Breakdowns in Machine Reading:

Attempting to De-privilege Modern English Print with the Power of Supercomputing and the DH Dashboard

The State(s) of the OCR Problem

Computer Science, circa 1990

- The OCR problem is solved! We can machine-read text with 99% accuracy.

The Humanities, circa 2010

- Wait, what? We still can't reliably perform digital searches on texts printed before 1800! And what about OCR for languages other than modern English?

little better herein, and tell he do it, the opinion of *E. P.* ſtands good, for all his pretended ſelf authoritative aſſertion, and for the cloſe of this Doctors *fifth Section*, *E. P.* is not ſo poſſeſſed with a Stoicall Apathy, as not to have a ſence of his unſavory Jeare (a Character of the goodneſs of his cauſe) and therefore tells him in his "own words in another place, upon another occaſion, That *he will* "*be ſilent, rather then imitate that part of his Spirit*, though there be in his vindication the advantages of many oppoſite retortions.

3. *Slave*, an unfit expreſſion to call the Jewiſh ſervant by.

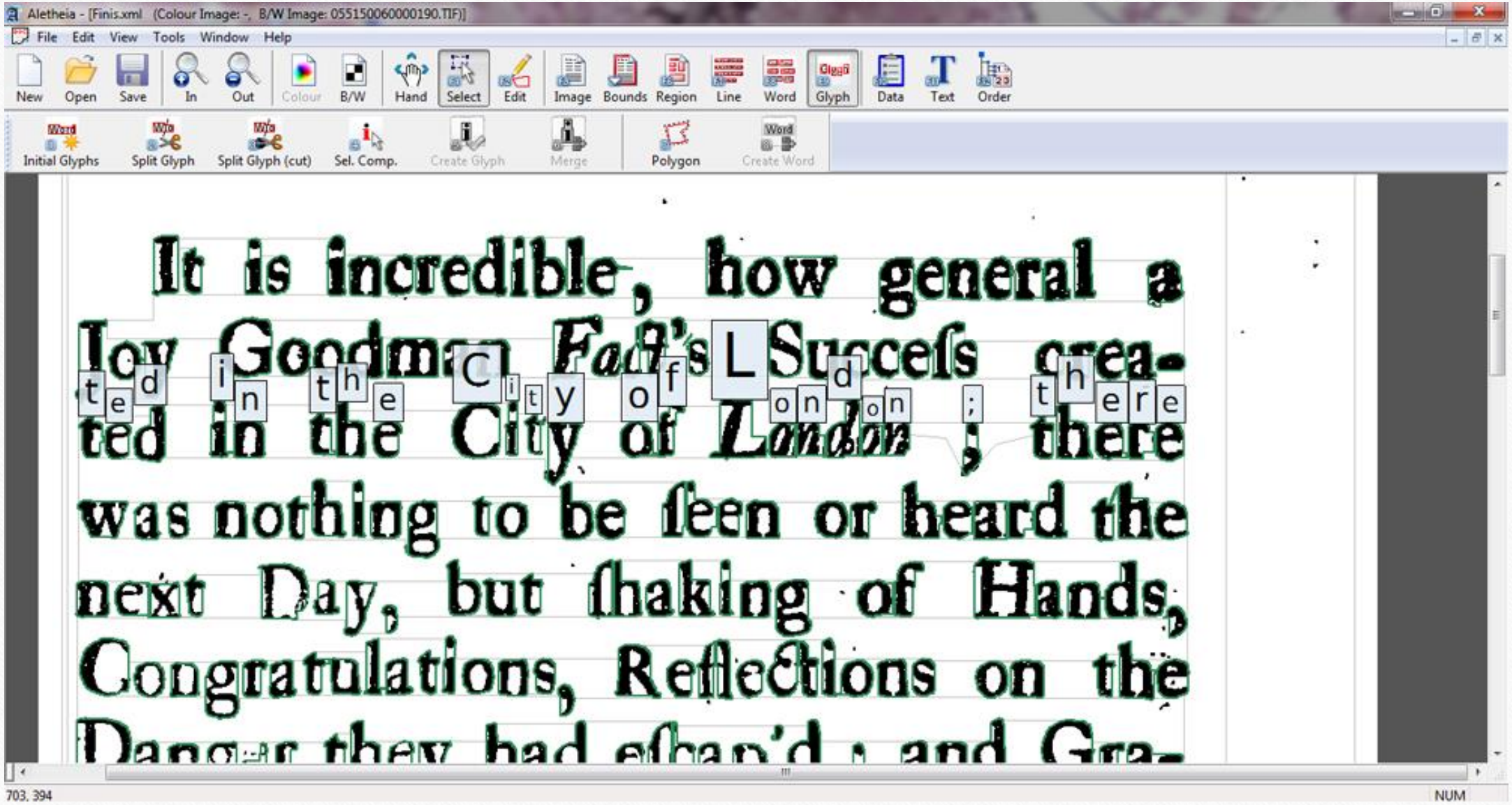
Guliel. Gibieuf. *de libert. Dei & creaturæ.* l. 1. p. 29. ca. 5. Sect. 5.

The third anſwer to the Doctors explanation of that place of *Exod.* 21. 6. *That the Iewiſh Servant did not give up himſelf to be a ſlave for ever, but his ſlavery did terminate in the yeare of Iubite*, here the Doctor obſtreperouſly vaunts over his opponent, yet he that takes ſeriouſly the weight of this mans arguments, will finde (I preſume) his virulency and malice poyzing more in the ſcale, then his reaſon: I muſt confeſs, that if a ſmooth dialect, and an exquiſite texture of words could overcloud a truth, this Doctor can do as much as any towards it: *He ſhall be a ſlave*, &c. it were well the Doctor could afford the ſervant *Iew* (whom the Chaldee renders a ſon of *Iſrael*) a better tearme then ſlave (this is a word well becoming them that would not out of neceſſity, but out of choice, and with all their induſtry make themſelves ſo) it may be the

