VISUAL STEP-BY-STEP CASE STUDY: PATENT TROLLS

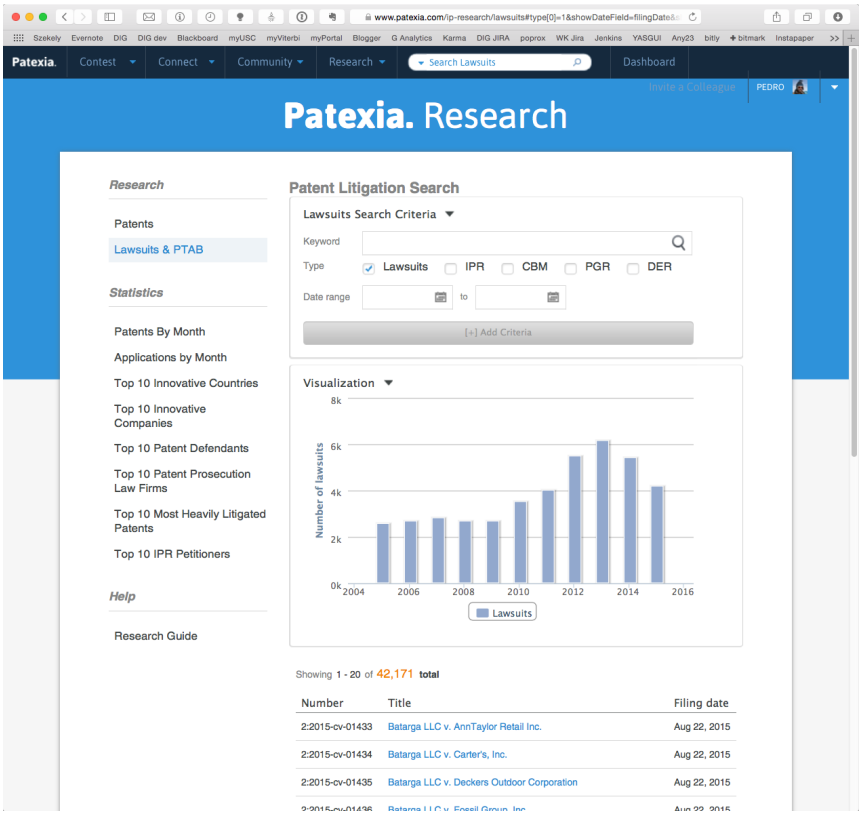
Let's apply the course concepts we have studied in an actual information extraction application

Opening prompt / problem statement

- Companies use patents to protect their intellectual property. Patent Trolls are companies that purchase patents and then threaten to sue many companies for patent infringement, looking to settle rather than go to court
    - Who are the patent trolls?
    - Which patents are being trolled?
    - Who are their victims?

