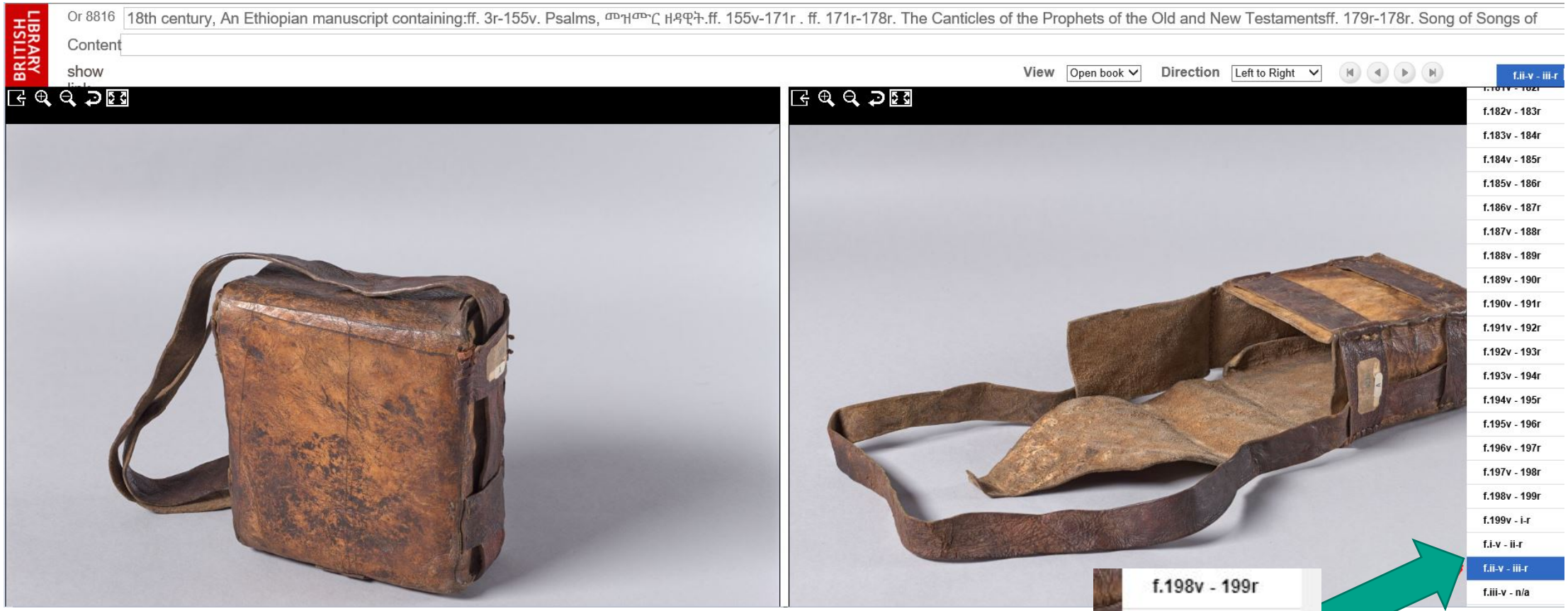
flyswot: garden-variety machine learning applications

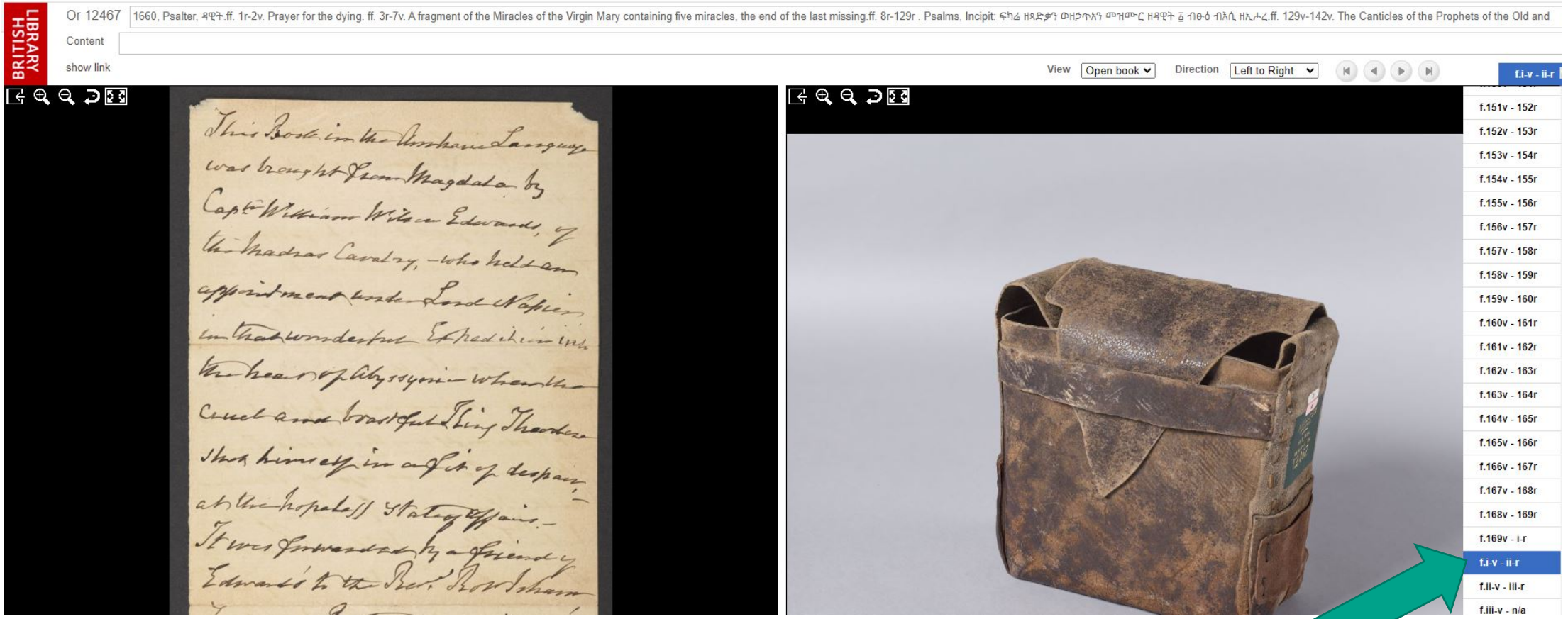
Catherine Cronin – Digitisation Workflow Officer
Daniel van Strien – Digital Curator
Andrew Longworth – Digitisation Workflow Manager