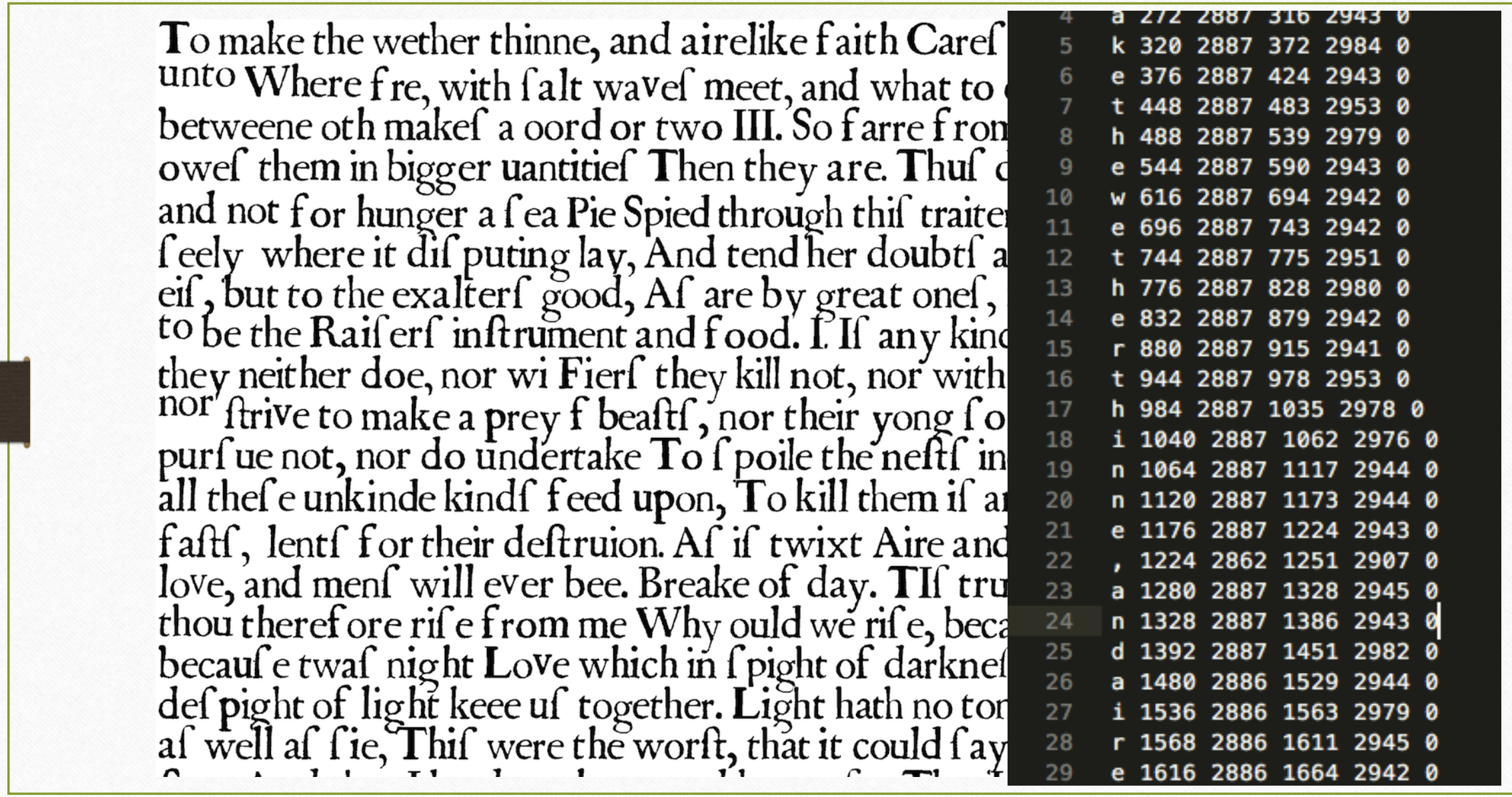
(to) littlēJ-Bffiëiheteinf -a-tidtelliheïj do'it'.the opinion of E. P;stands goods-lion-al.fhispretended feilds authoritative assertion'- and -for the JcloseWthiJs-Dojëtorijcistb SeHqun; E JI?, not,fo-poffii:ffed Mch.a Sfoizcallïliphthy'dasffiot Jto ii:'ive a fence'of his unsavory Jeare(a Cha-rafler ofthe goodness of. hisciufe j and therefore tells himin his Jcc own words inanot.lier place. upon another occasion' That he will u besilenr' pitcher then imitate rh.-zr.part of his Q.spir?t' though there be Jinghis vindication che advantages of many opposite terortions. -The third anfwer.t.o Jth'e Doctors-explanati'oh of that place of Mve:an 'Exod. 21.--6. That the Iewish.Servant did not give up himstlfto unfit ex- be a stavefor ever;'-but his slavery did terminate in the yeare oflubite, premonto hete rhe Do,flor obstreperously ,vaunts over his jopponen'r,

The Early Modern OCR Project (eMOP)

- Funded by a Mellon grant in 2012
- Train extant, open-source OCR engines to read Early Modern printed documents
- OCR documents from Early English Books Online (EEBO) and Eighteenth Century Collections Online (ECCO), or roughly **45 million page images** in two years.


Aletheia - [Finis.xml (Colour Image: -, B/W Image: 055150060000190.TIF)]
File Edit View Tools Window Help
New Open Save In Out Colour B/W Hand Select Edit Image Bounds Region Line Word Glyph Data Text Order
Initial Glyphs Split Glyph Split Glyph (cut) Sel. Comp. Create Glyph Merge Polygon Create Word
It is incredible, how general a Joy Goodman Fact's Succefs created in the City of London ; there was nothing to be ſeen or heard the next Day, but ſhaking of Hands, Congratulations, Reflections on the
703, 394
NUM


Font
Font Name
cbil
Line Height
0
Open Directories
Extracted Glyphs
TIF/Box Pairs
Save Font
Delete Font
Italic
Bold
Fixed
Serif
Fraktur
Selected Glyph Offsets
X Offset
Y Offset
0
0
Glyphs
a (U+0061)
Copy to Clipboard
Reclassify Glyph
Delete Glyph
Remove All
Delete Removed
Across all pages
Display Size
Prev Page
Next Page


To make the wether thinne, and airelike faith Caref
unto Where fre, with ſalt wavef meet, and what to
betweene oth makef a oord or two III. So farre fron
owef them in bigger uantitief Then they are. Thuf
and not for hunger a ſea Pie Spied through thif traite
ſeely where it diſputing lay, And tend her doubtf a
eif, but to the exalterf good, Af are by great onef,
to be the Raiſerf inſtrument and food. I. If any kin
they neither doe, nor wi Fierf they kill not, nor with
nor ſtrive to make a prey f beaſtf, nor their yong ſo
purſue not, nor do undertake To ſpoile the neſtf in
all thef e unkinde kindf feed upon, To kill them if a
faſtf, lentf for their deſtruion. Af if twixt Aire and
love, and menf will ever bee. Breake of day. TIf tru
thou therefore rif e from me Why ould we rif e, beca
becauſe twaf night Love which in ſpight of darknef
def pight of light keee uf together. Light hath no ton
af well af ſie, Thif were the worſt, that it could ſay
4 a 272 2887 316 2943 0
5 k 320 2887 372 2984 0
6 e 376 2887 424 2943 0
7 t 448 2887 483 2953 0
8 h 488 2887 539 2979 0
9 e 544 2887 590 2943 0
10 w 616 2887 694 2942 0
11 e 696 2887 743 2942 0
12 t 744 2887 775 2951 0
13 h 776 2887 828 2980 0
14 e 832 2887 879 2942 0
15 r 880 2887 915 2941 0
16 t 944 2887 978 2953 0
17 h 984 2887 1035 2978 0
18 i 1040 2887 1062 2976 0
19 n 1064 2887 1117 2944 0
20 n 1120 2887 1173 2944 0
21 e 1176 2887 1224 2943 0
22 , 1224 2862 1251 2907 0
23 a 1280 2887 1328 2945 0
24 n 1328 2887 1386 2943 0
25 d 1392 2887 1451 2982 0
26 a 1480 2886 1529 2944 0
27 i 1536 2886 1563 2979 0
28 r 1568 2886 1611 2945 0
29 e 1616 2886 1664 2942 0

