BULK REVIEWER

Tim Walsh
Digital Preservation Librarian, Concordia University Library
2018 Summer Fellow, Harvard Library Innovation Lab
tim.walsh@concordia.ca

Identify, review, and remove sensitive files

- In disk images and directories
- Regardless of file format
- Powered by bulk_extractor
- Review results, dismiss false positives, generate reports, and export files in prototype browser application
- Built using Django, Vue.js, bulk_extractor, DFXML & Docker

Social Security Numbers

Credit card numbers

Phone numbers

Email addresses

Internet history

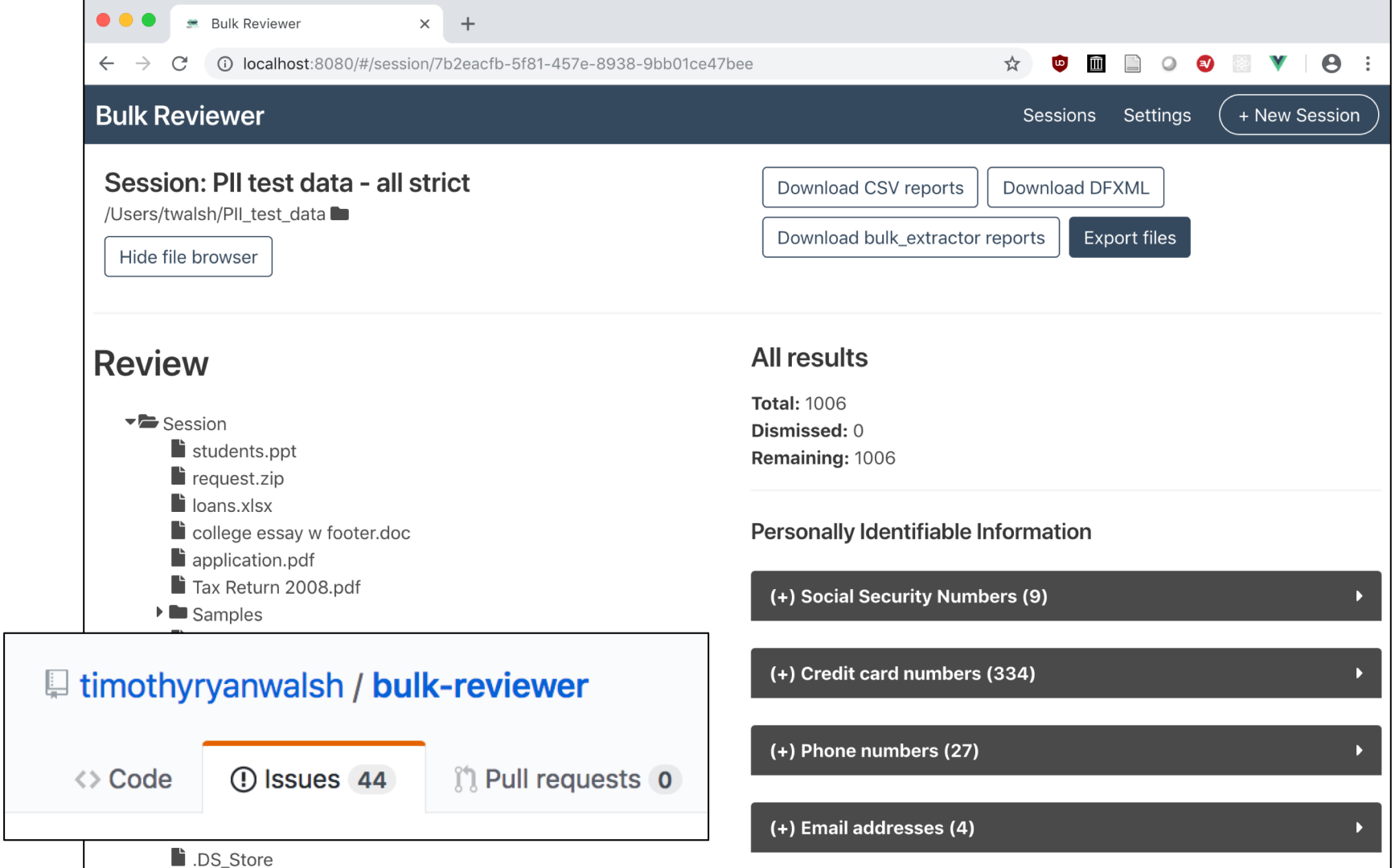
EXIF metadata

Geolocation (GPS) data

Custom search terms

Library Innovation Lab

Langdell Hall. Harvard University.


Bulk Reviewer
localhost:8080/#/session/7b2eacfb-5f81-457e-8938-9bb01ce47bee
Bulk Reviewer
Sessions
Settings
+ New Session
Session: PII test data - all strict
/Users/twalsh/PII_test_data
Hide file browser
Download CSV reports
Download DFXML
Download bulk_extractor reports
Export files
Review
Session
students.ppt
request.zip
loans.xlsx
college essay w footer.doc
application.pdf
Tax Return 2008.pdf
Samples
All results
Total: 1006
Dismissed: 0
Remaining: 1006
Personally Identifiable Information
(+) Social Security Numbers (9)
(+) Credit card numbers (334)
(+) Phone numbers (27)
(+) Email addresses (4)
timothyryanwalsh / bulk-reviewer
Code
Issues 44
Pull requests 0
.DS_Store

Development roadmap

Improvements on current tool (alpha):

- Add **new scanners**: Facebook, Outlook, identity-based
- Make it fit for **Canadian context**: SIN scanner, test data, internationalization/translation
- **Pre-scan OCR** for PDFs and standard image formats
- Scalability and security improvements
- UX/UI and accessibility improvements

Next-generation improvements (beta):

- **Machine learning** (context-based risk assessment)

Thank you!

Interested in getting involved? Get in touch!

tim.walsh@concordia.ca
@bitarchivist
https://github.com/timothyryanwalsh/bulk-reviewer