BRITISH LIBRARY

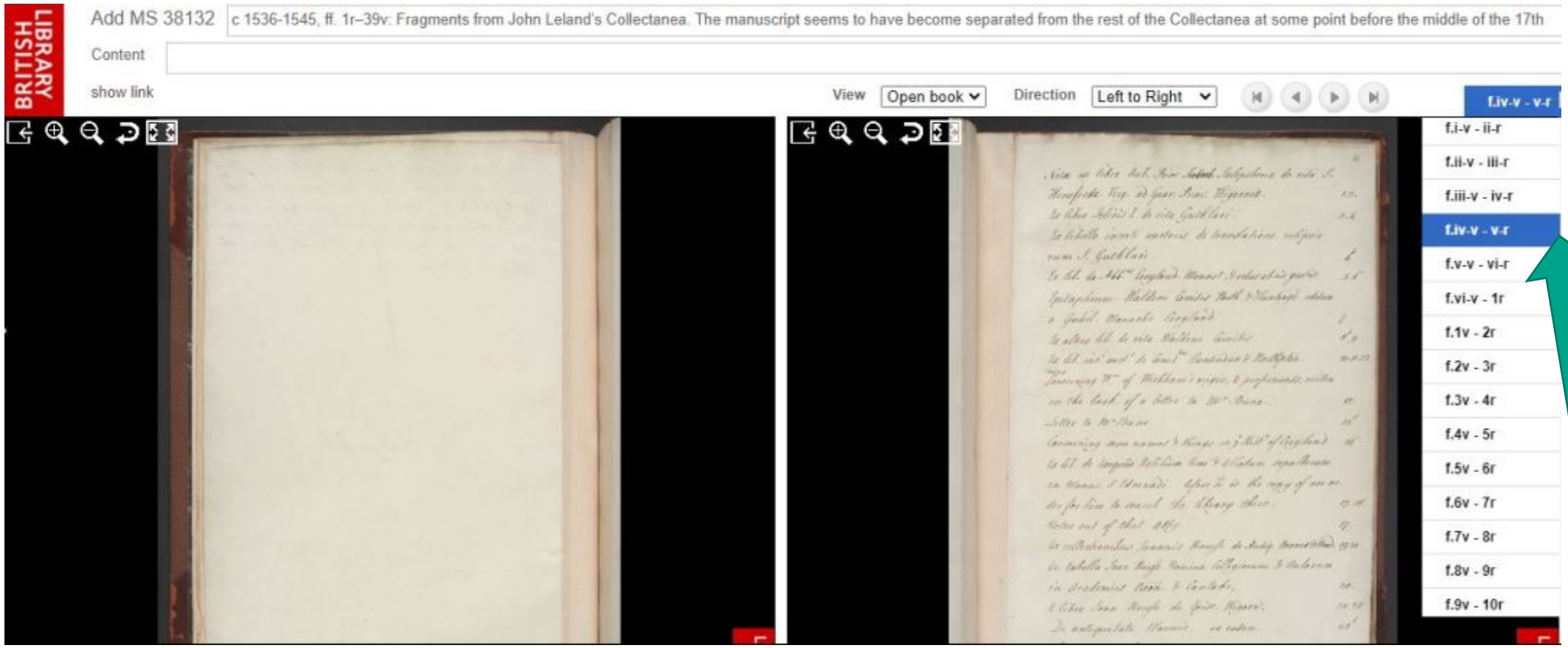
What do all these images have in common?



