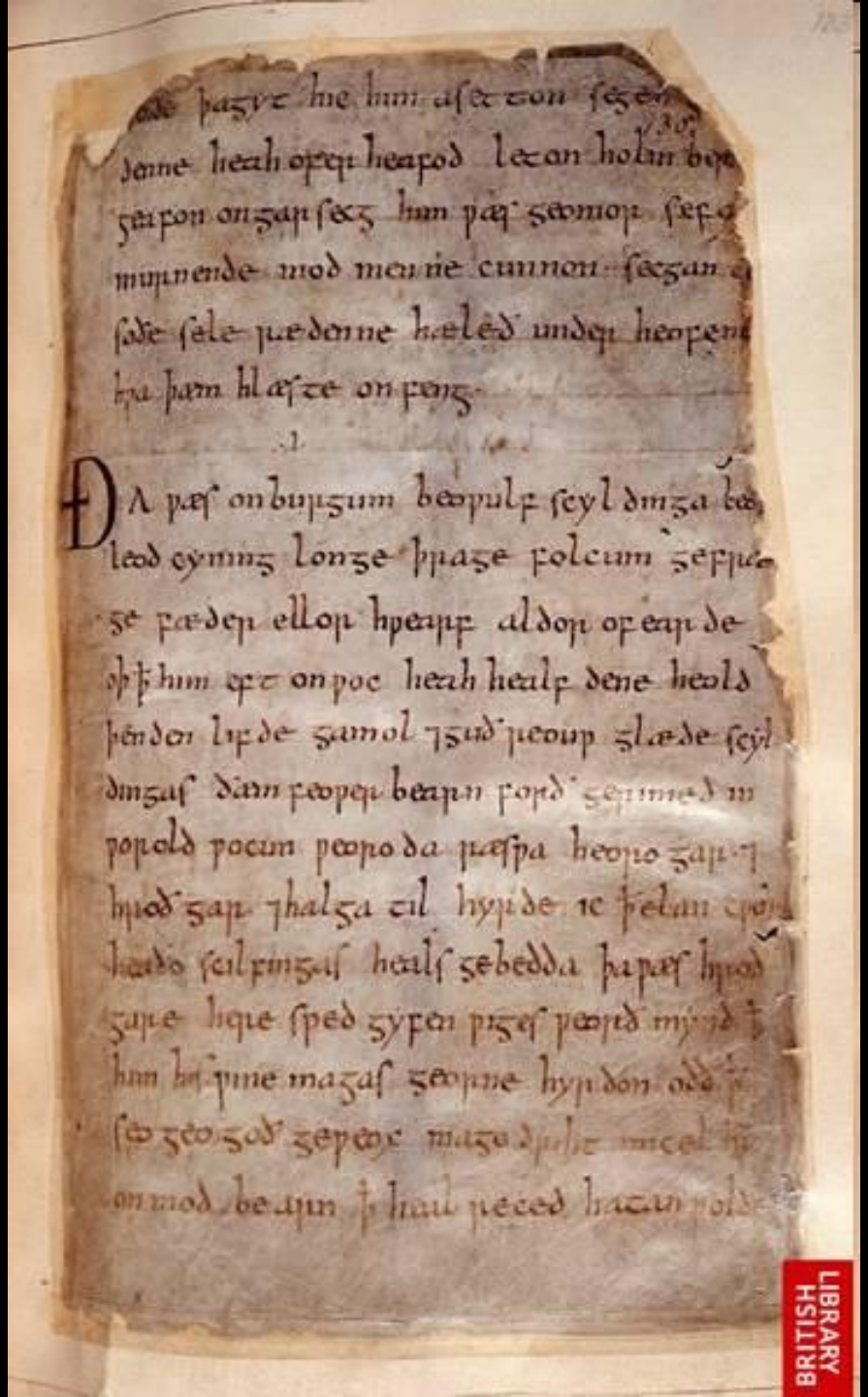
BRITISH LIBRARY

BRITISH LIBRARY

BRITISH LIBRARY

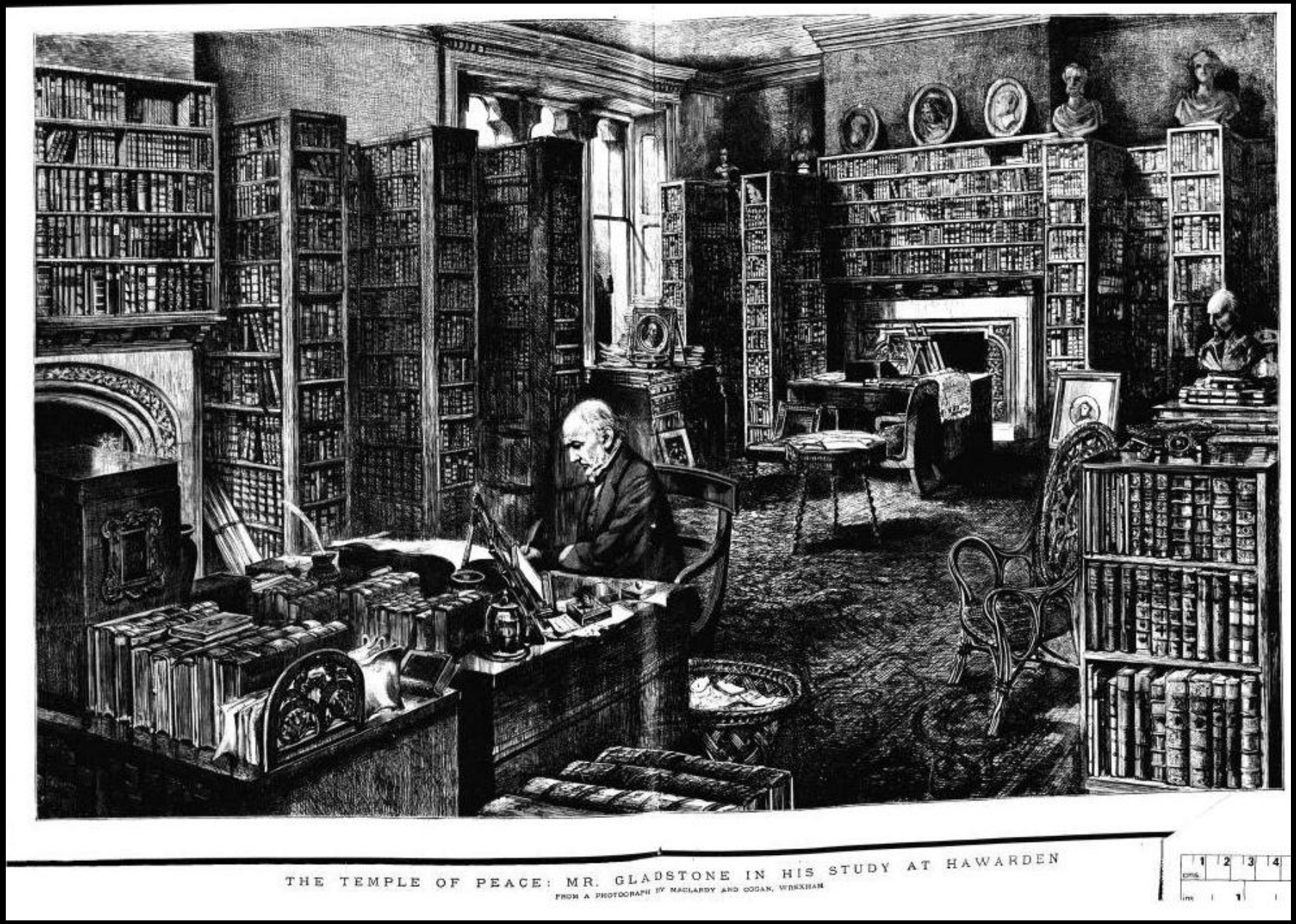
What have we done so far

BRITISH LIBRARY



Beowulf

1994