IDHMC
TEXAS A&M UNIVERSITY

eMOP Dashboard

Initiative for Digital Humanities, Media, and Culture

Results Filter

Ground Truth: All
Works: All
Data Set: All
OCR Batch: All
Print Font: All
OCR completed date:
From: To:
Reset Filter Apply Filter

Job Queue

Scheduled:	36
In-Progress:	0
Await Postprocess:	0
Failed:	24351
Ingested:	40842453
Ingest Failed:	0
TOTAL:	40866840

Select All | Schedule Selected | Schedule All | Set Print Font for Selected Works | Export

Search: Show 25 entries

	Status		Data Set	ID	TCP Number	Title	Author	Font	OCR Date	OCR Engine	OCR Batch	Juxta	RETAS
☐	0-0-63-0	🔎	EEBO	1		Alkinoou eis ta tou Platonos eisagoge Alcinoi in Platonicam philosophiam introductio.	Albinus.		03/13/2015 17:30	Tesseract	4: EEBO w/o GT (SC8b-R8-D2b)	N/A	N/A
☐	0-0-2-0	🔎	EEBO	2	A24326	An Account of a horrid and barbarous murder committed on the body of a young person supposed to be of a good quality in the fields beyond Whitechappel-Church in the Parish of Stepny ...	Anon.	all other EEBO w/GT	03/11/2015 16:04	Tesseract	3: EEBO with GT (SC8b-R8-D2b)	0.694	N/A
☐	0-0-137-0	🔎	EEBO	3		The way of the Spirit in bringing souls to Christ set forth in X sermons on John 16:7, 8, 9, 10 and chap 7:37 / by Mr. Thomas Allen, late pastor of a church in ... Norwich.	Allen, Thomas, 1608-1673.		03/13/2015 17:30	Tesseract	4: EEBO w/o GT (SC8b-R8-D2b)	N/A	N/A
☐	0-0-1-0	🔎	EEBO	4	A47215	A contemplation on Bassets-down-Hill by the most sacred adorer of the Muses, Mrs. A.K.	A. K., Mrs.	all other EEBO w/GT	03/11/2015 16:04	Tesseract	3: EEBO with GT (SC8b-R8-D2b)	0.046	N/A
☐	0-0-18-0	🔎	EEBO	5	A29730	A dissuasive from popery sent in a letter from A.B. to C.D.	A. B.	all other EEBO w/GT	03/11/2015 16:04	Tesseract	3: EEBO with GT (SC8b-R8-D2b)	0.752	N/A
☐	0-0-3-0	🔎	EEBO	6	A24571	An Account of Oliver Hawley and John Condon who were executed at Tyburn on Friday the 2d of July 1686 for robing His Majesties male near Ilford in the county of Essex.	Anon.	all other EEBO w/GT	03/11/2015 16:04	Tesseract	3: EEBO with GT (SC8b-R8-D2b)	0.279	N/A

eMOP Project OCR Accuracy

EEBO	ECCO
68%	**86%**

eMOP Assets (emop.tamu.edu)

Database Schema

A MySQL database initially designed for EEBO and ECCO document metadata, then expanded to include ground truth and OCR results.

Storage

Network Attached Storage designed to serve up many terabytes of data.

eMOP Controller

A Python software package for the parallelized running of tasks on the Brazos Supercomputing Cluster.

Training Data

High-res, cleaned manuscript images with each character meticulously delineated and identified for a multitude of fonts used by Early Modern printers.

eMOP Dashboard

A Ruby-on-Rails web app for queueing up OCR jobs and viewing results.

OCR Workflow

The establishment of a step-by-step, automated process for OCR'ing page images and then improving that OCR via post-processing.

Enter the PrimerosLibros project…

The IDHMC's first attempt to leverage eMOP assets toward a new endeavor ☺

Responding to PrimerosLibros Challenges

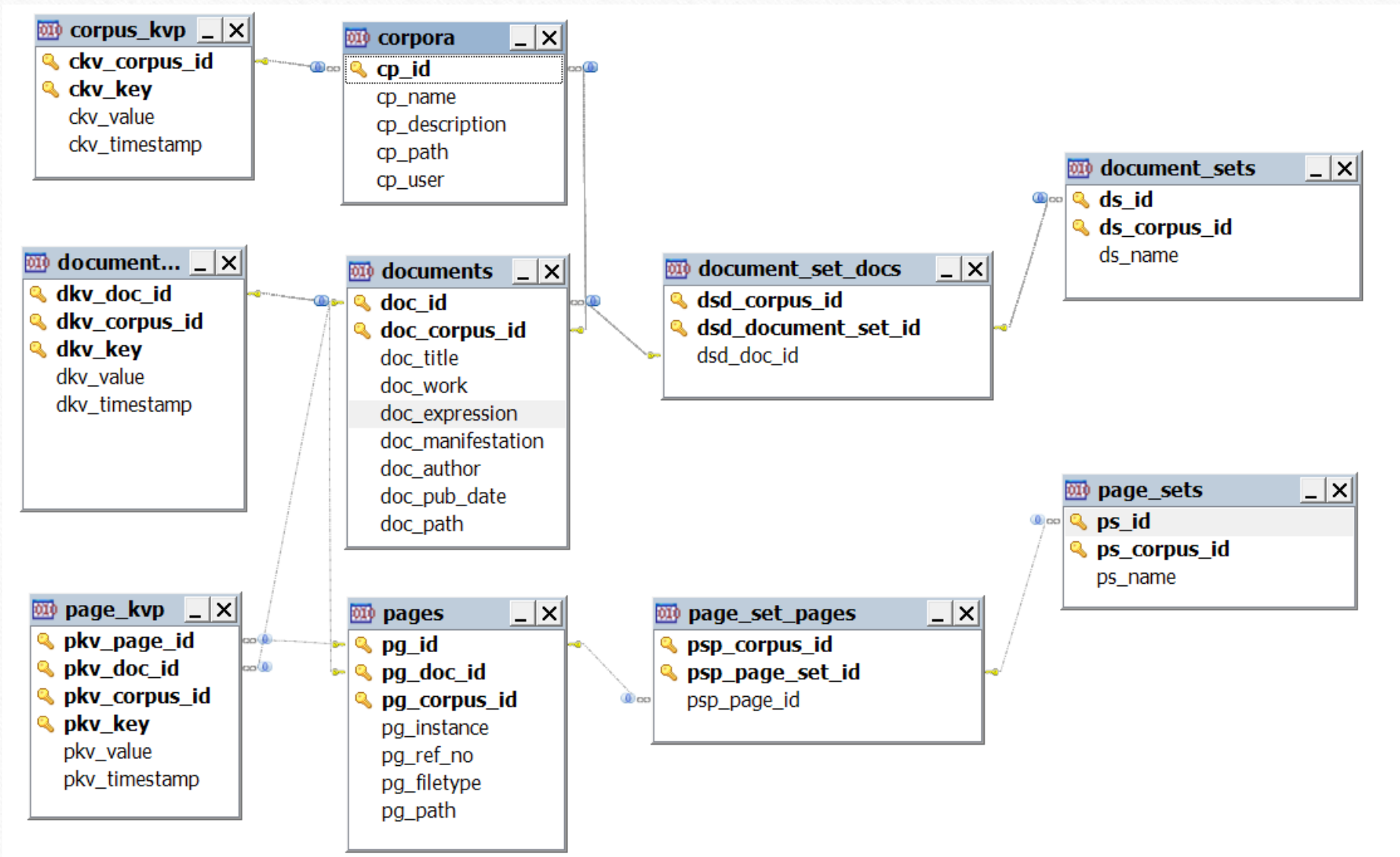
Machine-reading in languages other than English (Spanish, Latin, Nahuatl, Huastec, Mixtec, Otomi, Tarascan, and Zapotec)

