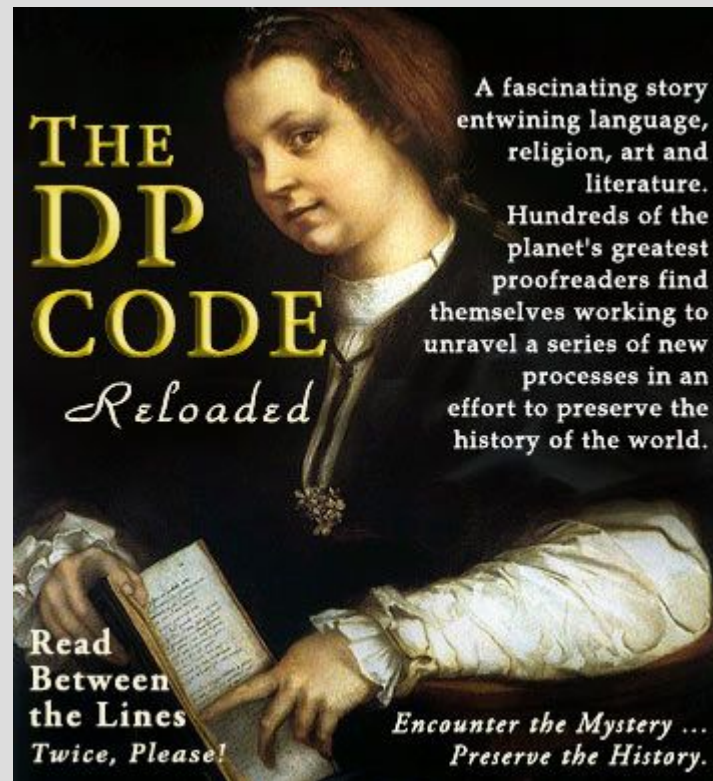
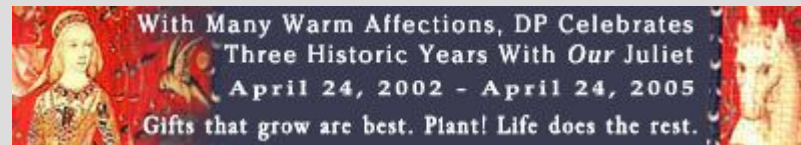


The Distributed Proofreaders Project



The Distributed Proofreaders Project

Juliet Sutherland
www.pgdp.net



What We Do

- Convert public domain content from printed formats into digital text editions.
- Distribute the work across many volunteers at both the page and project level

DP Celebrates the Birthday of the
Immortal Bard, William Shakespeare

*Not marble nor the gilded monuments
Of princes shall outlive this powerful rime.*



Distributed Proofreaders Is

- Completely volunteer
 - Worldwide participation
- Independent of commercial, corporate or academic affiliation
- Cooperatively and ideologically allied with Project Gutenberg
- Unhindered by any editorial or censoring policy
- Unrestricted by the limits of a fixed location



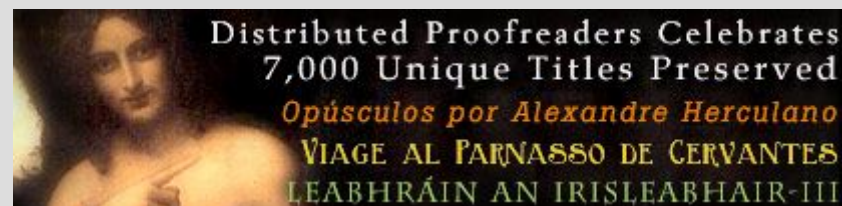
Why?