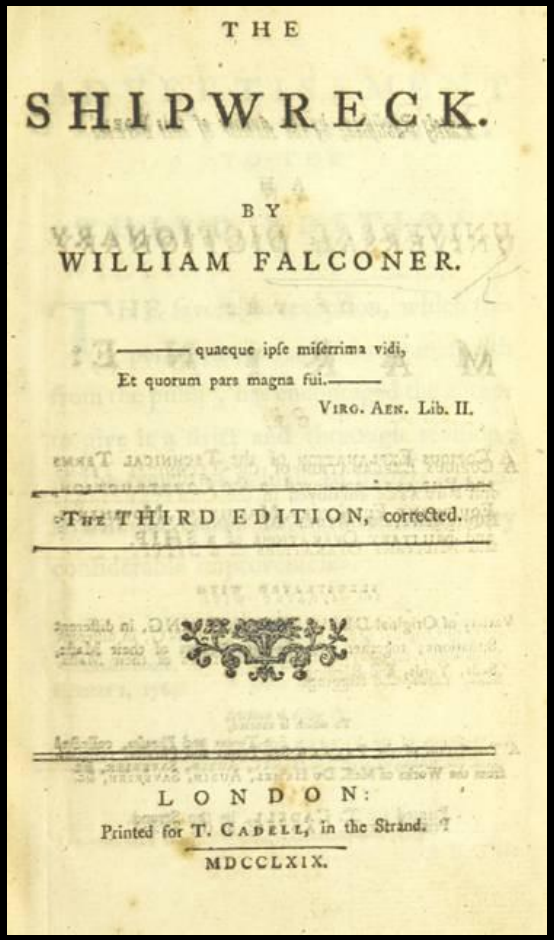


THE TEMPLE OF PEACE: MR. GLADSTONE IN HIS STUDY AT HAWARDEN

FROM A PHOTOGRAPH BY MACLARDY AND OGDAN, WREXHAM

4 million pages pre-1900 Newspapers

BRITISH LIBRARY



THE

SHIPWRECK.