No more code written in older versions of Python (2.7) that did not readily support Unicode out of the box. All tools developed for the PrimerosLibros project use Python 3.5.

Responding to PrimerosLibros Challenges

Supporting a New Database Schema

A complete redesign of the original eMOP database which supports (but does not enforce) a FRBR (work, expression, manifestation, instance) hierarchical structure, and otherwise allows for the storing of key/value pairs at the corpus, document, and page level, providing for a "schemaless" approach.


corpus_kvp
ckv_corpus_id
ckv_key
ckv_value
ckv_timestamp
corpora
cp_id
cp_name
cp_description
cp_path
cp_user
document_sets
ds_id
ds_corpus_id
ds_name
document...
dkv_doc_id
dkv_corpus_id
dkv_key
dkv_value
dkv_timestamp
documents
doc_id
doc_corpus_id
doc_title
doc_work
doc_expression
doc_manifestation
doc_author
doc_pub_date
doc_path
document_set_docs
dsd_corpus_id
dsd_document_set_id
dsd_doc_id
page_sets
ps_id
ps_corpus_id
ps_name
page_kvp
pkv_page_id
pkv_doc_id
pkv_corpus_id
pkv_key
pkv_value
pkv_timestamp
pages
pg_id
pg_doc_id
pg_corpus_id
pg_instance
pg_ref_no
pg_filetype
pg_path
page_set_pages
psp_corpus_id
psp_page_set_id
psp_page_id

DH Dashboard Corpus Manager Job Manager Hello, Bryan! Logout

All Documents

Show 25 out of 135 rows Search

	Document ID	FB ID	Title	Printer	Year	Training Set	View
☐	1000	pl_blac_004	***Doctrina cristiana para instrucción e informaci…***	J. Cromberger	1544	93	Pages Details
☐	1113	pl_blac_001	***Doctrina breve muy provechosa de las cosas q…***	J. Cromberger	1544	91	Pages Details
☐	1114	pl_blac_002	***Tripartito del cristianísimo y consolatorio docto…***	J. Cromberger	1544	158	Pages Details
☐	1115	pl_blac_003	***Este es un compendio breve que trata de la ma…***	J. Cromberger	1544	24	Pages Details
☐	1124	pl_houg_001	***Doctrina breve muy provechosa de las cosas q…***	J. Cromberger	1544	169	Pages Details
☐	893	pl_jcbl_002	***Cartilla para la enseñanza de la doctrina cristia…***	J. Pablos	1547	undefined	Pages Details
☐	1020	pl_jcbl_001	***Doctrina cristiana en lengua mexicana***	J. Cromberger /…	1547	40	Pages Details
☐	1252	pl_blac_020	***Regla cristiana breve para ordenar la vida y tie…***	undefined	1547	94	Pages Details
☐	1250	pl_blac_018	***Doctrina cristiana en lengua española y mexica…***	J. Pablos	1550	21	Pages Details
☐	1251	pl_blac_019	***Doctrina cristiana en lengua mexicana***	J. Pablos	1553	22	Pages Details
☐	901	pl_blac_005	***Recognitio,summularum reverendi patris Illdep…***	J. Pablos	1554	25	Pages Details
☐	902	pl_blac_006	***Dialectica resolutio cum textu Aristotelis***	J. Pablos	1554	26	Pages Details
☐	959	pl_uslp_001	***Recognitio,summularum reverendi patris Illdep…***	J. Pablos	1554	86	Pages Details