- Accurate, versatile text editions are in great demand
 - Can copy and paste
 - Can be used by vision-impaired people
 - Down-sizes well for use on small viewers like PDAs
 - Allows for the potentially rich uses of structured text
- We preserve material for posterity and make it freely available in the present



The DP Community

- Mostly from English speaking countries
- Graduate students to retirees
- 400-500 individuals login per 24 hours
- ~1000 individuals login per week
- Active forums and chat room
- As international as we can make it on a site that only uses English
- About 20% of our production is LOTE

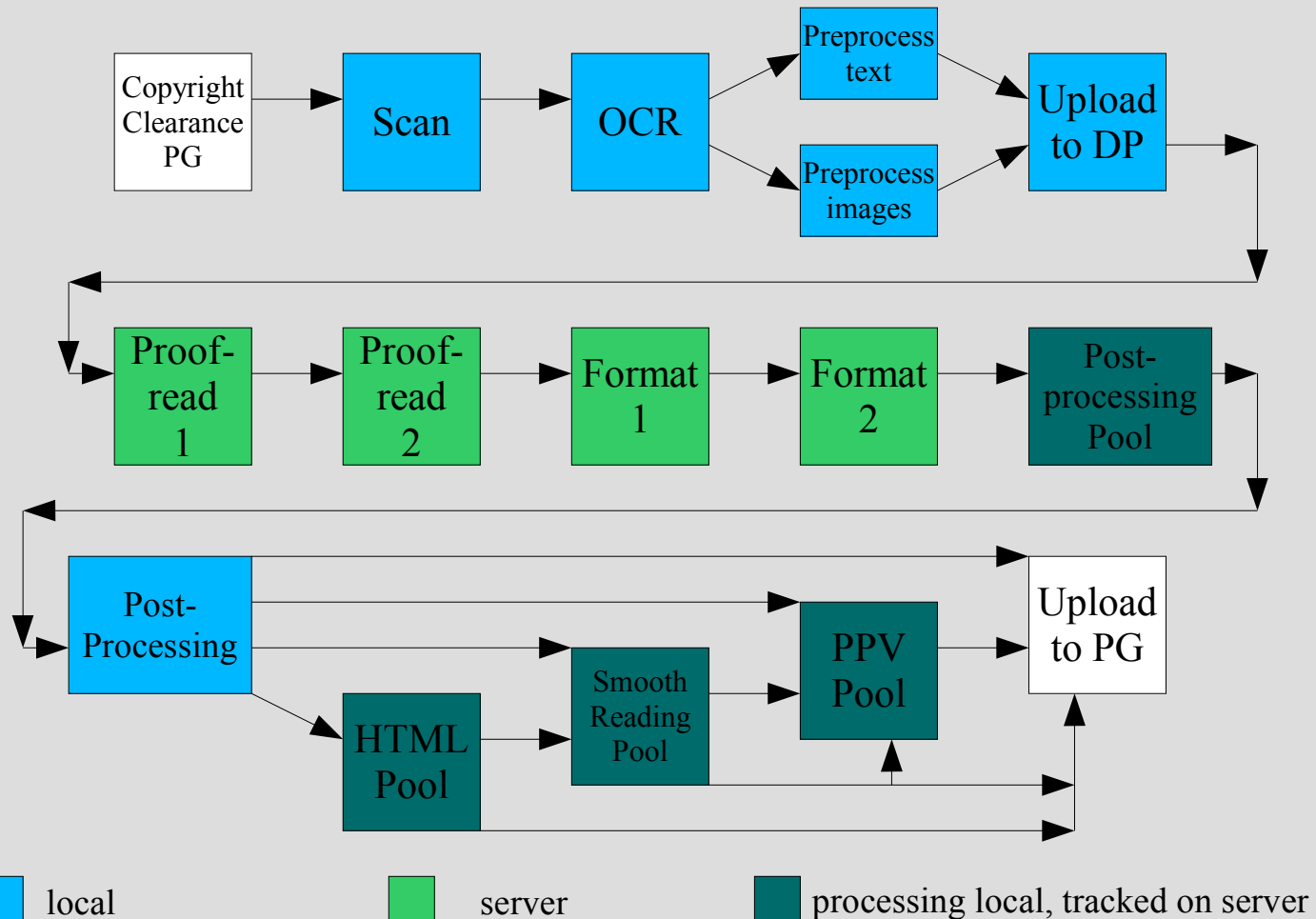


Numbers

- Unique titles produced: 7100+
- Projects in process: 3600+
- Pages proofed 2/24/03 to 5/31/05: 5 million
- Pages proofed since 6/3/05: ~6K/day
- Accounts ever: ~33,500
- Languages in progress: ~15
- Languages ever: ~40
- Forum posts/day: ~135