BY

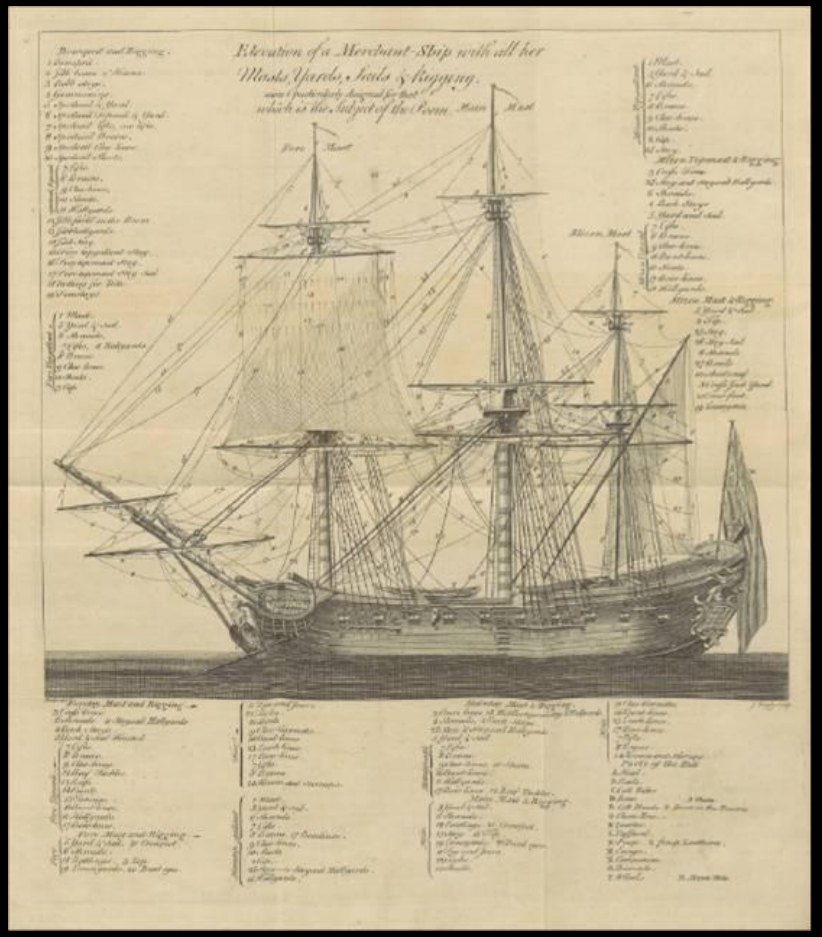
WILLIAM FALCONER.

——— quaeque ipse miserrima vidi,
Et quorum pars magna fui.———

VIRG. AEN. Lib. II.

THE THIRD EDITION, corrected.

LONDON:
Printed for T. CADELL, in the Strand.
MDCCLXIX.



