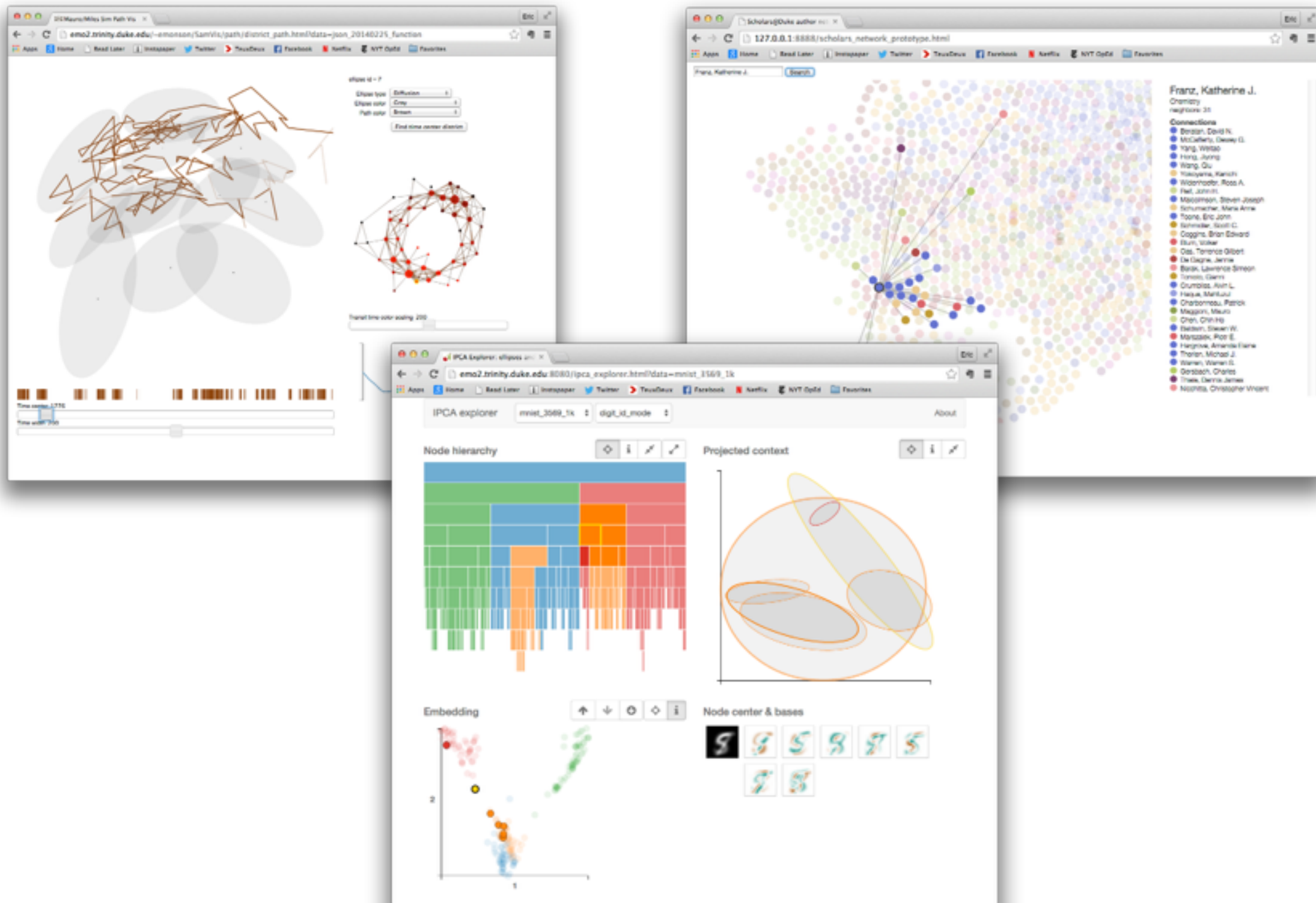


Putting a spit shine on the Getty Provenance Index

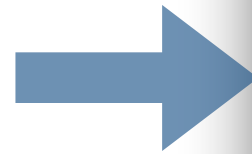
Eric E Monson

Visualization & Interactive Systems

Other work: Interactive web-based visualizations backed by lightweight data servers