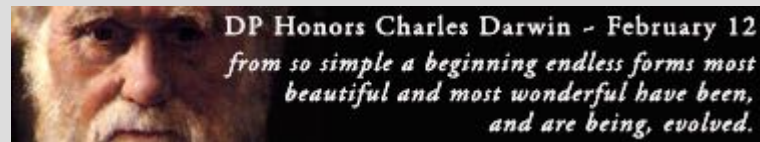


Current DP Process



Scans

- Mostly provided by volunteers using flatbed, inexpensive scanners
- Some are sourced from various image archives including: MBP, LoC, Gallica, Canadiana, Toronto collection, MoA, and many others
- Some from 2 high-speed, sheet-fed scanners
 - Books scanned this way are destroyed in the process



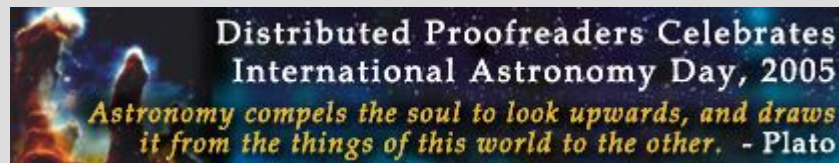
OCR

- Done locally, volunteers provide their own software
- Most use ABBYY Finereader Pro 5.0-7.0
- Some use other products
- OCR Pool for those without access to OCR software



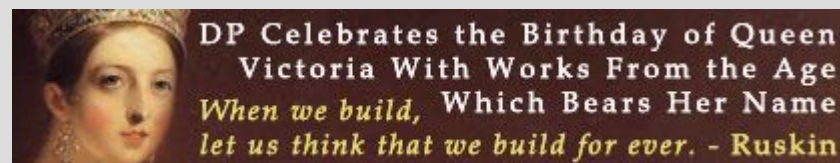
Preprocessing Text

- Fix the most common OCR problems
 - Spaced punctuation
 - 1-l-i, O-0 confusion
 - Common scannos: tbe, '11
- Extract some formatting information
- Volunteer developed guiprep is our primary tool
- Some volunteers use other ad hoc scripts and programs



Preprocessing Images

- Downsize for proofing if necessary
 - B&W, 300 dpi, 1000 pixels wide
- We try to keep the scans to ~50K/page, and definitely under 100K
- This step is optional



Proofing Rounds

- Make the text, at the character level, match the scanned image
- Find and correct
 - Scannos
 - Stealth scannos: arid/and, he/be
- Spell check
- DP Custom Mono font



Formatting Rounds

- Add markup as necessary
 - Chapters, sections, etc
 - Italics, bold, small caps
 - Poetry, block quotes, tables
 - LaTeX for math
- Home grown markup “tags”
 - Italics: `<i></i>`
 - Block quotes: `/# #/`
 - Footnotes: `[Footnote 1: text.]`



Horizontal User Interface

The screenshot shows a Microsoft Internet Explorer window with the title "[P1] Elizabethan sonnet-cycles: Phillis and Licia - Proofreading Interface - Microsoft Internet Explorer provided by Verizon On". The browser's menu bar includes File, Edit, View, Favorites, Tools, and Help. The main content area displays a sonnet titled "PHILLIS 15" with the following text:

PHILLIS 15

II

You sacred sea-nymphs pleasantly disporting
Amidst this wat'ry world, where now I sail;
If ever love, or lovers sad reporting,
Had power sweet tears from your fair eyes to
hail;
And you, more gentle-hearted than the rest,
Under the northern noon-stead sweetly stream-