25 million pages 19th Century Books

BRITISH LIBRARY

8,000 hours of Sound Recordings

BRITISH LIBRARY

What challenges do we face with our mass digitisation programme

BRITISH LIBRARY

…the amount of data we generate

BRITISH LIBRARY

…25 million pages of 19th century books

BRITISH LIBRARY

…old standard of uncompressed TIFF files

BRITISH LIBRARY

…creates over 2PB worth of data

BRITISH LIBRARY

…the solution for us was to use the JPEG2000 format instead

BRITISH LIBRARY

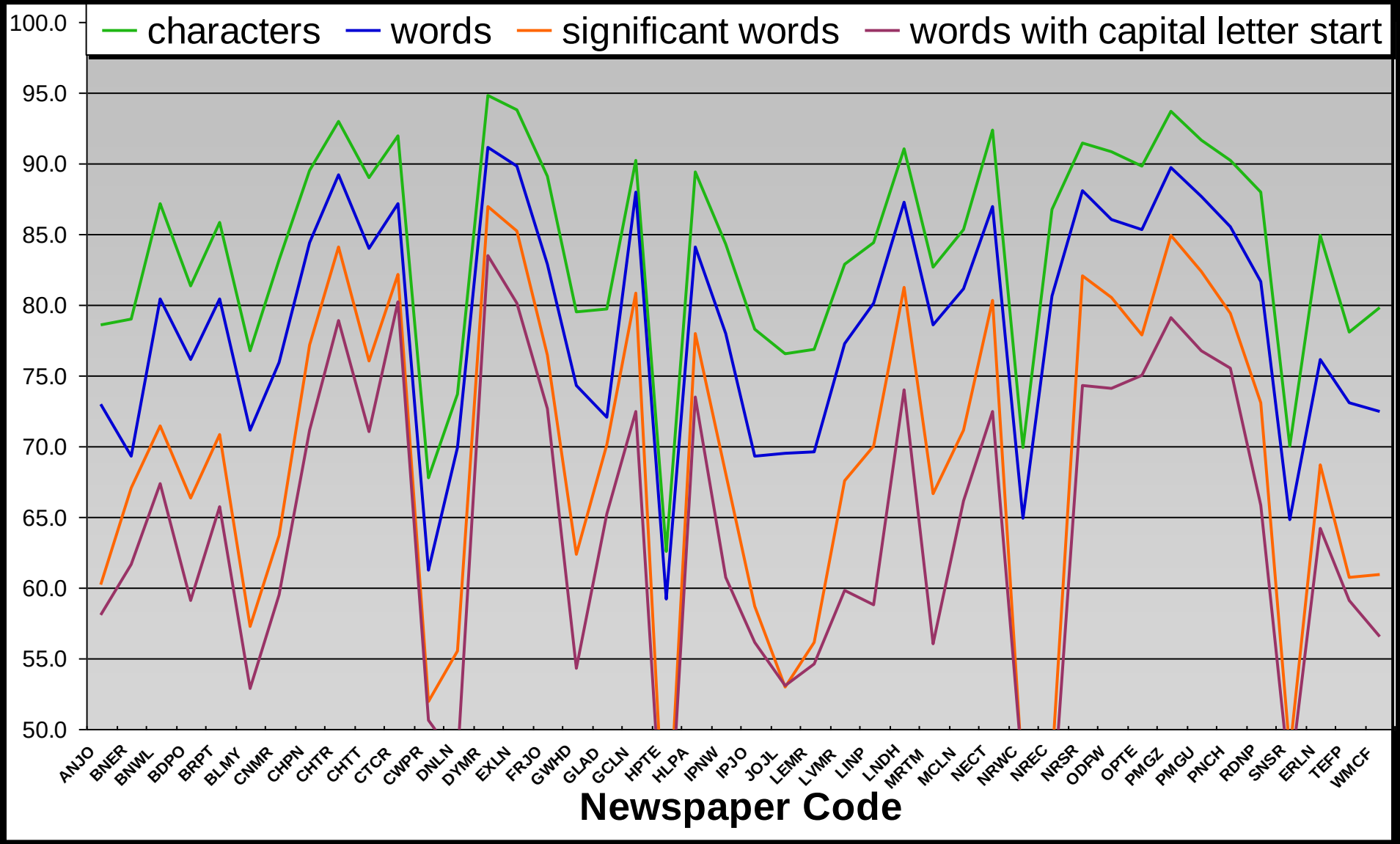
…using lossy! compression,
under 40TB

BRITISH LIBRARY

…is visually lossless compression the answer?

BRITISH LIBRARY

…Need for better text extraction



Measuring OCR accuracy

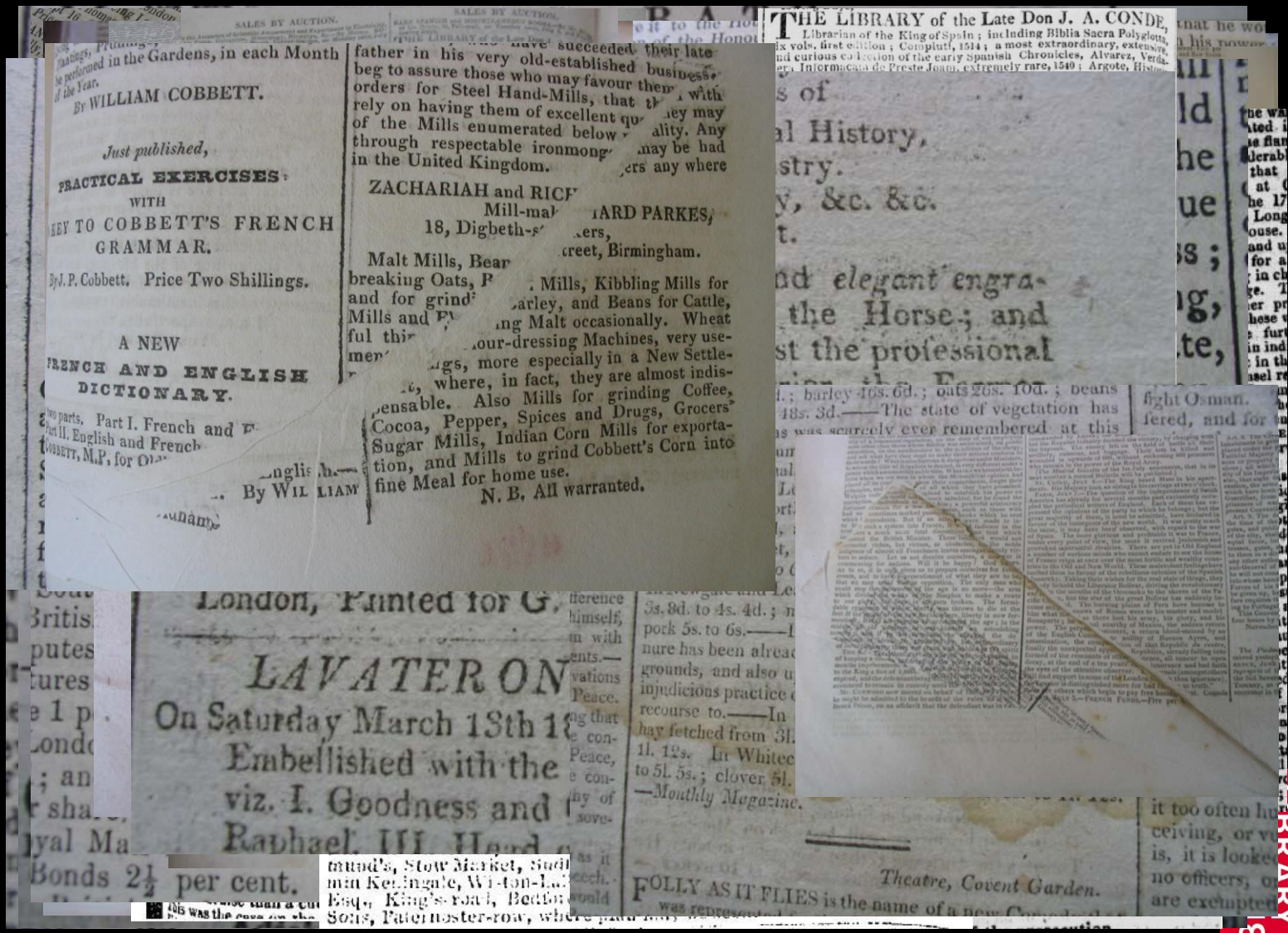
BRITISH LIBRARY

…four issues that affect performance of OCR?

