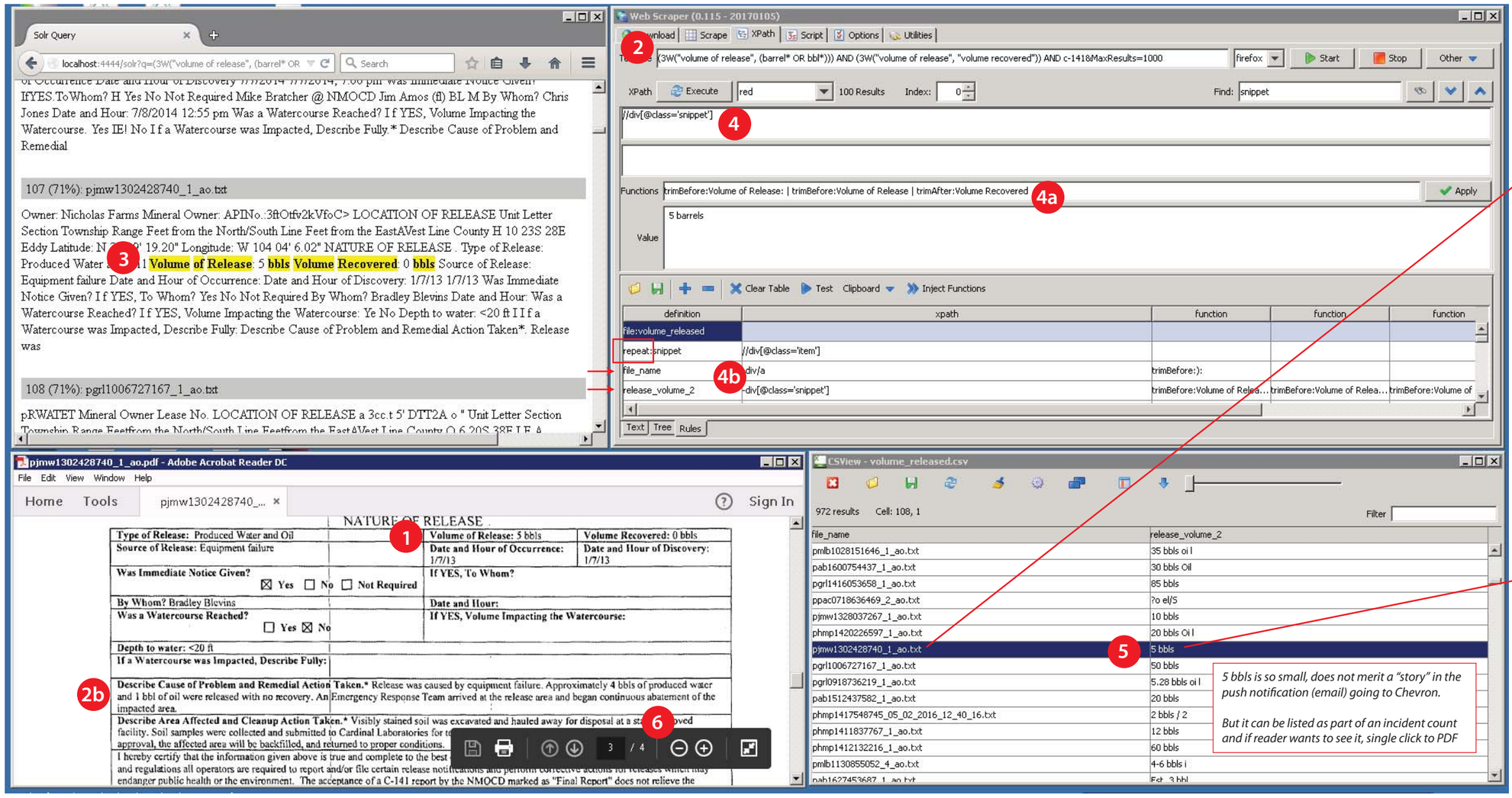
2a *within 3 words* *must have all criteria* *release we seek is in for c-141*

(3W("volume of release", (barrel* OR bbl*))) AND (3W("volume of release", "volume recovered")) AND C-141&MaxResults=1000

want release label and barrels to be close together *release and recovered need to be close also* *there may be other forms with same label*

this will be turned into a UI, so layperson does not need to know how to create Solr search strings



1. If we want to capture release volume, we ask Solr to search between the two volume (release/recovered) labels.

2. We construct a search to grab the contents between release and recovered label. In this case we also require a form of term barrels to be between the labels. We will do this for MCF also.

If we return a number (e.g. 5), we don't know if it's oil or gas. Don't want to take a chance. A null value is better than a wrong value.

2a. We will make it easy for a layperson to search for structured content. What is the left value (volume of release)? What is the right (volume recovered)? What must the results contain? (bbl* OR barrel*). Any other must-have criteria? C-141.

2b. Any user will be able to come in and via a UI grab data, e.g. the paragraph between remedial action taken and cleanup action taken. These will form narratives for touch-points (periodic alerts for executives).

4. Using the existing Scraper, we isolate each search result "snippet" returned by Lucene.

4a. We trim this snippet so all that is left is the content between the two (release and recovered) labels.

4b. Next, we build out the rule sheet. We want document name (to tie back to API) and the trimmed released volume. We want to repeat this scrape for all results (over 10K) from Lucene. We will grab all C-141 fields at one time.

5. We get structured data back in a list that can be used to populate a database table.

6. PDFs can be multiple pages, even multiple C-141 forms. These PDFs will be broken into single page PDFs for scraping.

7. Each PDF is scraped and converted to an XML file. This XML file is used to transform data into the database(s). The root element is the PDF filename. There is one element for each page (processed separately). Executive alerts can also be driven off XML.

7a. A layperson can do all the scrapping for C-141 forms with no programming knowledge. All 10K+ will be scraped at one time.

7b. If a <page> element does not have all required data, it will step higher in the XML until it finds relevant data. For example, the C-141 may not have API, but an earlier page might.

8. All company names can be parsed out of XML for "touchpoint" creation. Entire narratives can be sent to operator and/or third parties along with who at company handled.