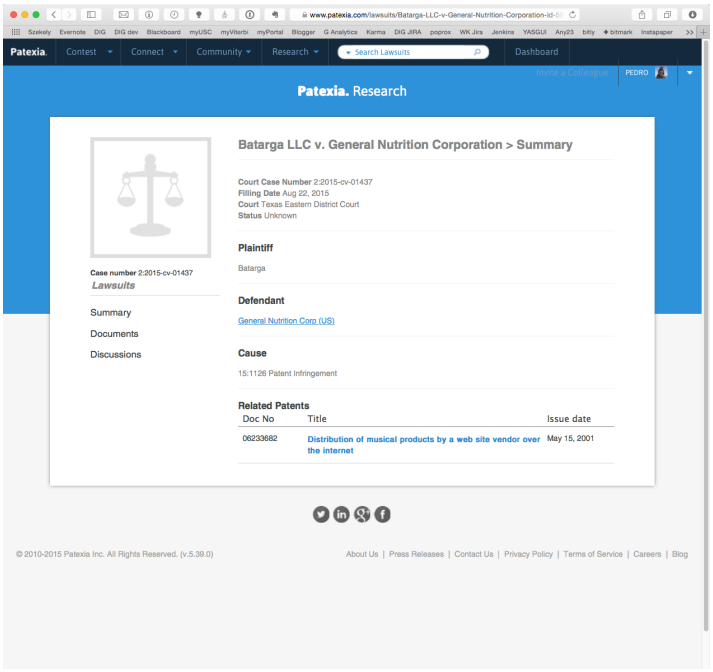
Context and opportunity

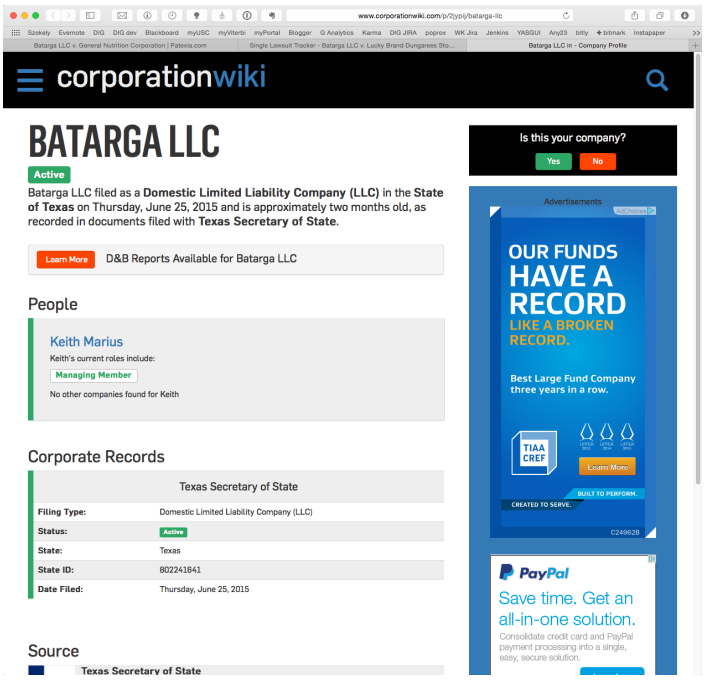
Patent research is a major application area



https://www.patexia.com/ip-research/lawsuits#type[0]=1&showDateField=filingDate&showCharts=1