BRITISH LIBRARY

PHYSICAL CHARACTERISTICS OF SOURCE MATERIAL

- Bleed through
- Stains
- Tight binding
- Holes/tears
- Printer errors
- Stamps
- Creases
- Dirt
- Animals
- Inconsistent inking
- Repairs
- Paper quality
- Lamination



SINGULAR AND SERIOUS ACCIDENT.—On Wednesday [illegible] noon Mr. Charles Wyber, of the Borough-road, went into the Fleet Prison to visit a friend, and joined a party in the Coffee-room, who entered into the foolish amusement of tossing up a penny-piece to the top of the room, and catching the same in the mouth upon its descent to the floor. Mr. Wyber was considered a perfect adept at this game; but the penny-piece at last found its way into the throat, where it remained for upwards of half an hour. A Surgeon tried to force it upwards, but being unable to do so, he contrived to move it downwards into the stomach. Mr. Wyber was comparatively much relieved by the penny-piece being removed out of the throat, and was enabled, in the evening, to be carried to his house in a hackney-coach.

I SINGULAR AND SERIOUS ACCIDENT.- -Oll WeU Iwtje e
noon Mr. Charles %Vyber, of the Borougll-roadt, V
Fleet Prison to visit a friend, and joined a party III
room, who entered into the foolish a seincitllt of l
pcnny-p)icce to the top of thle room, andl eatchivug
tle mothli upon its descent to the lloor. Mr.) L
sidered a perfect adept at this game bot time Ile"
last found its way into thle throat, where it V'tOtdt
wards of half an hour. A Surgeon tried tv folCe
but being ulnable to do so, lie contrived to mDOVC t
into the stomach. Mr. Wyyber was commmpariatLvl
lieved by the penny-piece being riemoved out ot the
was enabiled, in the evening, to be carried to
hackney-coach.