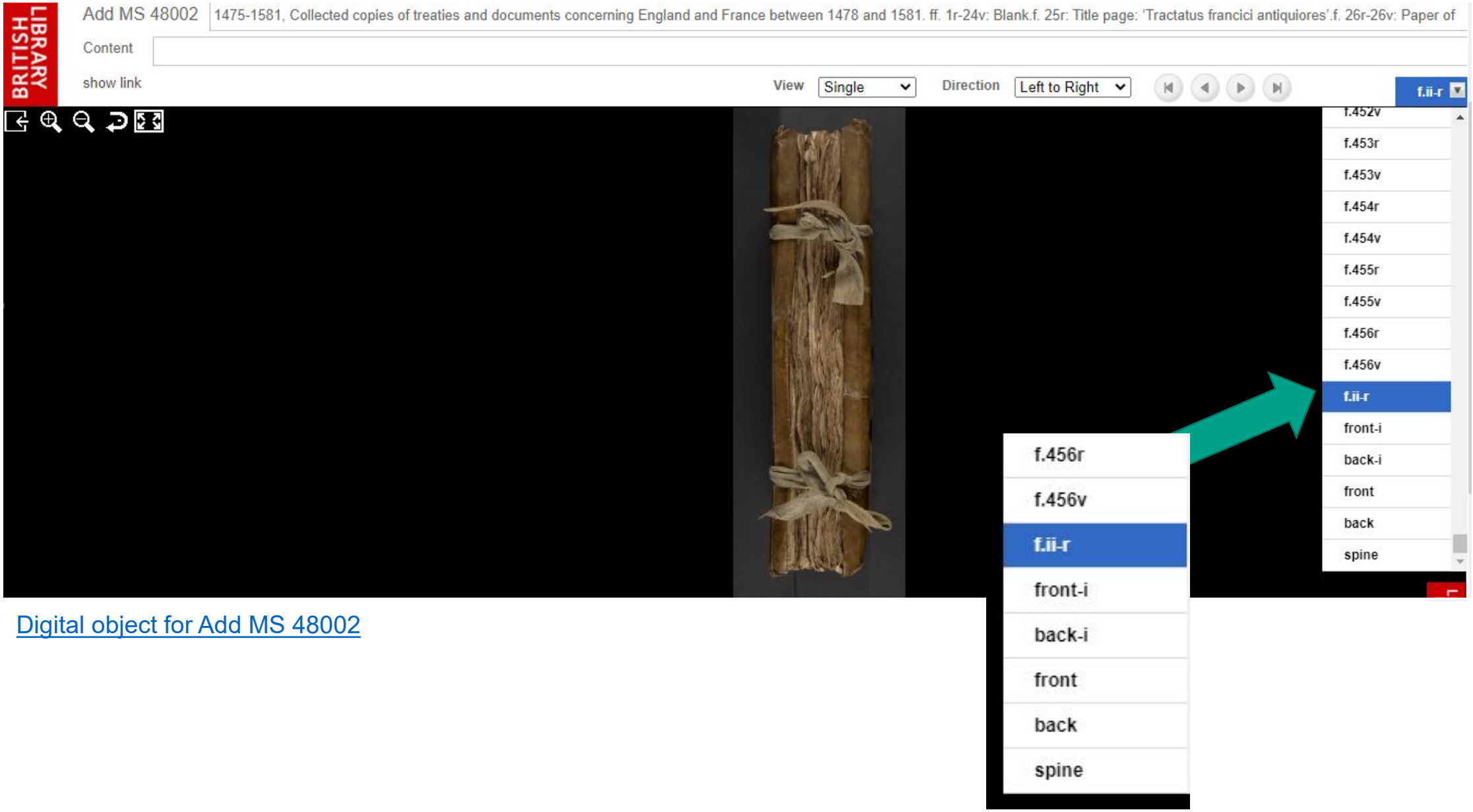
[Digital object for Or 8816]



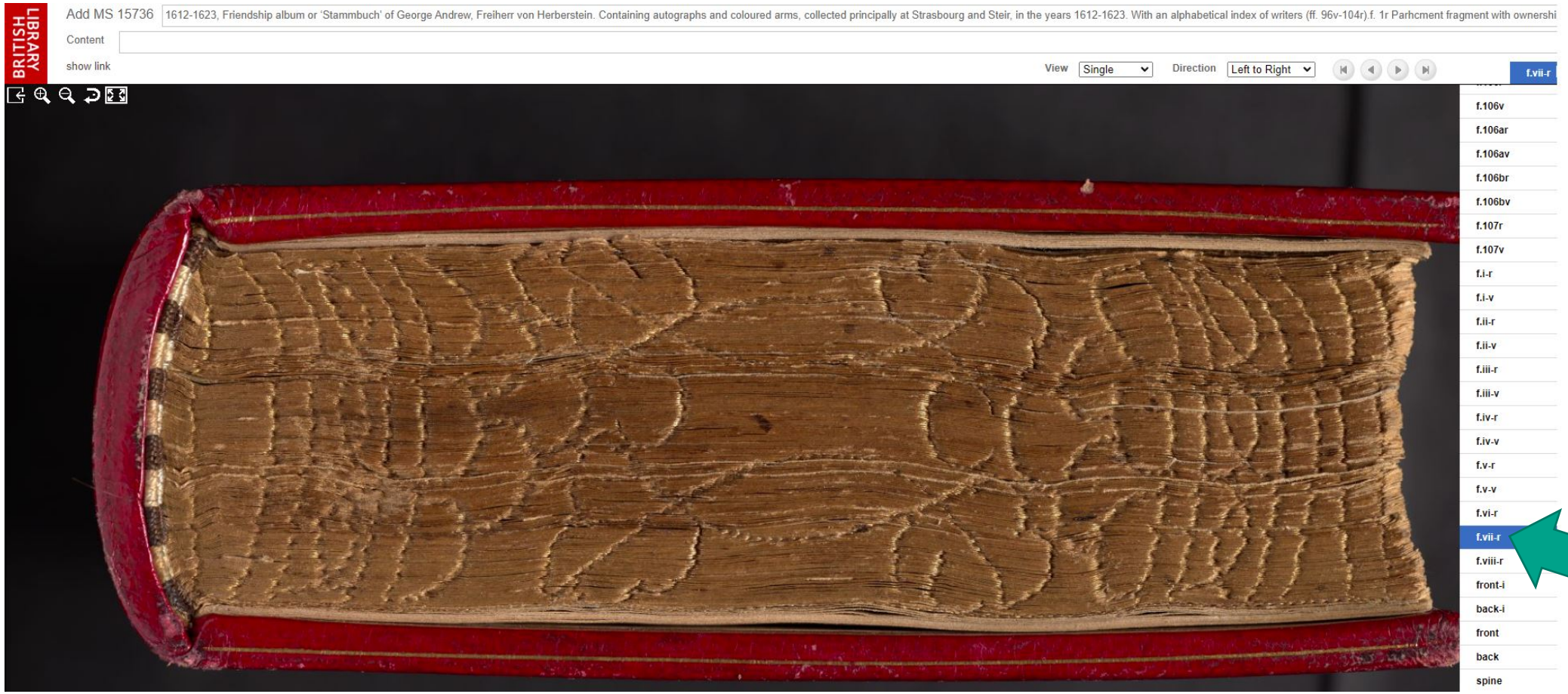
[Digital object for Or 12647]