DALMI: studying large-scale art markets



18th_cent_sales_contents_20150115.xlsx

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

Edit Font Alignment Number Format Cells Themes

Paste Calibri (Body) 12 B I U A

Align General Conditional Formatting Styles Actions Themes

K99 Rembrandt

	J	K	L	M	N	O	P	Q	R
82	202	Vander Neer	Neer, A.	NEER, AERT VAN DER	Dutch		Deux jolies vues de Village hollandais au		
83	201	Vander Does	Does	DOES	Dutch or Flemish		Un troupeau de différents animaux, gar		
84	200	Ezias Vandeveld	Velde, E. (I)	VELDE, ESAIAS VAN DE (Dutch		Un marché de campagne, avec divers gr		
85	199	Adrien Vandeveld	Velde, A.	VELDE, ADRIAEN VAN DE	Dutch		Un petit paysage plein d'esprit, où se tr		
86	198	Van Bloom	Bloemen	BLOEMEN	Flemish		Un Tartare monté sur un cheval qui cou		
87	197	Sarazin	Sarazin	SARAZIN	French or Swiss		Un paysage agréable & pittoresque, au		
88	196	Ruysdael	Ruisdael	RUISDAEL	Dutch		Une étude de plusieurs chaumières enti		
89	0195[c]	Pynacker	Pynacker	PYNACKER, ADAM	Dutch		Autre étude à l'encre de la Chine, de plu		
90	0195[b]	K. du Jardin	Dujardin, K.	DUJARDIN, KAREL	Dutch		Autre étude à l'encre de la Chine, de plu		
91	0195[a]	Ruysdael	Ruisdael	RUISDAEL	Dutch		Autre étude à l'encre de la Chine, de plu		
92	194	Ruysdael	Ruisdael	RUISDAEL	Dutch		Autre paysage, étude d'arbres & brouss		
93	193	Ruysdael	Ruisdael	RUISDAEL	Dutch		Un charmant Paysage, où se voit au mili		
94	192	Henry Roos	Roos, J.H.	ROOS, JOHANN HEINRIC	German		Un groupe de deux figures & dix animau		
95	191	Roland Roghman	Roghman, R.	ROGHMAN, ROELANT	Dutch		Deux petits paysages à l'encre de la Chir		
96	190	Roland Roghman	Roghman, R.	ROGHMAN, ROELANT	Dutch		Un paysage très pittoresque, & fait avec		
97	189	Marco Ricci	Ricci, M.	RICCI, MARCO	Italian		Un rocher mouillé par le pied d'une rivie		
98	188	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Tobie endormi; près de lui est l'Ange qu		
99	187	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Vingt quatre idem, dont Agar répudiée,		
100	186	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Seize idem, dont l'Ange devant Tobie, &		
101	185	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Dix sept idem, dont la Décolation de Sai		
102	184	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Quinze idem, dont Loth & ses Filles, Tob		
103	183	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Douze idem au pinceau trempé dans le		
104	182	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Quatre croquis, dont une Résurrection,		
105	181	Rembrandt	Rembrandt	REMBRANDT HARMENS	Dutch		Le bon Samaritain, sujet en travers, d'ur		
106	180	Raphael	Raffaello San	RAFFAELLO SANTI	Italian		Un groupe de trois têtes de femme & er		
107	179	Rademaker	Rademaker	RADEMAKER	Dutch		La vue d'un gros Village hollandais: trav		
108	178	Pierre	Pierre, J.B.M	PIERRE, JEAN-BAPTISTE	French		Une tête de Vieillard à barbe, vue de pr		
109	177	Piazzetta	Piazzetta	PIAZZETTA, GIOVANNI B	Italian		Une belle tête de Vieillard à barbe & chi		
110	176	Perignon	Perignon, N.	PERIGNON, NICOLAS	French		Un très joli Paysage bordé d'un grand Ci		
111	175	Perignon	Perignon, N.	PERIGNON, NICOLAS	French		Deux bouquets de différentes fleurs ble		
112	174	Parrocel	Parrocel	PARROCEL	French		Deux petits corps de garde & halte de S		
113	173	Parrocel	Parrocel	PARROCEL	French		Trois belles têtesw de Soldats, à la pierr		
114	172	Parrocel	Parrocel	PARROCEL	French		Douze croquis de compositions pour Ba		
115	171	Parrocel	Parrocel	PARROCEL	French		Un autre sujet pareillement exécuté, où		
116	170	Parrocel	Parrocel	PARROCEL	French		Deux sujets exécutés à la sanguine avec		
117	169	Parrocel	Parrocel	PARROCEL	French		Un groupe de six Cavaliers combattant e		
118	168	Parrocel	Parrocel	PARROCEL	French		Cinq quintes & études, dont une taboula		