The numbers

There are 152 words

61 words are incorrectly identified

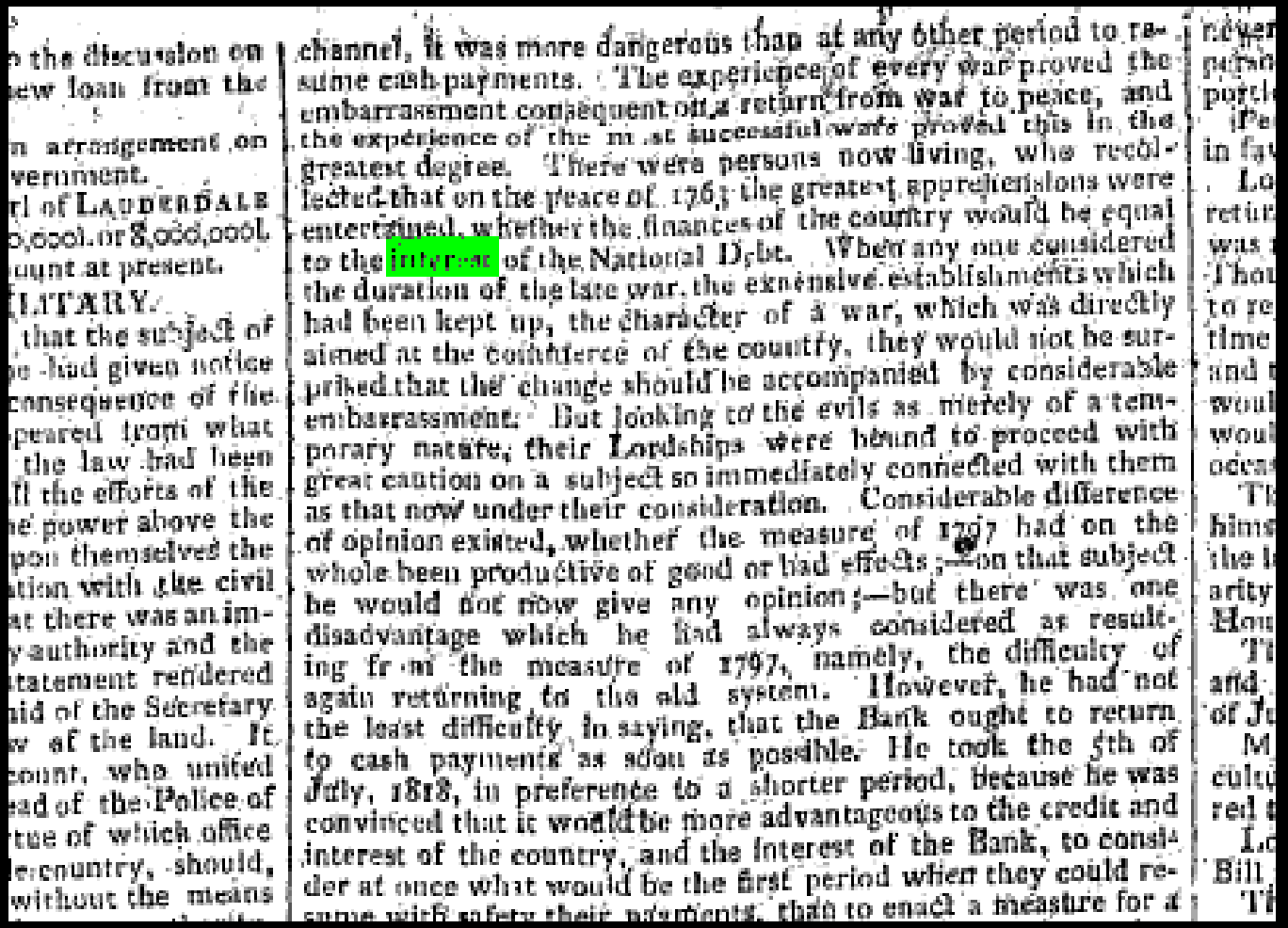
Giving us 60% word accuracy

But all words are not equal

BRITISH LIBRARY

SINGULAR AND SERIOUS ACCIDENT.—On Wednesday [illegible] noon Mr. Charles Wyber, of the Borough-road, went into the Fleet Prison to visit a friend, and joined a party in the Coffee-room, who entered into the foolish amusement of tossing up a penny-piece to the top of the room, and catching the same in the mouth upon its descent to the floor. Mr. Wyber was considered a perfect adept at this game; but the penny-piece at last found its way into the throat, where it remained for upwards of half an hour. A Surgeon tried to force it upwards, but being unable to do so, he contrived to move it downwards into the stomach. Mr. Wyber was comparatively much relieved by the penny-piece being removed out of the throat, and was enabled, in the evening, to be carried to his house in a hackney-coach.

They had the internet in 1816 !