[Digital object for Add MS 38132]



[Digital object for Add MS 48002]



[Digital object for Add MS 15736]

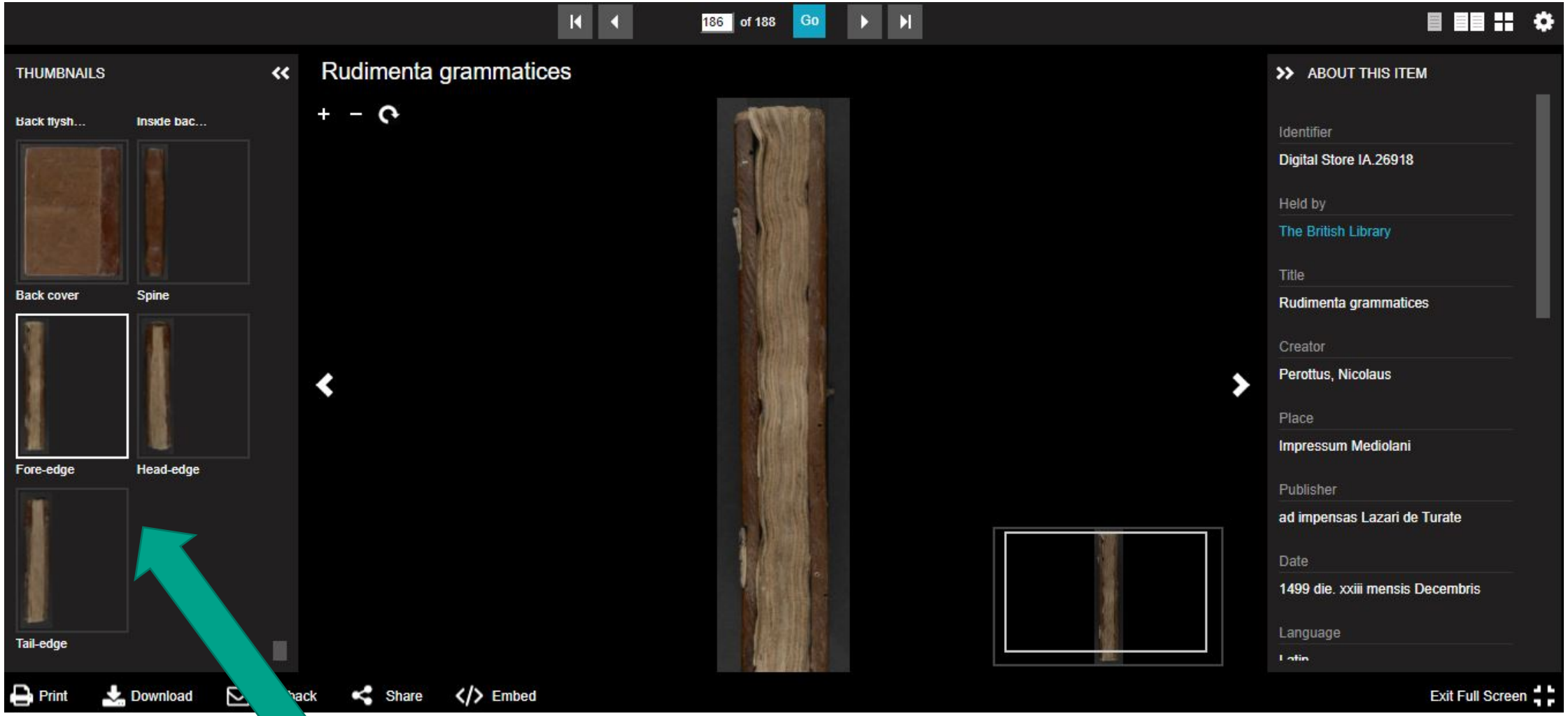
[Digital object Add MS 11829]

They are all FAKE FLYSHEETS

- Limited set of visible image labels
- Inflexible order sequence of images
- Tied to an old digital object platform
- Labels derived from image file names
- Many images incorrectly labelled as a flysheet

New IIIF Universal Viewer

- Digital objects migrated to new viewer
- Standardised vocabulary for image labels
- Flexibility in order sequence of images
- Ability to apply appropriate labels to ‘fake flysheet’ images