Normal View Ready

Getty Provenance Index



The screenshot shows a web browser window with the address bar displaying www.getty.edu/research/tools/provenance/search.html. The page header includes the text "Explore the Getty" and "Connect with Us | Shop". The main heading is "The Getty Research Institute". Below this is a navigation bar with links: "Exhibitions & Events", "Special Collections", "Library", "Search Tools & Databases", "Scholars & Projects", "Publications", and "About the GRI".

The "Search Tools & Databases" section is active, showing a sidebar with a list of search tools and databases. The main content area is titled "Search the Getty Provenance Index® Databases" and contains the following text:

This vast collection of digital records is expanded and enriched on a regular basis. The quantity and scope of research material that is available varies by region, period, and type of document. The databases currently include the following types of records:

- Archival Inventories**
Archival inventories are legal documents from private and public archives that list objects from a household. The inventories in this database list works of art from private collections in the Netherlands, Italy, Spain, and France from 1550 to 1840. The Inventory Descriptions section of this database contains 5,200 documents. From these documents, more than 270,000 individual records have been entered in the Inventory Contents section. Photocopies of most of these documents can be found in the [Collectors Files](#).
- Sales Catalogs**
Typically published by auction houses and dealers, sales catalogs list works of art for public auction. This database includes catalogs from major cities in Belgium, France, Germany, Great Britain, the Netherlands, and Scandinavia from 1650 to 1945. The Sale Descriptions section of this database contains more than 15,000 catalogs. From these catalogs, more than one million records have been entered into the Sale Contents section. This database also contains private contract sales through which collectors were able to acquire artworks during an extended period of exhibition. Photocopies of most of these documents can be found in the [Sales Catalogs Files](#).

Additionally, bibliographic information on more than 2,000 [German Sales Catalogs](#) from 1930 to 1945 is available in the Sale Descriptions section. More than 230,000 individual auction sales records for paintings, sculptures, and drawings have been extracted from these catalogs, and each record is linked to the full PDF of its corresponding catalog residing at the website of the Heidelberg University Library.

The sidebar on the left lists the following search tools and databases:

- Primo Search
- Getty Research Portal
- Collection Inventories & Finding Aids
- Photo Archive
- Research Guides & Bibliographies
- Digital Collections
- Article & Research Databases
- Collecting & Provenance Research
 - Search the Databases
 - Getty Provenance Index Database
 - Dealer Stock Books
 - Payments to Artists Database
 - See What's Covered
 - Using the Databases
 - Collaborators & Partners
 - Collectors Files
 - German Sales Catalogs, 1930-1945
 - Sales Catalogs Files
 - Events Related to the History of Collecting
- BHA & RILA
- Getty Vocabularies

On the right side of the page, there are sections for "See What's New:", "Inside Perspective", "Related Research Projects:", and "Related Research Guides:". The "See What's New:" section includes a link to "Dealer Stock Books (over 67,000 records)". The "Inside Perspective" section includes a link to "How and why the Dealer Stock Books database was created". The "Related Research Projects:" section includes links to "British Sales", "German Sales", and "Display of Art in Roman Palaces". The "Related Research Guides:" section includes a link to "German Sales Catalogs".

Getty Provenance Index

1.5 million records in multiple databases

- Archival Inventories
- Sales Catalogues
- Dealer Stock Books
- Payments to Artists
- Public Collections

Getty Provenance Index

The screenshot shows a web browser window with the address bar displaying `piprod.getty.edu/starweb/pi/servlet.starweb`. The page title is "The Getty Provenance Index® Databases". The navigation bar includes links for "Research Home", "Search Tools & Databases", "Collecting & Provenance Research", and "Provenance Databases". Below the navigation bar, there are three main sections: "Archival Inventories" (with links for "Search Inventory Contents" and "Search Inventory Descriptions"), "Sales Catalogs" (with links for "Search Sale Contents" and "Search Sale Descriptions"), and "Public Collections" (with links for "Search Public Collections" and "Search Provenance of Paintings"). A button labeled "Exit & Logout" is also present.