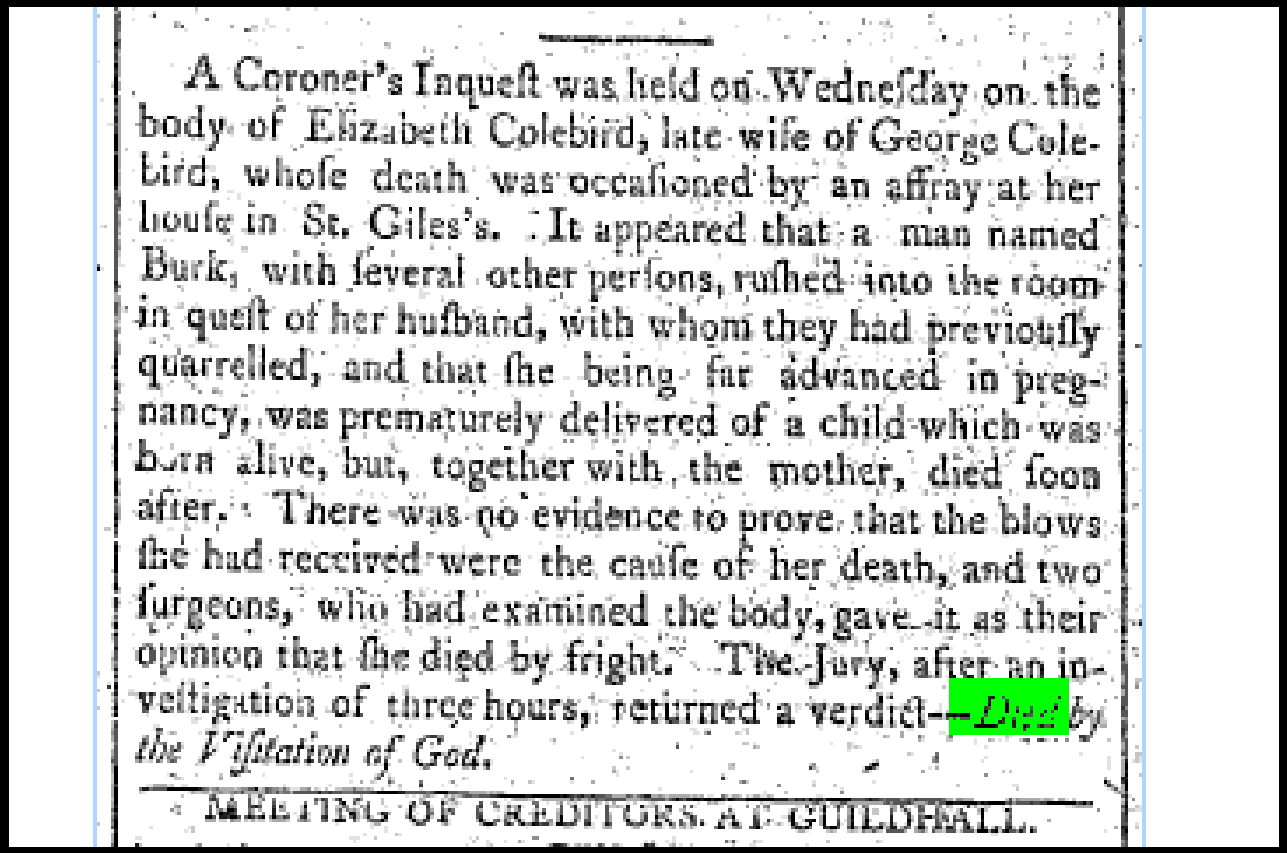


channel, it was more dangerous than at any other period to re-
sume cash-payments. The experience of every war proved the
embarrassment consequent on a return from war to peace, and
the experience of the most successful wars proved this in the
greatest degree. There were persons now living, who recol-
lected that on the peace of 1763, the greatest apprehensions were
entertained, whether the finances of the country would be equal
to the interest of the National Debt. When any one considered
the duration of the late war, the extensive establishments which
had been kept up, the character of a war, which was directly
aimed at the commerce of the country, they would not be sur-
prised that the change should be accompanied by considerable
embarrassment. But looking to the evils as merely of a tem-
porary nature, their Lordships were bound to proceed with
great caution on a subject so immediately connected with them
as that now under their consideration. Considerable difference
of opinion existed, whether the measure of 1797 had on the
whole been productive of good or bad effects;—on that subject
he would not now give any opinion;—but there was one
disadvantage which he had always considered as result-
ing from the measure of 1797, namely, the difficulty of
again returning to the old system. However, he had not
the least difficulty in saying, that the Bank ought to return
to cash payments as soon as possible. He took the 5th of
July, 1818, in preference to a shorter period, because he was
convinced that it would be more advantageous to the credit and
interest of the country, and the interest of the Bank, to consi-
der at once what would be the first period when they could re-
sume with safety their payments, than to enact a measure for a

The Morning Chronicle (London, England), Saturday, May 18, 1816; Issue 14678

BRITISH LIBRARY

and DVD in 1803!



A Coroner's Inquest was held on Wednesday on the body of Elizabeth Colebird, late wife of George Colebird, whose death was occasioned by an affray at her house in St. Giles's. It appeared that a man named Burk, with several other persons, rushed into the room in quest of her husband, with whom they had previously quarrelled, and that she being far advanced in pregnancy, was prematurely delivered of a child which was born alive, but, together with the mother, died soon after. There was no evidence to prove that the blows she had received were the cause of her death, and two surgeons, who had examined the body, gave it as their opinion that she died by fright. The Jury, after an investigation of three hours, returned a verdict—*Died by the Visitation of God.*

MEETING OF CREDITORS AT GUILDHALL.

The Morning Chronicle (London, England), Friday, June 10, 1803; Issue 10625

BRITISH LIBRARY

…Summary of issues

Geometric distortions lead to text being missed or incorrectly identified

Quality of source material has a notable impact on accuracy levels

The need to focus on significant words

False positives can be introduced by using modern lexicons

And there are others……..

BRITISH LIBRARY

How are we addressing the OCR issues?

BRITISH LIBRARY

IMPACT

Innovating OCR software and language technology

Sharing expertise and building capacity across Europe

Ensuring that tools and services will be sustained after the end of the project

Improving Access to Text
IMPACT

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

BRITISH LIBRARY

26 partners: Libraries, Research Institutes, Industry Partners

BRITISH LIBRARY

Facts and figures

- Project supported by the European Community under the FP7 ICT Work Programme.
- coordinated by the National Library of the Netherlands (KB)
- EU funding: € 11 500 000
- Start date: 1 January 2008
- Duration: 48 months
- From 2012: sustainable Centre of Competence
- Web site: www.impact-project.eu

BRITISH LIBRARY

…what will IMPACT deliver

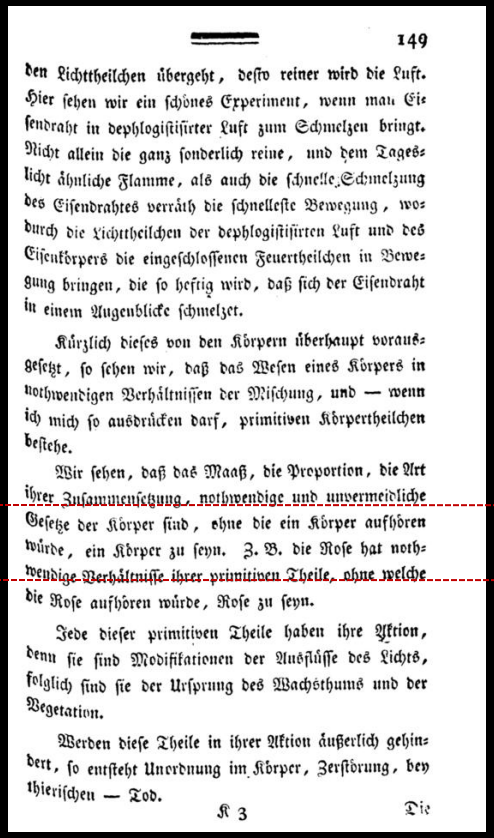
BRITISH LIBRARY

…advancement in the state of the art in image enhancement, segmentation and binarisation

BRITISH LIBRARY

Straightening the lines

Original image



149

den Lichttheilchen übergeht, desto reiner wird die Luft. Hier sehen wir ein schönes Experiment, wenn man Eisendraht in dephlogistisirter Luft zum Schmelzen bringt. Nicht allein die ganz sonderlich reine, und dem Tageslicht ähnliche Flamme, als auch die schnelle Schmelzung des Eisendrahtes verräth die schnelleste Bewegung, wodurch die Lichttheilchen der dephlogistisirten Luft und des Eisenkörpers die eingeschlossenen Feuertheilchen in Bewegung bringen, die so heftig wird, daß sich der Eisendraht in einem Augenblicke schmelzet.