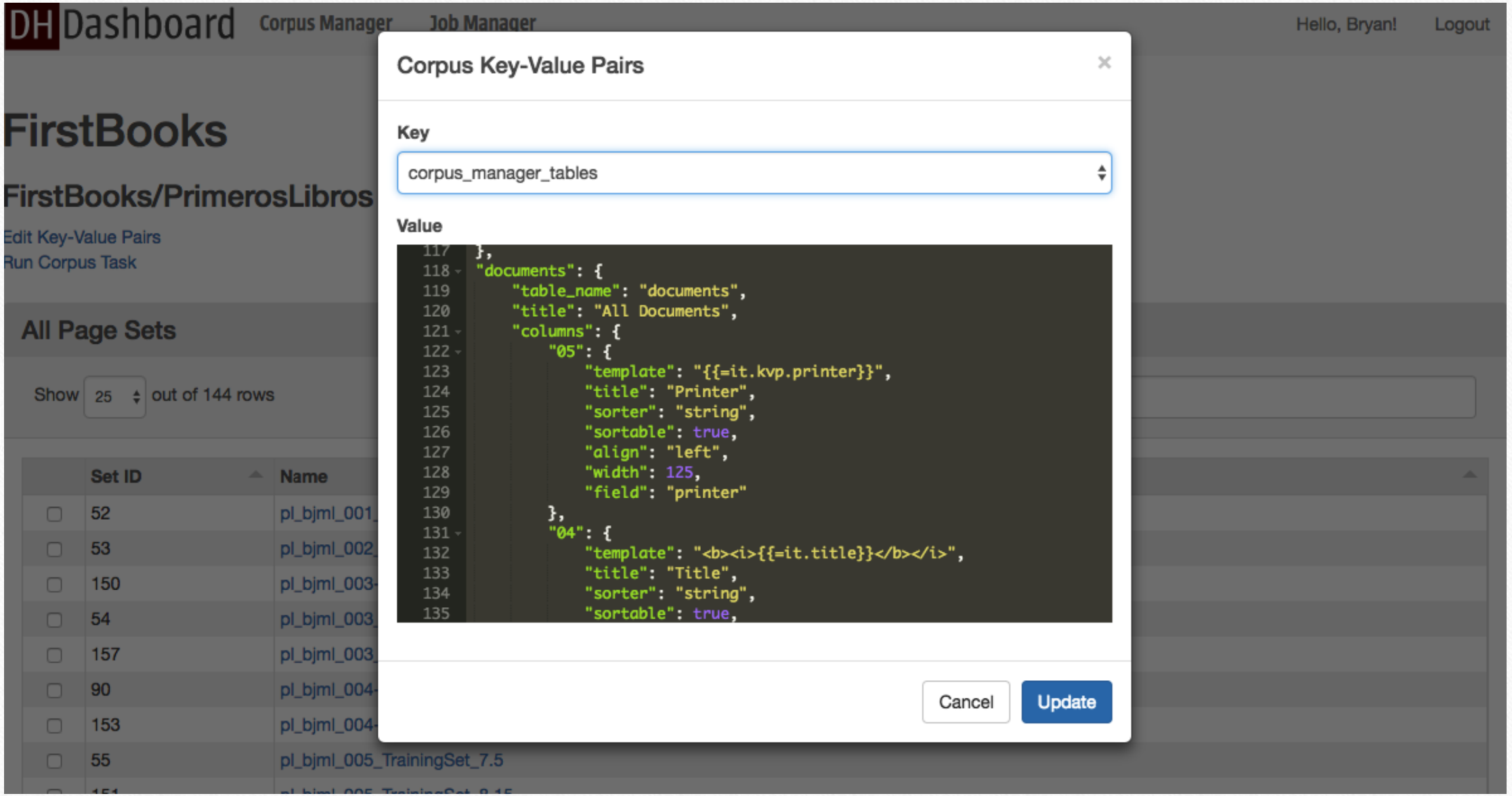
DH Dashboard Corpus Manager Job Manager Hello, Bryan! Logout

Documents from Set 3 Engine Comparison

Back to All Documents

Show : out of 61 rows Search

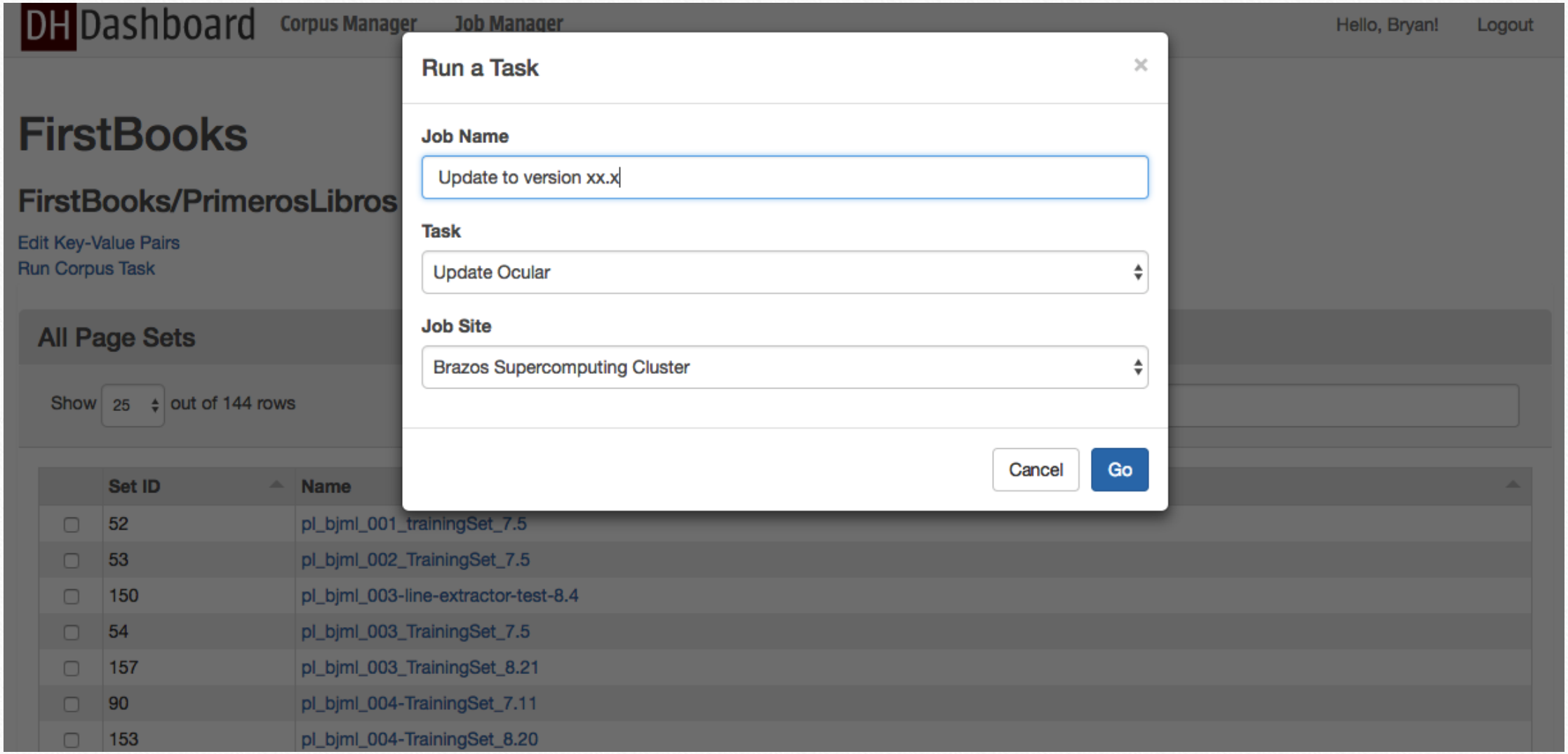
	Doc ID	Title	GCV Score	Tess4 Score	eMOP Score	Source	View
☐	40355	*Pasquils passe, and passeth not Set downe in three pees. His passe, precession, and progn…*	0.25	0.30	0.78	EEBO	Pages Details
☐	40356	*Much adoe about nothing As it hath been sundrie times publikely acted by the right honoura…*	0.50	0.29	0.81	EEBO	Pages Details
☐	40357	*The lettin[g] of humours blood in the head-vaine with a new morissco, daunced by seauen s…*	0.25	0.21	0.22	EEBO	Pages Details
☐	40363	*Inimicus amicus an excellent treatise, shewing, how a man may reape profit by his enemy.*	0.33	0.22	0.64	EEBO	Pages Details
☐	40364	*The exaltation of the kingdome and priesthood of Christ In certaine sermons vpon the 110. P…*	0.42	0.21	0.57	EEBO	Pages Details
☐	40366	*A treatise concerning the trinitie of persons in vnitie of the deitie Written to Thomas Manneri…*	0.21	0.24	0.68	EEBO	Pages Details
☐	40367	*Dialogicall discourses of spirits and divels declaring their proper essence, natures, dispositi…*	0.40	0.20	0.66	EEBO	Pages Details
☐	40368	*The mirror of diuine prouidence Containing a collection of Theodoret his arguments: declari…*	0.08	0.19	0.30	EEBO	Pages Details