Records retrieved:

☒ match all words ☐ match any word

(Optional) Refine your search using the fields below.

Artist Name (browse):	<input type="text"/>	(e.g., Jacob Jordaens or Cuyp or Seghers; Albani)
Artist Nationality:	<input type="text"/>	(e.g., German)
Lot Title/Description:	<input type="text"/>	(e.g., Venus or batailles)
Object Type:	<input type="text"/>	
Lot Sale Date or Range:	<input type="text"/> through <input type="text"/>	(e.g., 1751 04 23 or 1839 10* or 1930 through 1932)
Buyer or Seller Name:	<input type="text"/>	(e.g., Pembroke)
Transaction Type:	<input type="text"/>	
Auction House:	<input type="text"/>	(e.g., Christie's or Lebrun)
City of Sale:	<input type="text"/>	(e.g., Berlin)
Country of Sale:	<input type="text"/>	
Lugt #:	<input type="text"/>	(e.g., 6410)
Catalog #:	<input type="text"/>	(e.g., D-108)
Lot #:	<input type="text"/>	(e.g., 0002)

Subject searches will further limit your results to British 18th-c. and German 20th-c. sales only:

Subjects: (use ctrl-click to make multiple selections)

Allegory/Allegorie
Animals/Tiere

[Overview](#) • [About the Databases](#) • [Contact Us](#)

Sales Catalogue DB designed for web display (although now downloadable w/limits)

The screenshot shows a web browser window with the address bar displaying `piprod.getty.edu/starweb/pi/servlet.starweb`. The page title is "The Getty Provenance Index® Databases". The navigation bar includes links for "Research Home", "Search Tools & Databases", "Collecting & Provenance Research", and "Provenance Databases". Below the navigation bar, there are three main sections: "Archival Inventories", "Sales Catalogs", and "Public Collections". The "Sales Catalogs" section is active, showing links for "Search Sale Contents" and "Search Sale Descriptions".

Below the navigation bar, there are several links: "Back to Search Results", "Back to Search", "Catalog Location Codes", "Previous Record", and "Next Record".

The main content area displays a search result for "Lot 0187 from Sale Catalog F-A458". The result is presented in a table-like format with the following fields:

Artist Name	REMBRANDT HARMENSZ. VAN RIJN (Dutch) <i>from catalog:</i> Rembrandt
Title / Description	Vingt quatre idem, dont Agar répudiée, Notre Seigneur en Jardinier, &c.
Object Type	Dessin
Seller	Bruny de La Tour d'Aigues, Jean-Pierre-Alexandre de <i>from catalog:</i> De La Tour Daigues
Transaction	Inconnue, 9 livres 1 sol pour les lots nos.187 & 198
Sale Date	1777 May 15 and following days (<i>This Lot:</i> May 15)
Expert	Basan (Pierre François)
Commissaire-Priseur	Hayot de Longpré (Philippe)
Sale Location	Paris, France
Lot Number	2694
See Also	Sale Description

At the bottom of the page, there are links for "Back to Search Results", "Back to Search", and "Catalog Location Codes". The footer contains links for "Overview", "About the Databases", and "Contact Us".

Data cleaning / wrangling – typical tasks

- Fix typos (*data entry or OCR errors*)
- Make data formats consistent (*e.g. dates*)
- Normalize names (*people, places, things*)

Getty Sales data is good for these (authority lists), but doesn't have artwork identification

- Granularity Problems

Make the data such that computers can use the entered values to sum, average, map, plot, etc.

Want only one type of data per column / field!