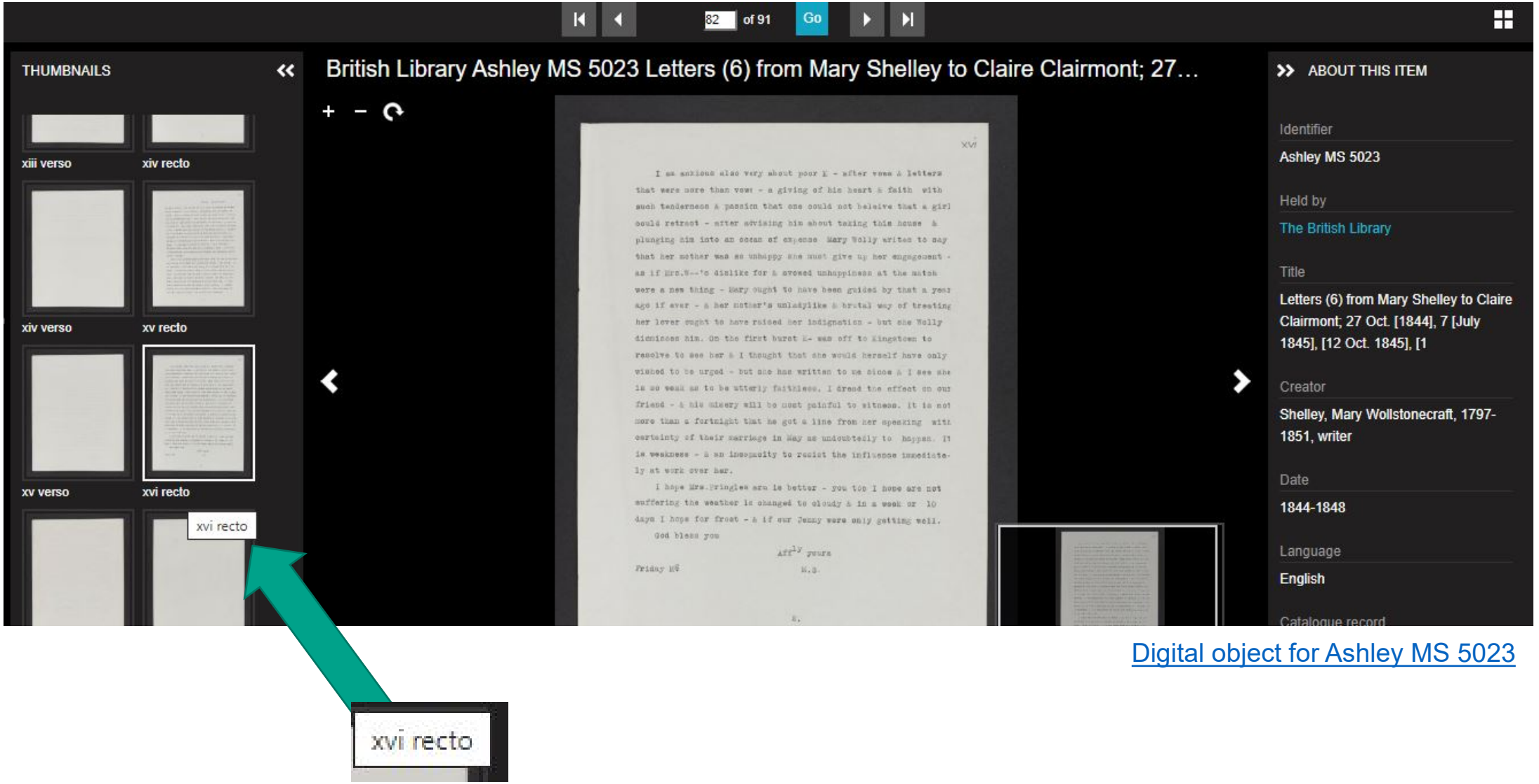


Fore-edge

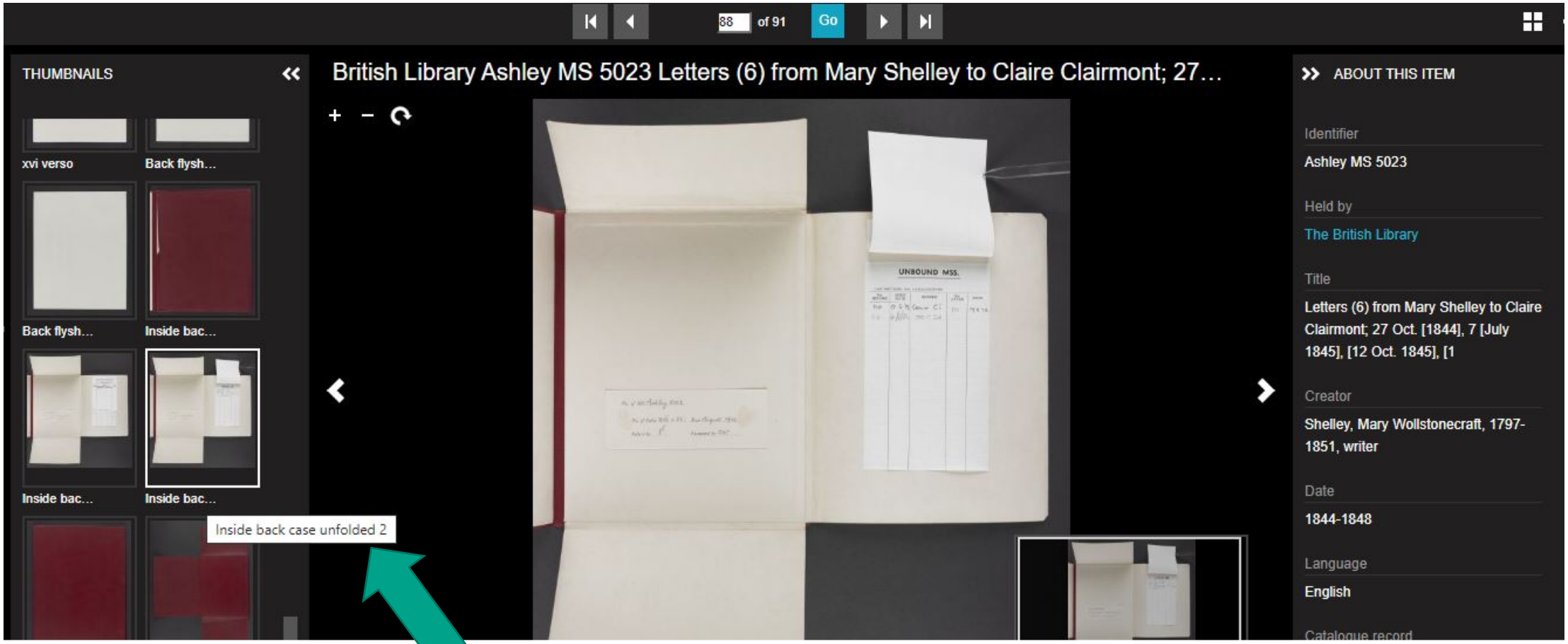
Head-edge

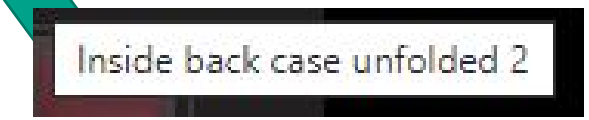
Tail-edge

Digital object for IA.26918

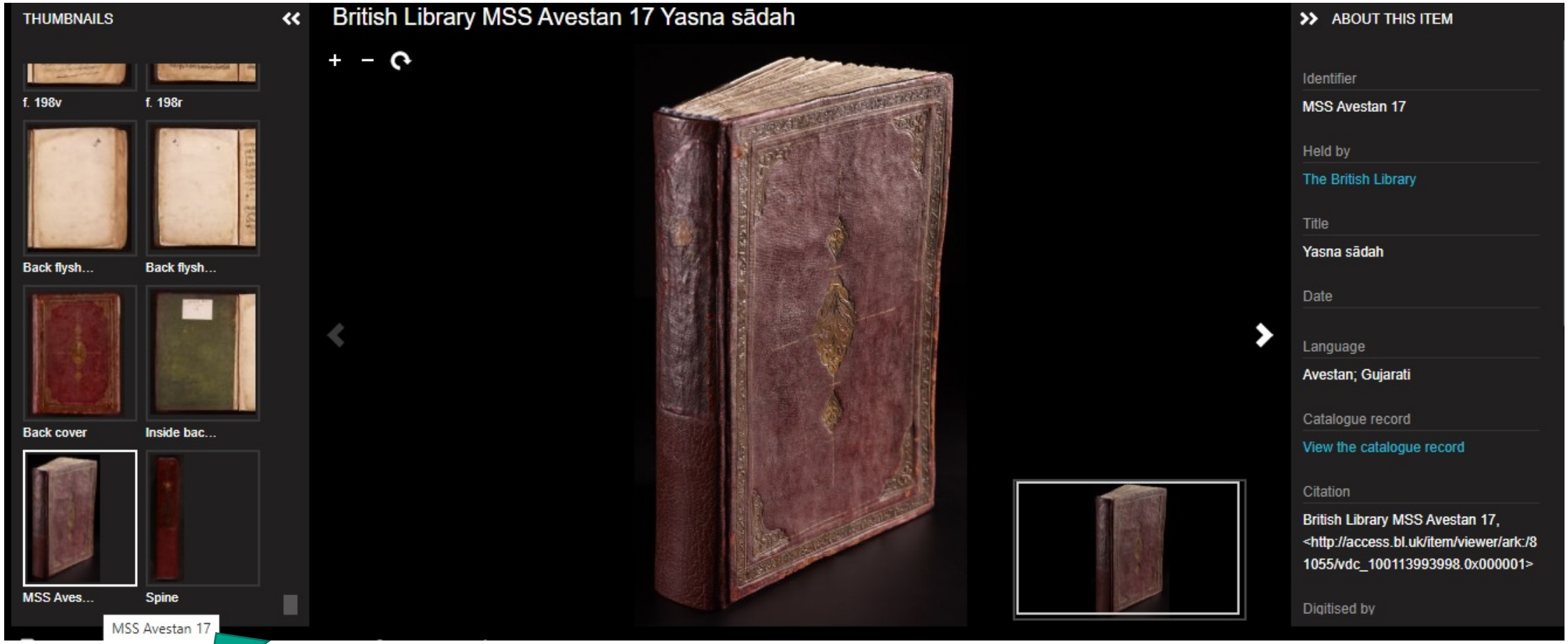


[Digital object for Ashley MS 5023]





Digital object for Ashley MS 5023



Digital object for MSS Avestan 17

We have a mountain of digitised content:

- Tens of thousands of images
- About 1 petabyte (1,000 terabytes)