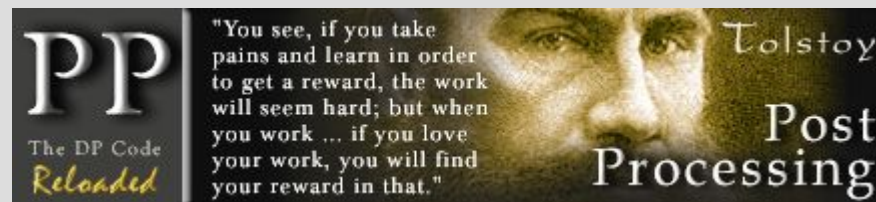
Below the text, there is a scrollable area containing a plain-text version of the sonnet with markup characters like ^ and <. At the bottom of the interface, there are several buttons: "Save as 'In Progress'", "Save as 'Done' & Proofread Next Page", "Save as 'Done'", "Stop Proofreading", "Switch to Vertical", "Return Page to Round", "Report Bad Page", and "Spell Check". Below these buttons, it says "Page: 027 View: [Project Comments](#) | [Image](#)" and "Image Resize: -25% +25% Original". The browser's status bar at the very bottom shows a toolbar with icons for text formatting (A, E, I, O, U, +, -) and a "HELP----> ?" link. At the bottom of the page, there is a footer with "Markup shortcuts: [Greek:] | [Sidnote:] | [Illustration:] | * | [] | [Footnote #:] | /* */ | /# #/ | * * * * * | [Blank Page]" and "Reference Information: [[Proofreading](#) and [Formatting](#) Guidelines] Proofreading Diagrams: [[High Res](#)] [[Medium Res](#)] [[Low Res](#)]"

Vertical User Interface

The screenshot displays a web browser window with the title "[P1] Elizabethan sonnet-cycles: Phillis and Licia - Proofreading Interface - Microsoft Internet Explorer provided by Verizon On". The browser's address bar shows "File Edit View Favorites Tools Help". The main content area is split into two vertical panels. The left panel shows a sonnet titled "PHILLIS 15" with the text: "You sacred sea-nymphs pleasantly disporting / Amidst this wat'ry world, where now I sail; / If ever love, or lovers sad reporting, / Had power sweet tears from your fair eyes to hail; / And you, more gentle-hearted than the rest, / Under the northern noon-stead sweetly streaming, / Lend those moist riches of your crystal crest, / To quench the flames from my heart's Etna streaming; / And thou, kind Triton, in thy trumpet relish / The ruthful accents of my discontent, / That midst this travel desolate and hellish, / Some gentle wind that listens my lament / May prattle in the north in Phillis' ears: / "Where Phillis wants, Damon consumes in tears." The right panel shows the same text in a vertical layout, with the title "PHILLIS 15" at the top. Below the text are several buttons: "Save as 'In Progress'", "Save as 'Done' & Proofread Next Page", "Save as 'Done'", "Stop Proofreading", "Switch to Horizontal", "Return Page to Round", "Report Bad Page", "Spell Check". Below the buttons, it says "Page: 027 View: [Project Comments](#) | [Image](#)" and "Image Resize: -25% +25% Original". At the bottom of the browser window, there is a toolbar with icons for font size, bold, italic, underline, and a "HELP----> ?" button. Below the toolbar, there is a footer with markup shortcuts and reference information: "Markup shortcuts: [Greek:] | [Sidenote:] | [Illustration:] | * | [] | [Footnote #:] | /* */ | /# #/ | * * * * * | [Blank Page] Reference Information: [[Proofreading](#) and [Formatting](#) Guidelines] Proofreading Diagrams: [[High Res](#)] [[Medium Res](#)] [[Low Res](#)]"

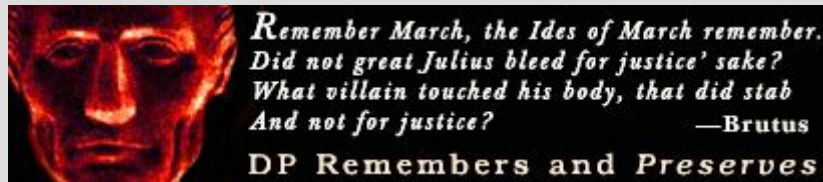
Post-processing

- Text files are concatenated and zipped
- Image files are zipped
- Project checked out by a volunteer
- Work done locally
- Final product either
 - To DP for further processing
 - Directly to PG



Post-Processing-cont.

- Global spell check
- Global scanno/stealth scanno check
- Consistent formatting
- Remove page separators
- Generate HTML, other formats if required:
UTF-8, pdf, etc.