Sales Catalogues granularity problems

	A	B	C
1	Catalogue No	Lot Number	Artist Name 1
52	F-A338	0012[c]	[ANONYMOUS] (Unknown)
53	F-A338	13	VERMEER, JAN (II) (Dutch)
54	F-A338	0014[a]	LUYKEN, JAN (Dutch)
55	F-A338	0014[b]	SCHELLINKS (Dutch)
56	F-A338	0015[a]	VELDE, WILLEM VAN DE (Dutch)
57	F-A338	0015[b]	GOYEN, JAN JOSEPHSZ. VAN (Dutch)
58	F-A338	16	FARGUE, PAULUS CONSTANTIJN LA (Dutch)
59	F-A338	17	FARGUE, PAULUS CONSTANTIJN LA (Dutch)
60	F-A338	18	FARGUE, PAULUS CONSTANTIJN LA (Dutch)
61	F-A338	19	PATEL, PIERRE (I) (French)
62	F-A338	20	DROST, ANTHONIE (Dutch)
63	F-A338	21	BOULLONGNE (French)
64	F-A338	61	BISSCHOP, ABRAHAM (Dutch), copie par
65	F-A338	62	GUERCINO (GIOVANNI FRANCESCO BARBIERI) (Italian)
66	F-A338	63	MARATTI, CARLO (Italian)
67	F-A338	64	SNYDERS, FRANS (Flemish)
68	F-A338	65	FYT, JAN (Flemish)
69	F-A338	66	VELDE, ADRIAEN VAN DE (Dutch)
70	F-A338	67	OSTADE, ADRIAEN VAN (Dutch)
71	F-A338	68	HUYSUM, JAN VAN (Dutch)
72	F-A338	0069[a]	BRIL, PAUL (Flemish)
73	F-A338	0069[b]	BAUDOUIN, PIERRE ANTOINE (French)
74	F-A338	70	BRIL, PAUL (Flemish)
75	F-A338	71	HULST (Dutch or Flemish)
76	F-A338	72	HASTOLZ, J. C. (Dutch)
77	F-A338	73	NIEULANDT (Dutch)
78	F-A338	0074[a]	WILDENS, JAN (Flemish)
79	F-A338	0074[b]	VELDE, WILLEM VAN DE (Dutch)
80	F-A338	75	MAES, DIRK (Dutch)
81	F-A338	76	KIEFT, PIETER (Dutch)

Sales Catalogues granularity problems

U
Present Location
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Dec 09
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Dec 09
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Dec 09
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Dec 09
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Dec 09
1773 Dec 09
1773 Dec 09
1773 Dec 09
1773 Dec 09
1773 Dec 09
1773 Dec 09
1773 Apr 14 - 1773 Apr 21 (This Lot: Apr 14)
1773 Dec 09
1773 Dec 09
1773 Dec 09
1773 Dec 09