DH Dashboard
Corpus Manager
Job Manager
Hello, Bryan!
Logout
FirstBooks
FirstBooks/PrimerosLibros
Edit Key-Value Pairs
Run Corpus Task
All Page Sets
Show 25 out of 144 rows
Set ID
Name
52
53
150
54
157
90
153
55
pl_bjml_005_TrainingSet_7.5
Corpus Key-Value Pairs
Key
corpus_manager_tables
Value
},
"documents": {
"table_name": "documents",
"title": "All Documents",
"columns": {
"05": {
"template": "{{=it.kvp.printer}}",
"title": "Printer",
"sorter": "string",
"sortable": true,
"align": "left",
"width": 125,
"field": "printer"
},
"04": {
"template": "{{=it.title}}",
"title": "Title",
"sorter": "string",
"sortable": true,
Cancel
Update

Responding to PrimerosLibros Challenges

Supporting Ocular (cont.)

Because Ocular is under active development, PrimerosLibros collaborators needed a way to pull the latest code down from their git repo, compile, and then install the new version for use.


DH Dashboard
Corpus Manager
Job Manager
Hello, Bryan!
Logout
FirstBooks
FirstBooks/PrimerosLibros
Edit Key-Value Pairs
Run Corpus Task
All Page Sets
Show 25 out of 144 rows
Set ID
Name
52 pl_bjml_001_trainingSet_7.5
53 pl_bjml_002_TrainingSet_7.5
150 pl_bjml_003-line-extractor-test-8.4
54 pl_bjml_003_TrainingSet_7.5
157 pl_bjml_003_TrainingSet_8.21
90 pl_bjml_004-TrainingSet_7.11
153 pl_bjml_004-TrainingSet_8.20
Run a Task
Job Name
Update to version xx.x
Task
Update Ocular
Job Site
Brazos Supercomputing Cluster
Cancel
Go

Responding to PrimerosLibros Challenges

Supporting a New OCR Engine (Ocular)

Unlike Tesseract, which requires a tedious, manual training process involving pre-processed images, etc., Ocular allows for the unsupervised training of its classifier given a set of exemplary images for each document.

DH Dashboard Corpus Manager Job Manager Hello, Bryan! Logout

All Page Sets

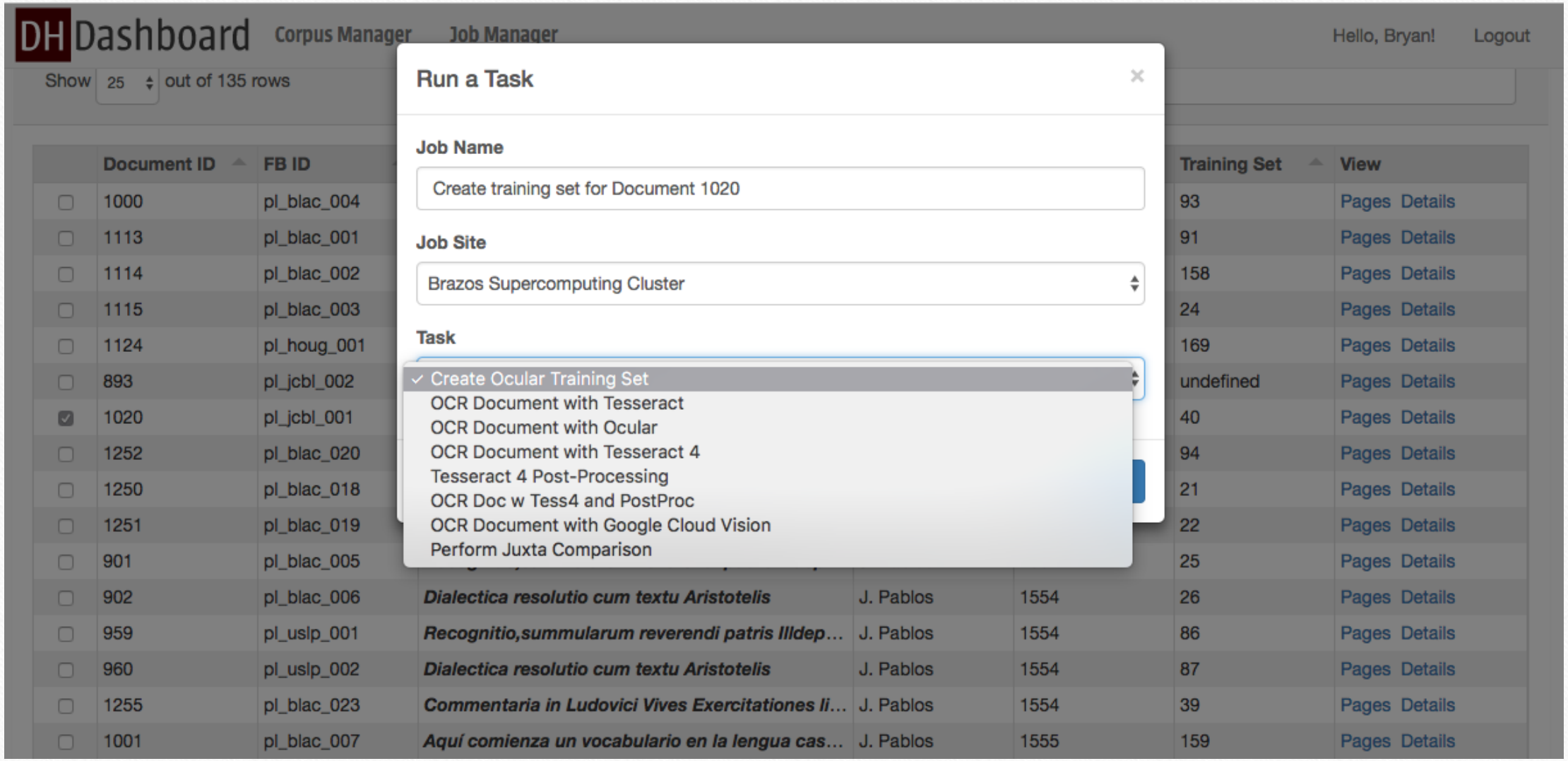
Show 25 out of 144 rows Search

	Set ID	Name
☐	52	pl_bjml_001_trainingSet_7.5
☐	53	pl_bjml_002_TrainingSet_7.5
☐	150	pl_bjml_003-line-extractor-test-8.4
☐	54	pl_bjml_003_TrainingSet_7.5
☐	157	pl_bjml_003_TrainingSet_8.21
☐	90	pl_bjml_004-TrainingSet_7.11
☐	153	pl_bjml_004-TrainingSet_8.20
☐	55	pl_bjml_005_TrainingSet_7.5
☐	151	pl_bjml_005_TrainingSet_8.15
☐	154	pl_bjml_005_TrainingSet_8.20
☐	56	pl_bjml_006_TrainingSet_7.5
☐	165	pl_bjml_006_TrainingSet_8.23
☐	91	pl_blac_001-TrainingSet_7.11
☐	92	pl_blac_002-TrainingSet_7.11