Post-processing Tools

- Can be done with text editor, spell checker, and image viewer/processor, gutcheck
- Specialized tools make the job quicker
 - Guiguts (perl)
 - Gut-axe, Gut-hammer, etc (Basic)
 - pg2html

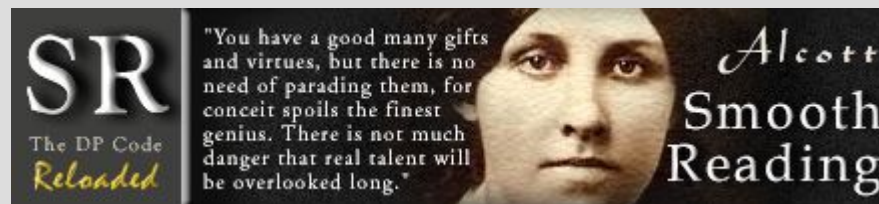
Distributed Proofreaders Celebrates
WORLD POETRY DAY

*O wild and loose to my soul—
O wondrous singer!*



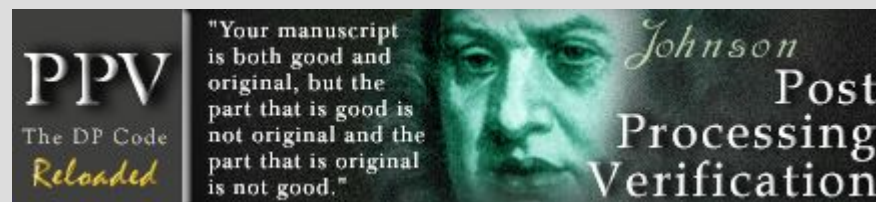
Smooth Reading

- Completed projects are made available in a pool for specified periods of time
- Volunteers read for pleasure, noting anything unusual or jarring
- Text or html version
- Anyone can download, DP account required to upload
- A final relaxed read through is effective at catching odd errors



Post-Processing Verification

- Mentoring process for new post-processors
- Quality check
- Direct posting allowed to PG
- Role granted based on consensus of existing PPVers.



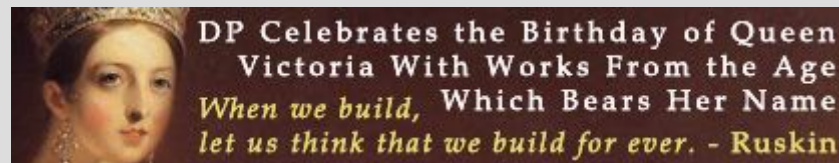
Server Software

- Linux
- Apache
- php
- MySQL
- phpMyAdmin
- phpBB
- CVS
- Sourceforge



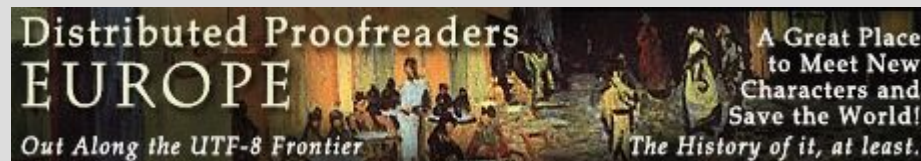
Server Details

- Production Server
 - Hosted at The Planet in Houston
 - Dual CPU
 - Mirrored 80G disk
- Test Server
 - Provided and hosted by The Internet Archive



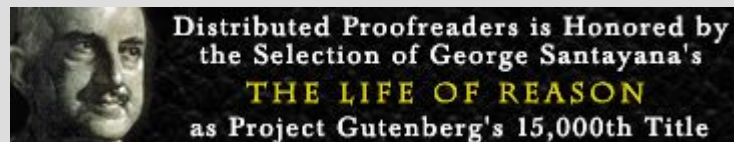
Distributed Proofreaders Europe

- Hosted by Project Rastko in Belgrade, Serbia
- Started in January, 2004
- UTF-8 based: works with Cyrillic and non-Western alphabets
- More localized due to translation of site interface into many languages
- Public domain rules are life+50 years
- About 1/10th the size of PGDP
- www.dp.rastko.net



Project Gutenberg

- PGDP's fiscal ally
- Provides copyright clearance services
- Provides archive and distribution of finished ebooks
- www.gutenberg.org