Kürzlich dieses von den Körpern überhaupt vorausgesetzt, so sehen wir, daß das Wesen eines Körpers in nothwendigen Verhältnissen der Mischung, und — wenn ich mich so ausdrücken darf, primitiven Körpertheilchen bestehe.

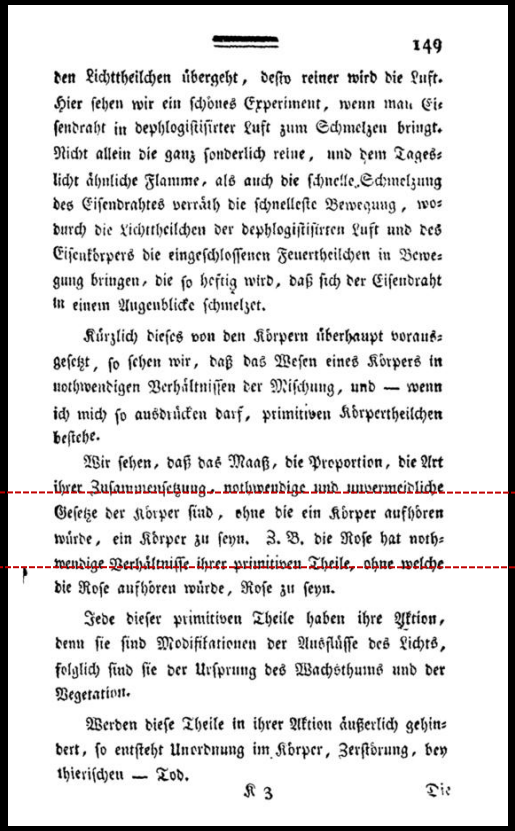
Wir sehen, daß das Maaß, die Proportion, die Art ihrer Zusammensetzung, nothwendige und unvermeidliche Gesetze der Körper sind, ohne die ein Körper aufhören würde, ein Körper zu seyn. Z. B. die Rose hat nothwendige Verhältnisse ihrer primitiven Theile, ohne welche die Rose aufhören würde, Rose zu seyn.

Jede dieser primitiven Theile haben ihre Aktion, denn sie sind Modifikationen der Ausflüsse des Lichts, folglich sind sie der Ursprung des Wachsthums und der Vegetation.

Werden diese Theile in ihrer Aktion äußerlich gehindert, so entsteht Unordnung im Körper, Zerstörung, bey thierischen — Tod.

K 3 Die

Dewarping v.1.1



149

den Lichttheilchen übergeht, desto reiner wird die Luft. Hier sehen wir ein schönes Experiment, wenn man Eisendraht in dephlogistisirter Luft zum Schmelzen bringt. Nicht allein die ganz sonderlich reine, und dem Tageslicht ähnliche Flamme, als auch die schnelle Schmelzung des Eisendrahtes verräth die schnelleste Bewegung, wodurch die Lichttheilchen der dephlogistisirten Luft und des Eisenkörpers die eingeschlossenen Feuertheilchen in Bewegung bringen, die so heftig wird, daß sich der Eisendraht in einem Augenblicke schmelzet.

Kürzlich dieses von den Körpern überhaupt vorausgesetzt, so sehen wir, daß das Wesen eines Körpers in nothwendigen Verhältnissen der Mischung, und — wenn ich mich so ausdrücken darf, primitiven Körpertheilchen bestehe.

Wir sehen, daß das Maaß, die Proportion, die Art ihrer Zusammensetzung, nothwendige und unvermeidliche Gesetze der Körper sind, ohne die ein Körper aufhören würde, ein Körper zu seyn. Z. B. die Rose hat nothwendige Verhältnisse ihrer primitiven Theile, ohne welche die Rose aufhören würde, Rose zu seyn.

Jede dieser primitiven Theile haben ihre Aktion, denn sie sind Modifikationen der Ausflüsse des Lichts, folglich sind sie der Ursprung des Wachsthums und der Vegetation.

Werden diese Theile in ihrer Aktion äußerlich gehindert, so entsteht Unordnung im Körper, Zerstörung, bey thierischen — Tod.

K 3 Die

…improvements with OCR products

BRITISH LIBRARY

…development of language tools

BRITISH LIBRARY

Building Named Entities resources (Source Material 34,500 pages)

Dictionary of National Biography

Alumni Oxonienses: the members of the University of Oxford 1500-1714

Alumni Oxonienses: the members of the University of Oxford, 1715-1886

Wilson's Mercantile Directory of Great Britain and Ireland

Cassell’s Gazetteer of Great Britain and Ireland

BRITISH LIBRARY

…tools to build capability

BRITISH LIBRARY

…and a research dataset

BRITISH LIBRARY

Thank You

Improving Access to Text
IMPACT

http://www.impact–project.eu/

Twitter: @impactocr

http://impactocr.wordpress.com/

www.bl.uk

aly.conteh@bl.uk

twitter: @aconteh

BRITISH LIBRARY