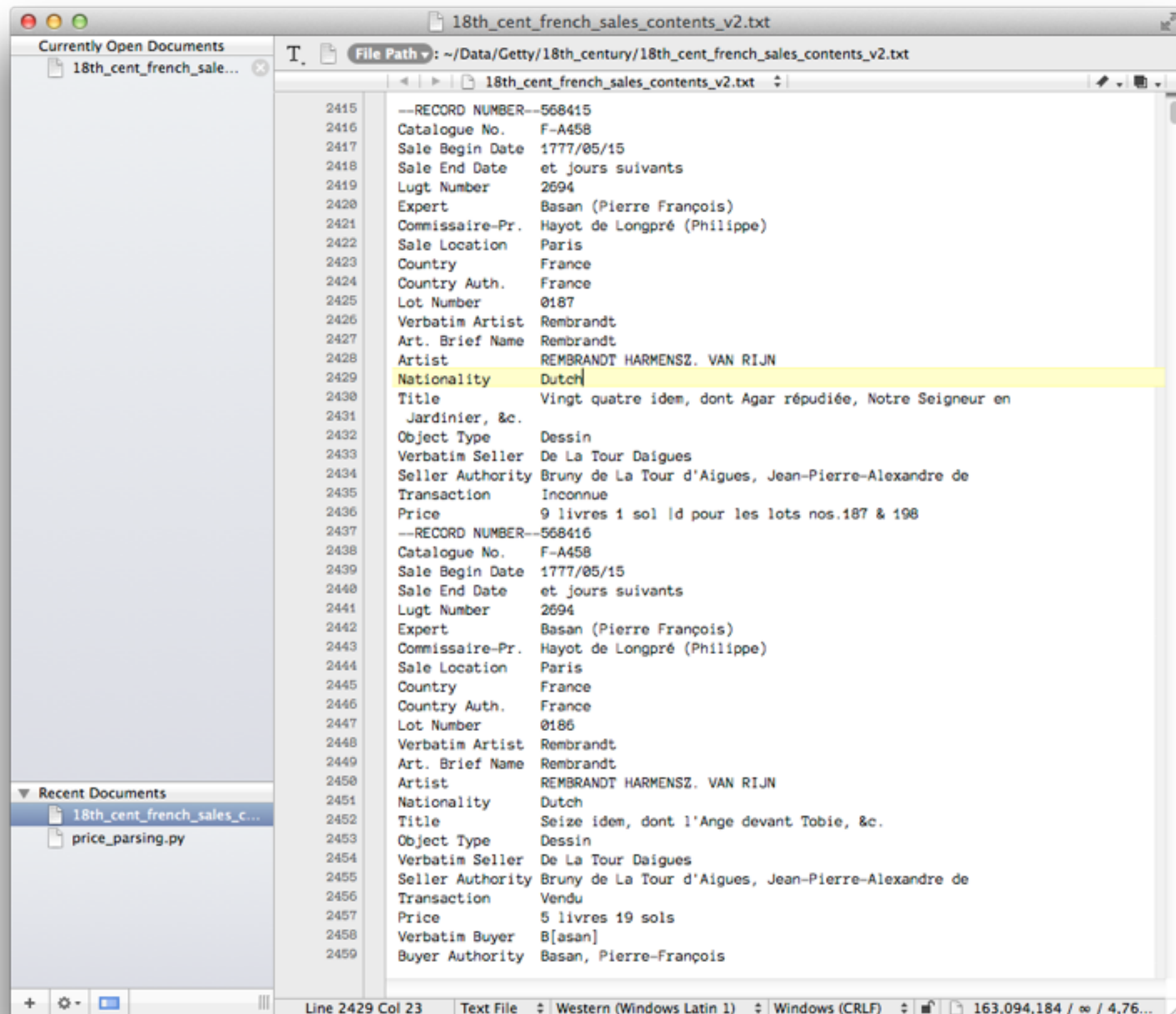
Sales Catalogues granularity problems

N
Transaction
Inconnue, 24 livres pour les lots 12[a-c]
Inconnue, 8 livres 1 sol
Inconnue, 7 livres 19 sols pour les lots nos. 14[a-b] & 15[a-b]
Inconnue, 7 livres 19 sols pour les lots nos. 14[a-b] & 15[a-b]
Inconnue, 7 livres 19 sols pour les lots nos. 14[a-b] & 15[a-b]
Inconnue, 7 livres 19 sols pour les lots nos. 14[a-b] & 15[a-b]
Inconnue, 5 livres 12 sols
Inconnue, 7 livres
Inconnue, 13 livres 12 sols
Inconnue, 13 livres 10 sols
Vendu, 12 livres 5 sols
Vendu, 19 livres 1
Vendu, 48 livres 19 sols
Vendu, 18 livres 19 sols
Vendu, 100 livres
Vendu, 48 livres
Vendu, 49 livres 19 sols
Vendu, 85 livres 1 sol
Vendu, 40 livres 4 sols
Vendu, 18 livres
Vendu, 16 livres 12 sols pour les lots 69[a] & [b]
Vendu, 16 livres 12 sols pour les lots 69[a] & [b]
Inconnue, 13 livres 4 sols
Vendu, 60 livres
Vendu, 100 livres 1 sol
Vendu, 10 livres 4 sols
Vendu, 37 livres 1 sol pour les lots 74[a] & [b]
Vendu, 37 livres 1 sol pour les lots 74[a] & [b]
Vendu, 60 livres 1 sol
Vendu, 36 livres 1 sol
Vendu, 36 livres
Vendu, 80 livres 1 sol
Vendu, 14 livres
Vendu, 13 livres 2 sols
Inconnue, 20 livres

“Raw text dump” has finer granularity

- 190k records for 18th Century French
- Directly from their STAR database
- Solves some granularity problems
- Needed a license agreement!