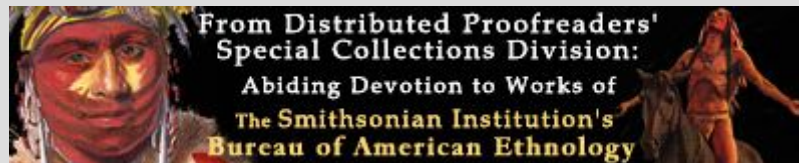
Other Archives

- Gradually building relationships with other archives
 - Library of Congress
 - Posner Memorial Collection
 - The Internet Archive
 - Inquires from many others



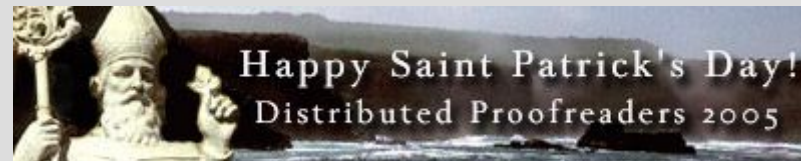
Some Large Projects

- Slave Narratives (30+ parts)
- History of the Philippines (22/50 vols so far)
- Copyright clearances
- Cornell Math Library
- Shakespeare in French
- Periodicals
 - Punch
 - Scientific American
 - As Farpas
 - De Aarde en haar volken



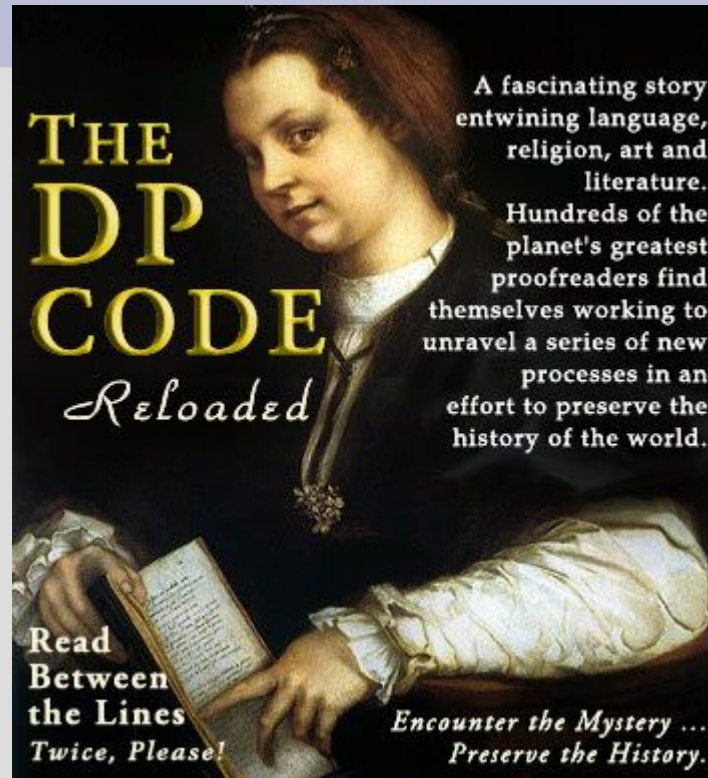
Budget

- What budget?
- Dollars from PG: \$3-10K/year
 - Server, server hosting, high speed scanner setups and maintenance
- In kind donations: Internet Archive
- Volunteers: time, material, tools
 - Paid my own way here
- Could expand with a financial inflow but not hindered by a present hunger



Coase's Penguins

- *Coase's Penguins, or Linux and the Nature of the Firm*, by Yochai Benkler, 2002
 - Abstract at <http://www.benkler.org/CoasesPenguin.html>
- Commons based peer production
- Low barrier to entry, with room for growth
- Small & large chunks available
- Integration of chunks is key
- Issues with over-estimators



**THE
DP
CODE**
Reloaded

A fascinating story
entwining language,
religion, art and
literature.
Hundreds of the
planet's greatest
proofreaders find
themselves working to
unravel a series of new
processes in an
effort to preserve the
history of the world.

**Read
Between
the Lines**
Twice, Please!

*Encounter the Mystery ...
Preserve the History.*

Banners by Thierry Alberto