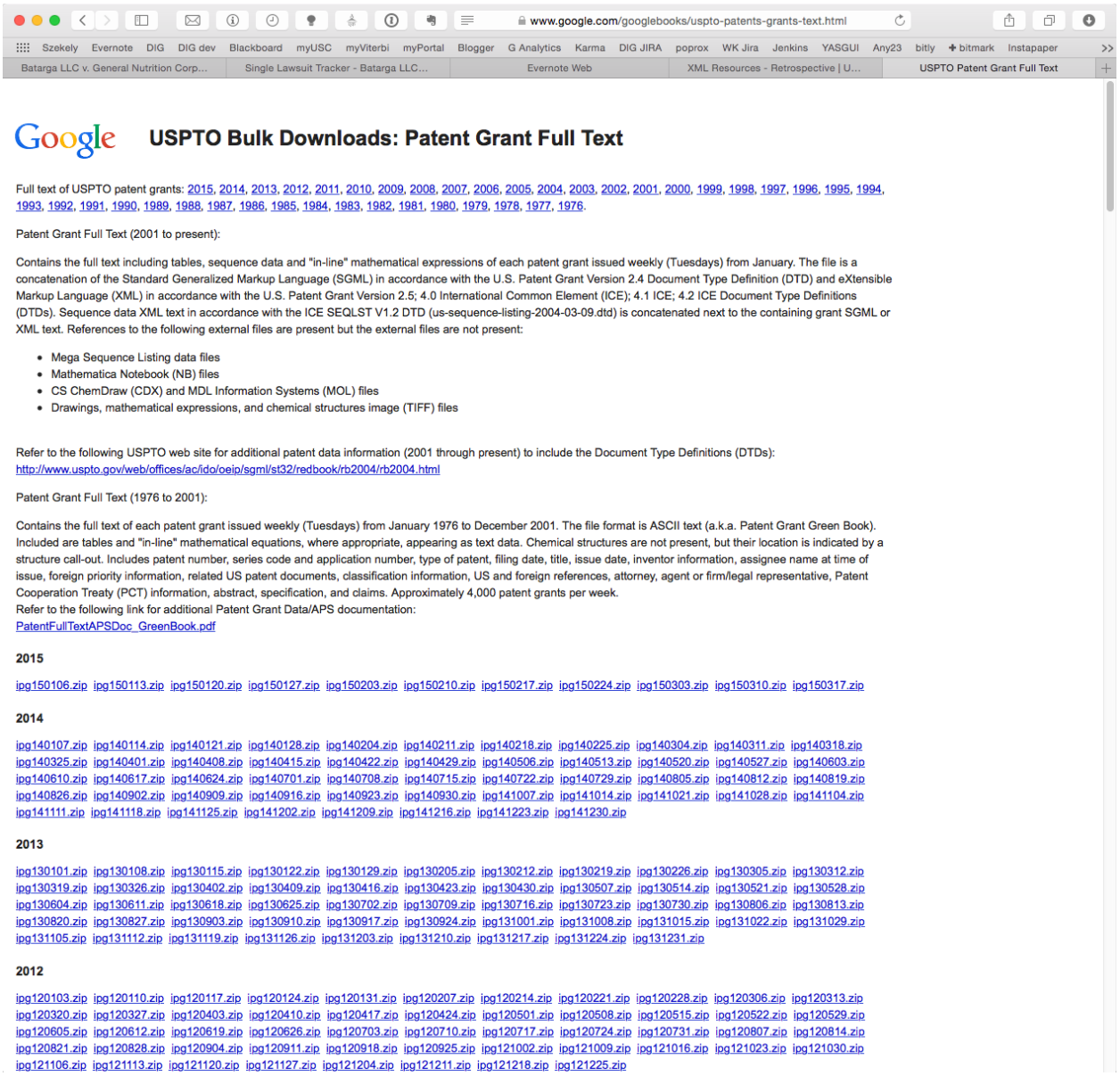
Example lawsuit





http://www.corporationwiki.com/p/2jypij/batarga-llc

Downloading US Patent Office patents in bulk



Google **USPTO Bulk Downloads: Patent Grant Full Text**

Full text of USPTO patent grants: 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000, 1999, 1998, 1997, 1996, 1995, 1994, 1993, 1992, 1991, 1990, 1989, 1988, 1987, 1986, 1985, 1984, 1983, 1982, 1981, 1980, 1979, 1978, 1977, 1976.

Patent Grant Full Text (2001 to present):

Contains the full text including tables, sequence data and "in-line" mathematical expressions of each patent grant issued weekly (Tuesdays) from January. The file is a concatenation of the Standard Generalized Markup Language (SGML) in accordance with the U.S. Patent Grant Version 2.4 Document Type Definition (DTD) and eXtensible Markup Language (XML) in accordance with the U.S. Patent Grant Version 2.5; 4.0 International Common Element (ICE); 4.1 ICE; 4.2 ICE Document Type Definitions (DTDs). Sequence data XML text in accordance with the ICE SEQLST V1.2 DTD (us-sequence-listing-2004-03-09.dtd) is concatenated next to the containing grant SGML or XML text. References to the following external files are present but the external files are not present:

- Mega Sequence Listing data files
- Mathematica Notebook (NB) files
- CS ChemDraw (CDX) and MDL Information Systems (MOL) files
- Drawings, mathematical expressions, and chemical structures image (TIFF) files

Refer to the following USPTO web site for additional patent data information (2001 through present) to include the Document Type Definitions (DTDs):
http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/redbook/rb2004/rb2004.html

Patent Grant Full Text (1976 to 2001):