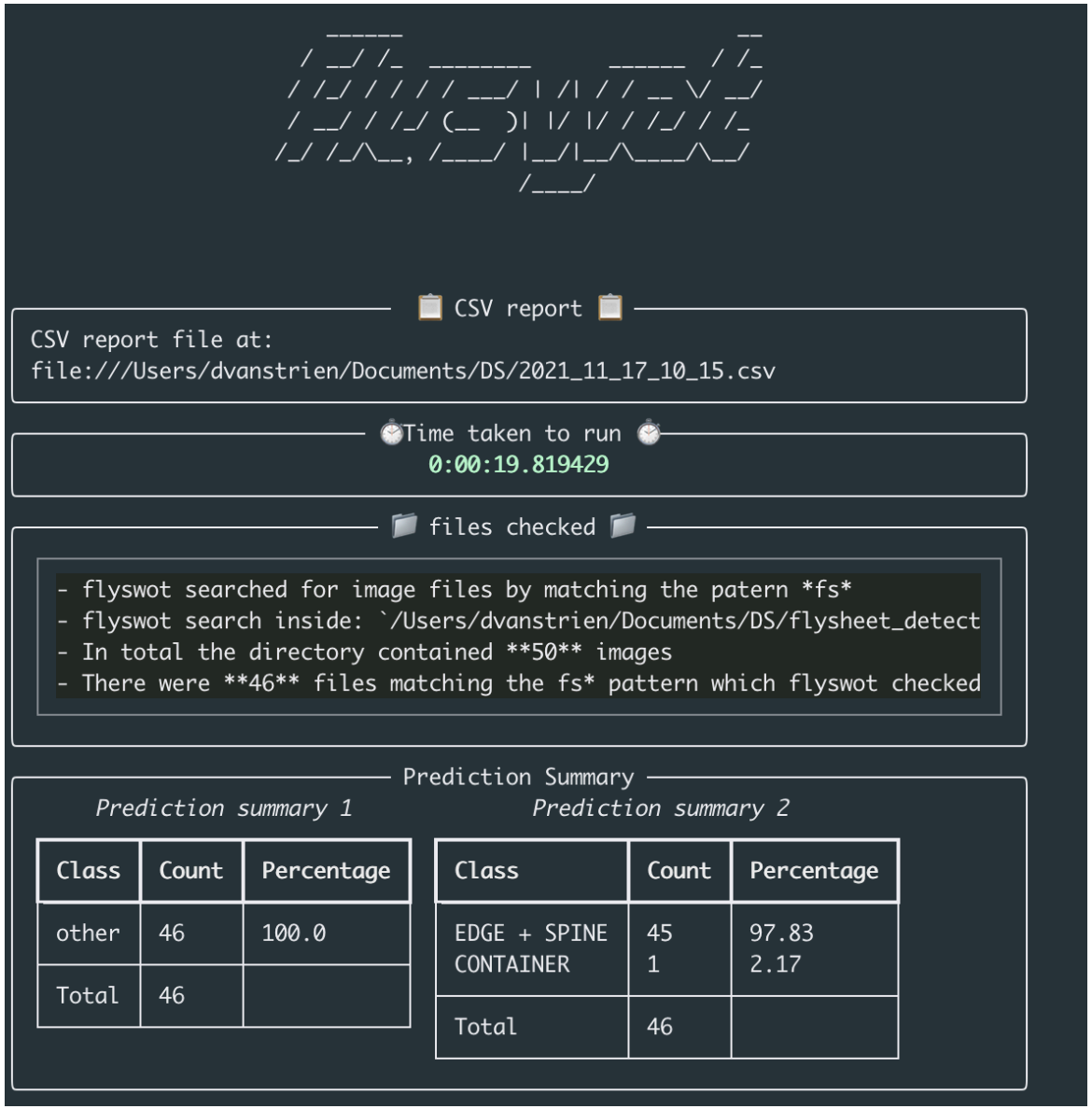
How do we identify which images are ‘fake flysheets’?

Using computer vision to detect fake flysheets

- Created training data by putting images into folders (integrating into existing workflows)
- Trained ‘boring’ CV models (ResNets)
- DVC and other tools used to make pipeline (semi) reproducible when new training data is available

flyswot

- Command Line Application used to ‘deploy’ our models
- Searches for images with a flysheet filename
- Predicts whether they are an actual flysheet or ‘other’ via a CSV report



BRITISH LIBRARY

Incorporating flyswot into the workflow

- Silver bullet!
- Unfortunately we have a large and varied collection
- flysheet vs other
- Interpret results using Power BI
- Allows us to assess the state of the image set and the amount of resource required

TIFF prediction 1

1 flysheet vs other	Total
flysheet	434
other	58
Total	**492**

TIFF prediction 2

2 label category	Total
flysheet	446
page + folio	28
cover	10
other	4
container	2
edge + spine	2
Total	**492**

Number of items analysed which contain flysheets

48

Total with "other" or flysheet < 80%

37

Interpreting the results

- Allows us to focus on the items with fake flysheets

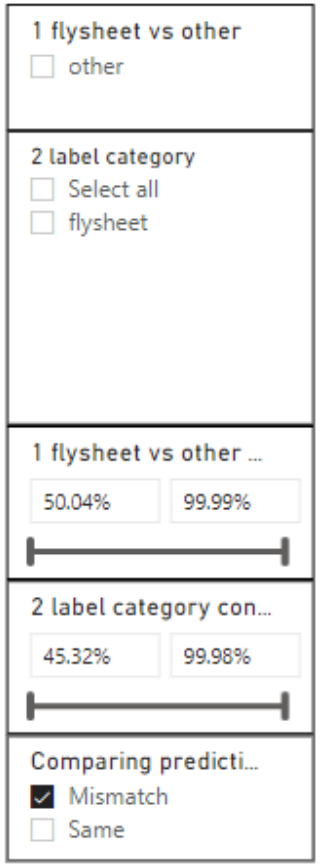
Items and files = "other" and/or flysheet with less than 80% confidence