Responding to PrimerosLibros Challenges

Supporting Ocular (cont.)

Whereas the old eMOP Dashboard supported a single, fixed workflow for OCR'ing documents, Ocular requires a two-step process for reading each document (font training on exemplary page sets, transcription).

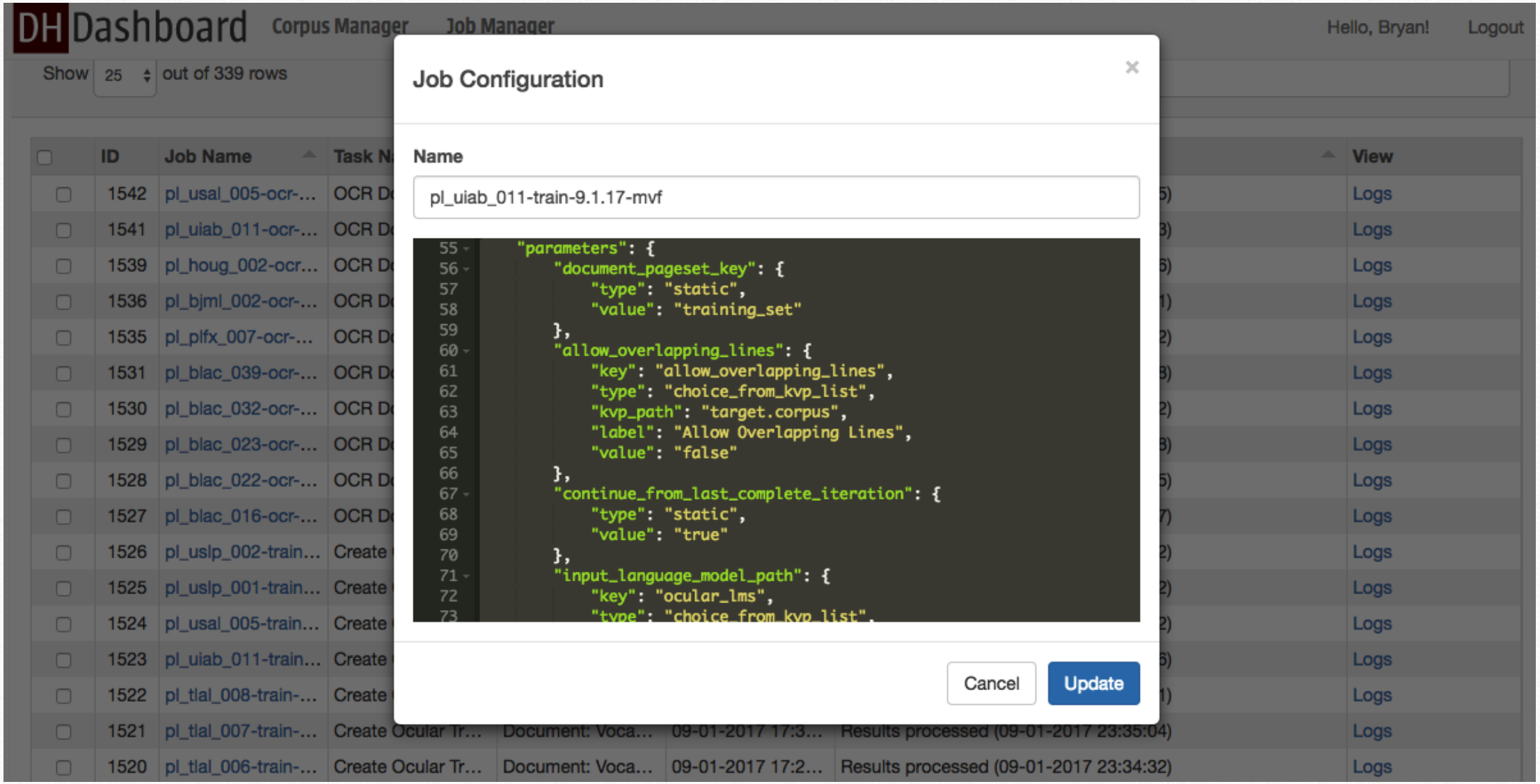

DH Dashboard
Corpus Manager
Job Manager
Hello, Bryan!
Logout
Show 25 out of 135 rows
Run a Task
Job Name
Create training set for Document 1020
Job Site
Brazos Supercomputing Cluster
Task
Create Ocular Training Set
OCR Document with Tesseract
OCR Document with Ocular
OCR Document with Tesseract 4
Tesseract 4 Post-Processing
OCR Doc w Tess4 and PostProc
OCR Document with Google Cloud Vision
Perform Juxta Comparison
Document ID
FB ID
Training Set
View
1000 pl_blac_004 93 Pages Details
1113 pl_blac_001 91 Pages Details
1114 pl_blac_002 158 Pages Details
1115 pl_blac_003 24 Pages Details
1124 pl_houg_001 169 Pages Details
893 pl_jcbl_002 undefined Pages Details
1020 pl_jcbl_001 40 Pages Details
1252 pl_blac_020 94 Pages Details
1250 pl_blac_018 21 Pages Details
1251 pl_blac_019 22 Pages Details
901 pl_blac_005 25 Pages Details
902 pl_blac_006 Dialectica resolutio cum textu Aristotelis J. Pablos 1554 26 Pages Details
959 pl_uslp_001 Recognitio,summularum reverendi patris IIIdep… J. Pablos 1554 86 Pages Details
960 pl_uslp_002 Dialectica resolutio cum textu Aristotelis J. Pablos 1554 87 Pages Details
1255 pl_blac_023 Commentaria in Ludovici Vives Exercitationes li… J. Pablos 1554 39 Pages Details
1001 pl_blac_007 Aquí comienza un vocabulario en la lengua cas… J. Pablos 1555 159 Pages Details