Contains the full text of each patent grant issued weekly (Tuesdays) from January 1976 to December 2001. The file format is ASCII text (a.k.a. Patent Grant Green Book). Included are tables and "in-line" mathematical equations, where appropriate, appearing as text data. Chemical structures are not present, but their location is indicated by a structure call-out. Includes patent number, series code and application number, type of patent, filing date, title, issue date, inventor information, assignee name at time of issue, foreign priority information, related US patent documents, classification information, US and foreign references, attorney, agent or firm/legal representative, Patent Cooperation Treaty (PCT) information, abstract, specification, and claims. Approximately 4,000 patent grants per week.
Refer to the following link for additional Patent Grant Data/APS documentation:
PatentFullTextAPSDoc_GreenBook.pdf

2015

ipg150106.zip ipg150113.zip ipg150120.zip ipg150127.zip ipg150203.zip ipg150210.zip ipg150217.zip ipg150224.zip ipg150303.zip ipg150310.zip ipg150317.zip

2014

ipg140107.zip ipg140114.zip ipg140121.zip ipg140128.zip ipg140204.zip ipg140211.zip ipg140218.zip ipg140225.zip ipg140304.zip ipg140311.zip ipg140318.zip
ipg140325.zip ipg140401.zip ipg140408.zip ipg140415.zip ipg140422.zip ipg140429.zip ipg140506.zip ipg140513.zip ipg140520.zip ipg140527.zip ipg140603.zip
ipg140610.zip ipg140617.zip ipg140624.zip ipg140701.zip ipg140708.zip ipg140715.zip ipg140722.zip ipg140729.zip ipg140805.zip ipg140812.zip ipg140819.zip
ipg140826.zip ipg140902.zip ipg140909.zip ipg140916.zip ipg140923.zip ipg140930.zip ipg141007.zip ipg141014.zip ipg141021.zip ipg141028.zip ipg141104.zip
ipg141111.zip ipg141118.zip ipg141125.zip ipg141202.zip ipg141209.zip ipg141216.zip ipg141223.zip ipg141230.zip

2013

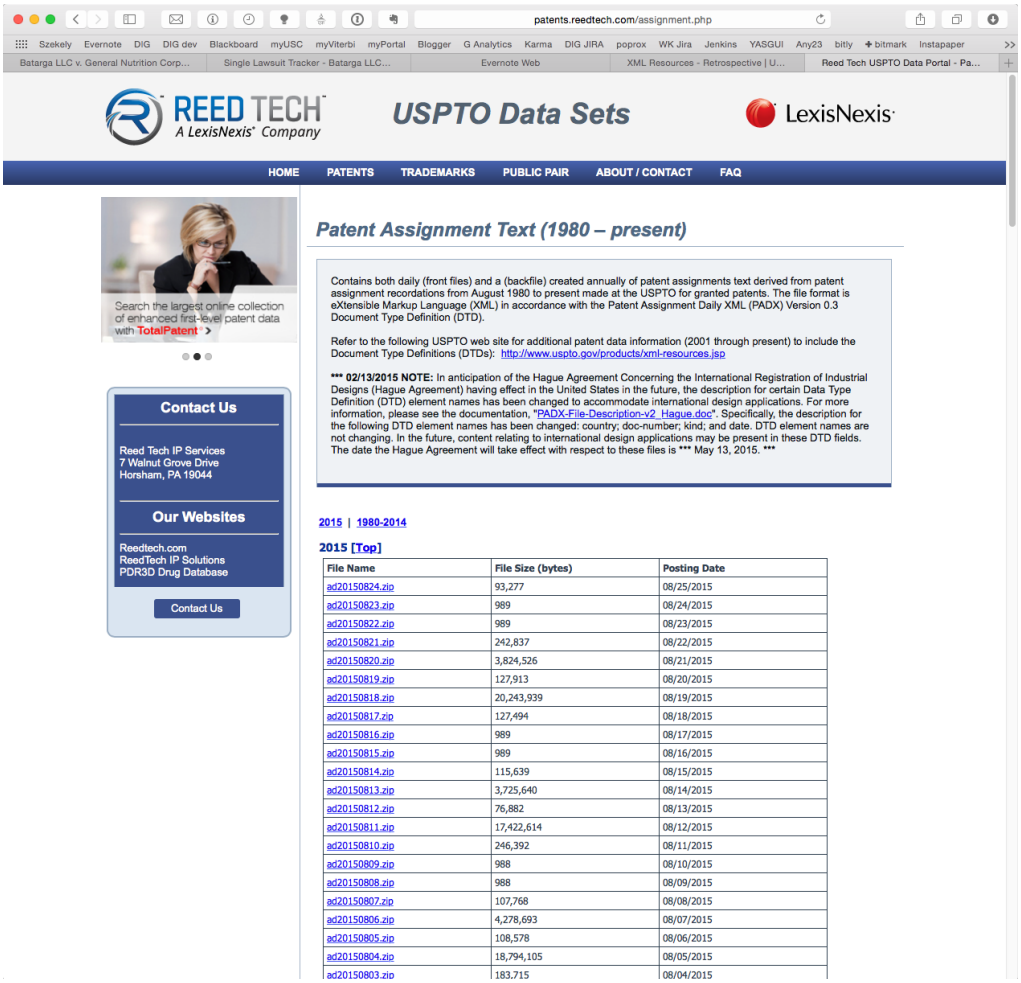
ipg130101.zip ipg130108.zip ipg130115.zip ipg130122.zip ipg130129.zip ipg130205.zip ipg130212.zip ipg130219.zip ipg130226.zip ipg130305.zip ipg130312.zip
ipg130319.zip ipg130326.zip ipg130402.zip ipg130409.zip ipg130416.zip ipg130423.zip ipg130430.zip ipg130507.zip ipg130514.zip ipg130521.zip ipg130528.zip
ipg130604.zip ipg130611.zip ipg130618.zip ipg130625.zip ipg130702.zip ipg130709.zip ipg130716.zip ipg130723.zip ipg130730.zip ipg130806.zip ipg130813.zip
ipg130820.zip ipg130827.zip ipg130903.zip ipg130910.zip ipg130917.zip ipg130924.zip ipg131001.zip ipg131008.zip ipg131015.zip ipg131022.zip ipg131029.zip
ipg131105.zip ipg131112.zip ipg131119.zip ipg131126.zip ipg131203.zip ipg131210.zip ipg131217.zip ipg131224.zip ipg131231.zip