directory	1 flysheet vs other	confidence	2 label category	confidence
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_10302\Tiff				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_10302\Tiff\add_ms_10302_fs001r.tif	other	91.08%	page + folio	71.10%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_11414\tiffs				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_11414\tiffs\add_ms_11414_fs001r.tif	flysheet	68.07%	flysheet	90.01%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_11414\tiffs\add_ms_11414_fse001r.tif	flysheet	50.88%	flysheet	83.79%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_15705\tiffs				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_15705\tiffs\add_ms_15705_fse007r.tif	other	95.38%	page + folio	94.66%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18144\Tiffs 300 ppi				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18144\Tiffs 300 ppi\add_ms_18144_fs001r.tif	flysheet	76.97%	flysheet	85.57%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18144\Tiffs 300 ppi\add_ms_18144_fs001v.tif	other	93.15%	other	48.20%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18144\Tiffs 300 ppi\add_ms_18144_fse005r.tif	other	99.93%	edge + spine	95.56%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18719\Tiffs 300 ppi				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18719\Tiffs 300 ppi\add_ms_18719_fs001r.tif	other	98.79%	cover	97.64%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18719\Tiffs 300 ppi\add_ms_18719_fse003r.tif	other	90.11%	page + folio	79.97%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_18719\Tiffs 300 ppi\add_ms_18719_fse012v.tif	other	99.53%	cover	92.90%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_19553\Tiff				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_19553\Tiff\add_ms_19553_fs001v.tif	other	99.85%	page + folio	99.02%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_21247\Tiff				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_21247\Tiff\add_ms_21247_fs004v.tif	other	75.18%	page + folio	62.36%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_25435\tiffs				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_25435\tiffs\add_ms_25435_fse002v.tif	flysheet	71.88%	flysheet	92.67%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_27695\tiffs				
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_27695\tiffs\add_ms_27695_fs002r.tif	other	97.07%	page + folio	96.34%

Continuous improvement

- Highlights any unexpected predictions

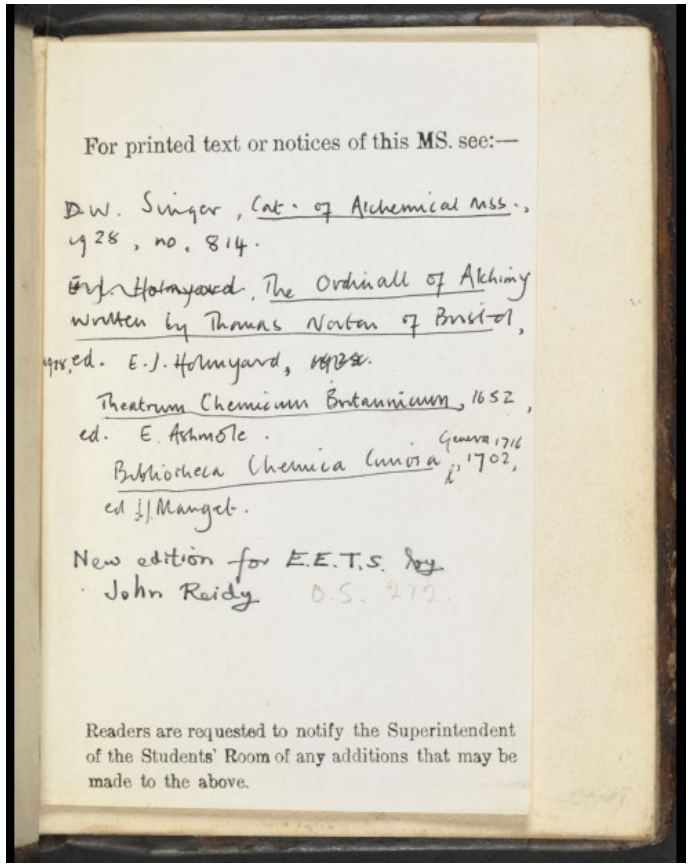
file	1 flysheet vs other	confidence	2 label category	confidence
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_27695\tiffs\add_ms_27695_fse007v.tif	other	63.67%	flysheet	49.16%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_65194\Tiff\add_ms_65194_fse001v.tif	other	50.83%	flysheet	88.31%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_65194\Tiff\add_ms_65194_fse002r.tif	other	86.22%	flysheet	57.86%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\add_ms_89309\300ppi tiffs\add_ms_89309_fs001r.tif	other	74.94%	flysheet	61.65%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\cotton_ms_tiberius_a_vii!1\tiffs\cotton_ms_tiberius_a_vii!1_fs001r.tif	other	50.04%	flysheet	76.67%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\egerton_ms_2709\tiffs\egerton_ms_2709_fs003r.tif	other	51.47%	flysheet	59.65%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\egerton_ms_3018\tiffs\egerton_ms_3018_fs001r.tif	other	87.07%	flysheet	54.23%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\harley_ms_6205\tiffs\harley_ms_6205_fse004v.tif	other	57.01%	flysheet	50.62%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\royal_ms_17_a_xliv\tiffs\royal_ms_17_a_xliv_fs001v.tif	other	71.32%	flysheet	77.77%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\royal_ms_17_a_xliv\tiffs\royal_ms_17_a_xliv_fs002v.tif	other	75.53%	flysheet	81.42%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\royal_ms_17_a_xliv\tiffs\royal_ms_17_a_xliv_fse003v.tif	other	73.33%	flysheet	81.51%
\\ad\collections\Preservation Digitisation Bids 2014-15\Finished Manuscripts 2\royal_ms_17_a_xliv\tiffs\royal_ms_17_a_xliv_fse004v.tif	other	67.30%	flysheet	82.32%



- Refine the thresholds, give it examples to learn from and improve accuracy

Flexible and understanding

- One of many tools we use as part of an adaptable workflow
- Manual intervention is still required but this is desirable
- Building trust and confidence in machine learning



add_ms_10302_fs001r.tif | 3 | Flysheet | f. [i-r]

Thank you

BRITISH LIBRARY

Catherine.Cronin@bl.uk

Daniel.Van-Strien@bl.uk

Andrew.Longworth@bl.uk