DH Dashboard Corpus Manager Job Manager Hello, Bryan! Logout

All Jobs

Show 25 out of 339 rows Search

	ID	Job Name	Task Name	Target	Submitted	Status	View
☐	1542	pl_usal_005-ocr-...	OCR Document ...	Document: Vida ...	09-04-2017 11:3...	Results processed (09-07-2017 18:45:55)	Logs
☐	1541	pl_uiab_011-ocr-...	OCR Document ...	Document: Prem...	09-04-2017 11:3...	Results processed (09-07-2017 00:29:53)	Logs
☐	1539	pl_houg_002-ocr...	OCR Document ...	Document: Provi...	09-04-2017 11:3...	Results processed (09-07-2017 18:48:46)	Logs
☐	1536	pl_bjml_002-ocr-...	OCR Document ...	Document: Tabul...	09-04-2017 10:4...	Results processed (09-05-2017 04:30:11)	Logs
☐	1535	pl_plfx_007-ocr-...	OCR Document ...	Document: Trata...	09-04-2017 10:4...	Results processed (09-05-2017 04:31:42)	Logs
☐	1531	pl_blac_039-ocr-...	OCR Document ...	Document: Psal...	09-03-2017 11:0...	Results processed (09-04-2017 06:50:48)	Logs
☐	1530	pl_blac_032-ocr-...	OCR Document ...	Document: Arte ...	09-03-2017 11:0...	Results processed (09-04-2017 01:47:32)	Logs
☐	1529	pl_blac_023-ocr-...	OCR Document ...	Document: Com...	09-03-2017 11:0...	Results processed (09-04-2017 00:14:18)	Logs
☐	1528	pl_blac_022-ocr-...	OCR Document ...	Document: Manu...	09-03-2017 11:0...	Results processed (09-03-2017 16:12:25)	Logs
☐	1527	pl_blac_016-ocr-...	OCR Document ...	Document: Arte ...	09-03-2017 11:0...	Results processed (09-03-2017 12:39:07)	Logs
☐	1526	pl_uslp_002-train...	Create Ocular Tr...	Document: Diale...	09-02-2017 01:3...	Results processed (09-02-2017 09:07:32)	Logs
☐	1525	pl_uslp_001-train...	Create Ocular Tr...	Document: Reco...	09-02-2017 01:3...	Results processed (09-02-2017 05:58:32)	Logs
☐	1524	pl_usal_005-train...	Create Ocular Tr...	Document: Vida ...	09-02-2017 01:3...	Results processed (09-02-2017 15:13:32)	Logs


DH Dashboard
Corpus Manager
Job Manager
Hello, Bryan!
Logout
Show 25 out of 339 rows
ID
Job Name
Task N
View
1542 pl_usal_005-ocr-... OCR D Logs
1541 pl_uiab_011-ocr-... OCR D Logs
1539 pl_houg_002-ocr... OCR D Logs
1536 pl_bjml_002-ocr-... OCR D Logs
1535 pl_plfx_007-ocr-... OCR D Logs
1531 pl_blac_039-ocr-... OCR D Logs
1530 pl_blac_032-ocr-... OCR D Logs
1529 pl_blac_023-ocr-... OCR D Logs
1528 pl_blac_022-ocr-... OCR D Logs
1527 pl_blac_016-ocr-... OCR D Logs
1526 pl_uslp_002-train... Create Logs
1525 pl_uslp_001-train... Create Logs
1524 pl_usal_005-train... Create Logs
1523 pl_uiab_011-train... Create Logs
1522 pl_tlal_008-train-... Create Logs
1521 pl_tlal_007-train-... Create Ocular Tr... Document: Voca... 09-01-2017 17:3... Results processed (09-01-2017 23:35:04) Logs
1520 pl_tlal_006-train-... Create Ocular Tr... Document: Voca... 09-01-2017 17:2... Results processed (09-01-2017 23:34:32) Logs
Job Configuration
Name
pl_uiab_011-train-9.1.17-mvf
"parameters": {
"document_pageset_key": {
"type": "static",
"value": "training_set"
},
"allow_overlapping_lines": {
"key": "allow_overlapping_lines",
"type": "choice_from_kvp_list",
"kvp_path": "target.corpus",
"label": "Allow Overlapping Lines",
"value": "false"
},
"continue_from_last_complete_iteration": {
"type": "static",
"value": "true"
},
"input_language_model_path": {
"key": "ocular_lms",
Cancel
Update

DH Dashboard Corpus Manager Job Manager Hello, Bryan! Logout

		1000_transcription_normalized.txt **Ocular Transcription:** pl_blac_019_00011-1000_transcription.txt
Page Num: 12 **ID:** 101890 **Image**: pl_blac_019_00012-1000.jpg **fb_page_id**: pl_blac_019_00012	pl_blac_019-ocr-7.19.17-mvf(1325)	**Ocular Comparison:** pl_blac_019_00012-1000_comparisons.txt **Ocular DIPL XML:** pl_blac_019_00012-1000_dipl.alto.xml **Ocular NORM XML:** pl_blac_019_00012-1000_norm.alto.xml **Ocular Normalized Transcription:** pl_blac_019_00012-1000_transcription_normalized.txt **Ocular Transcription:** pl_blac_019_00012-1000_transcription.txt
	pl_blac_019-OCR_5.26-haa-v03-2(1113)	**Ocular Comparison:** pl_blac_019_00012-1000_comparisons.txt **Ocular DIPL XML:** pl_blac_019_00012-1000_dipl.alto.xml **Ocular NORM XML:** pl_blac_019_00012-1000_norm.alto.xml **Ocular Normalized Transcription:** pl_blac_019_00012-1000_transcription_normalized.txt **Ocular Transcription:** pl_blac_019_00012-1000_transcription.txt
Page Num: 13 **ID:** 101891 **Image**: pl_blac_019_00013-1000.jpg **fb_page_id**: pl_blac_019_00013	pl_blac_019-ocr-7.19.17-mvf(1325)	**Ocular Comparison:** pl_blac_019_00013-1000_comparisons.txt **Ocular DIPL XML:** pl_blac_019_00013-1000_dipl.alto.xml **Ocular NORM XML:** pl_blac_019_00013-1000_norm.alto.xml **Ocular Normalized Transcription:** pl_blac_019_00013-1000_transcription_normalized.txt **Ocular Transcription:** pl_blac_019_00013-1000_transcription.txt

In conclusion…

- In order to change the perception of the state of the OCR problem, humanists must be aware of and emphasize the poor state of machine-reading for anything other than documents created using modern printing practices.
- UNICODE.
- Tools like the DH Dashboard need to de-privilege specific OCR workflows by allowing for flexibility in terms of both processes and data.

Thank you!

Bryan Tarpley

Initiative for Digital Humanities, Media, and Culture

Texas A&M University

bptarpley@tamu.edu