2012

ipg120103.zip ipg120110.zip ipg120117.zip ipg120124.zip ipg120131.zip ipg120207.zip ipg120214.zip ipg120221.zip ipg120228.zip ipg120306.zip ipg120313.zip
ipg120320.zip ipg120327.zip ipg120403.zip ipg120410.zip ipg120417.zip ipg120424.zip ipg120501.zip ipg120508.zip ipg120515.zip ipg120522.zip ipg120529.zip
ipg120605.zip ipg120612.zip ipg120619.zip ipg120626.zip ipg120703.zip ipg120710.zip ipg120717.zip ipg120724.zip ipg120731.zip ipg120807.zip ipg120814.zip
ipg120821.zip ipg120828.zip ipg120904.zip ipg120911.zip ipg120918.zip ipg120925.zip ipg121002.zip ipg121009.zip ipg121016.zip ipg121023.zip ipg121030.zip
ipg121106.zip ipg121113.zip ipg121120.zip ipg121127.zip ipg121204.zip ipg121211.zip ipg121218.zip ipg121225.zip

https://www.google.com/googlebooks/uspto-patents-grants-text.html

http://patents.reedtech.com/assignment.php



Case Questions

1) The last visualization shows that the bulk patent dataset is divided up into many files. How would you go about automatically downloading all of those using a script?
2) Patent files contain images, abstracts and text, among other elements. What aspects are expected to be useful for this case?
3) Draw an information integration pipeline (the class diagram is a guide only) and specify the technologies you will use at each stage of the pipeline.
4) Recall the three original questions we specified in the prompt. Using the material provided in the case study, and the architecture you drew in Q3, how would you go about providing decision support for those questions to an interested user?
5) Is a traditional database (i.e. relational database) or knowledge graph infrastructure more appropriate for this problem? Why or why not?