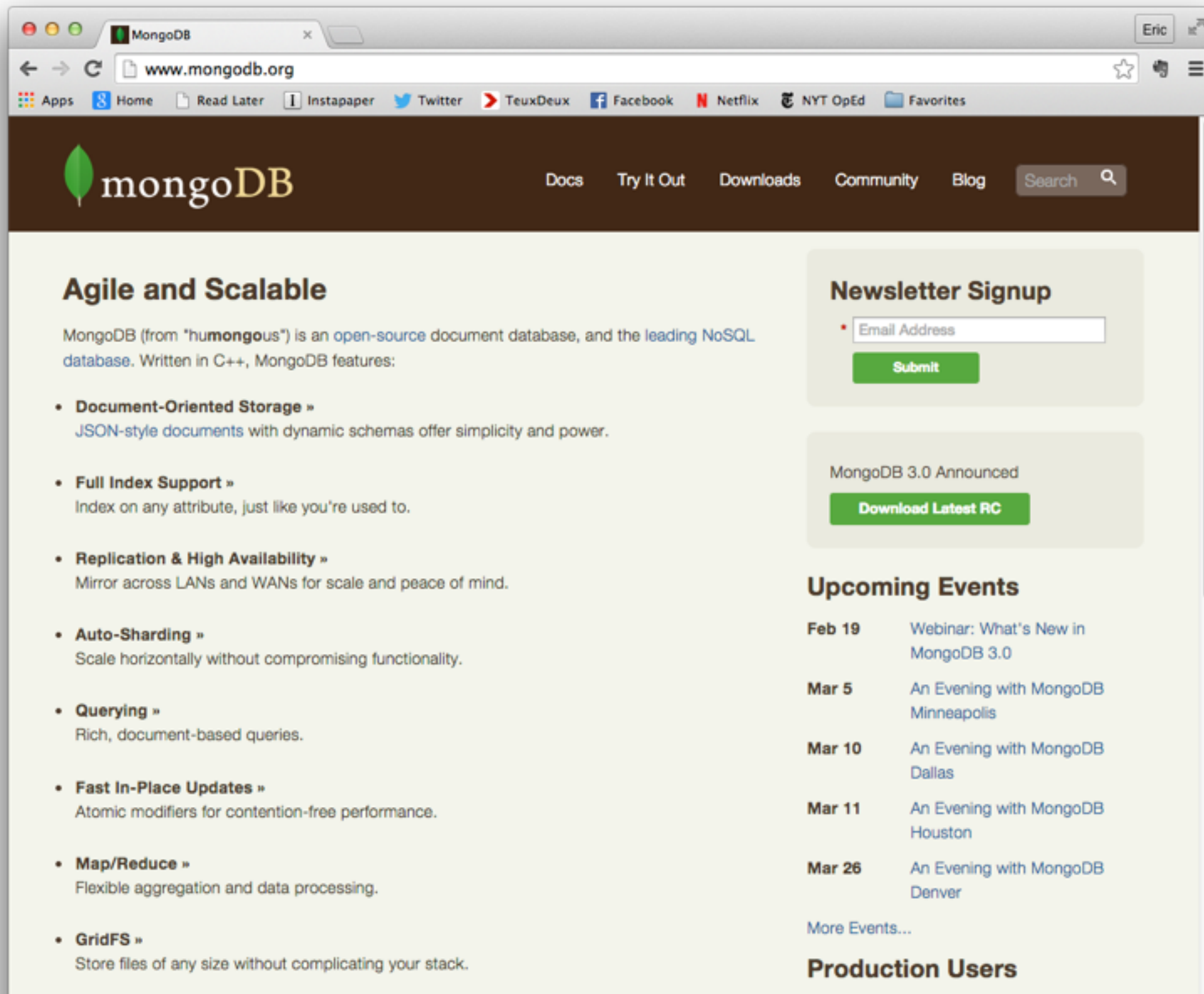
“Raw text dump” has finer granularity



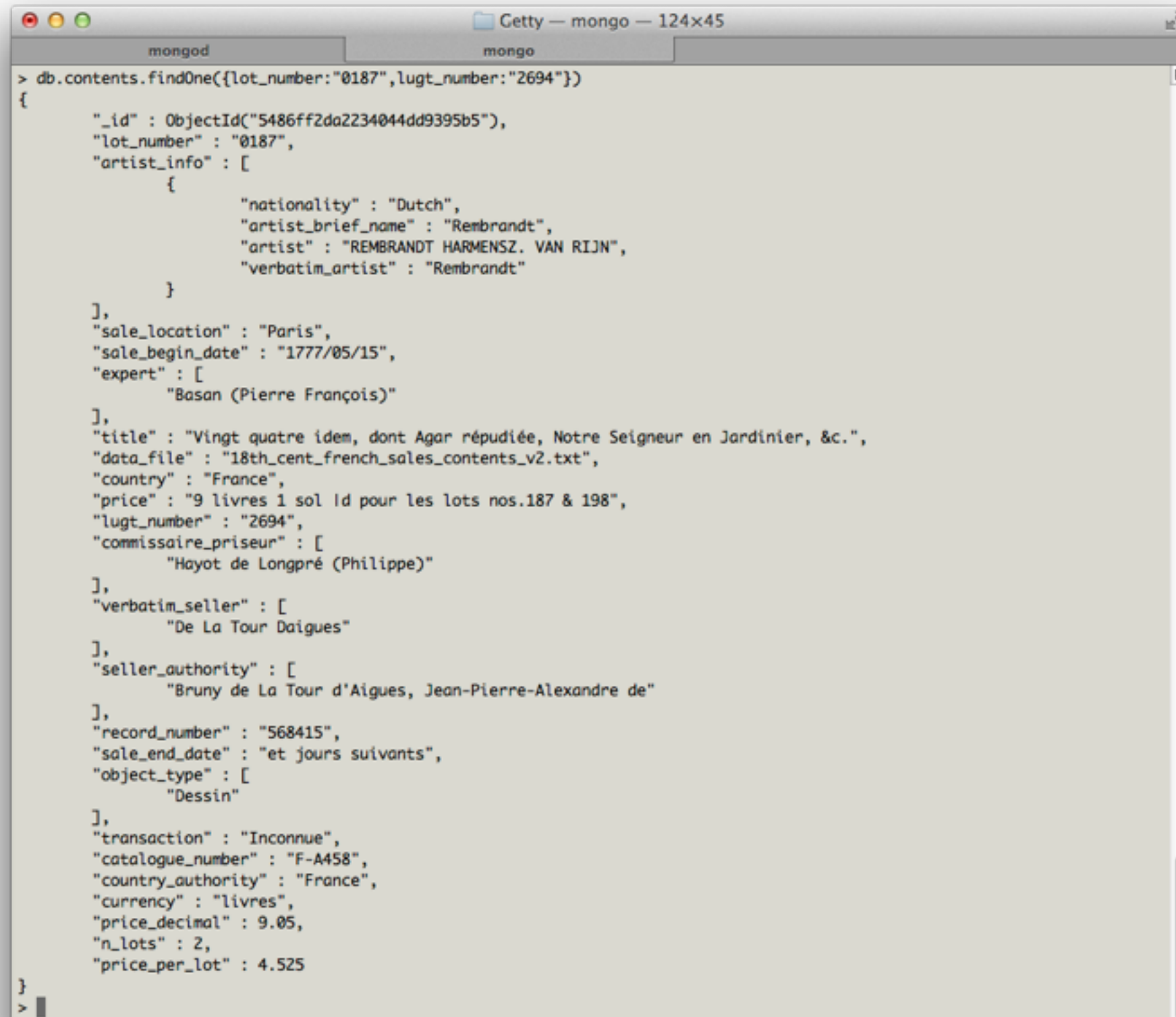
Raw text dump field descriptions

	A	B	C	D	E	
1	star_field_name	human_field_name	db_field_name	can_repeat	block	subfields_or_type
11	Commissaire-Pr.	Commissaire-Priseur	commissaire_priseur	yes		
12	Comm.-Pris.Auth.	Commissaire-Priseur Authority	commissaire_priseur_authority	yes		
13	Sale Location	Sale Location	sale_location			
14	Country	Country	country			
15	Country Auth.	Country Authority	country_authority			
16	C'ntry Authority	Country Authority	country_authority			
17	City Authority	City Authority	city_authority			
18	Lot Number	Lot Number	lot_number			int
19	Verbatim Artist	Verbatim Artist	verbatim_artist	yes	artist_info	
20	Art. Brief Name	Artist Brief Name	artist_brief_name	yes	artist_info	
21	Artist	Artist	artist	yes	artist_info	
22	Artist Authority	Artist Authority	artist_authority	yes	artist_info	
23	Nationality	Nationality	nationality	yes	artist_info	
24	Attribution Mod.	Attribution Modifier	attribution_modifier	yes	artist_info	
25	Birth Date	Birth Date	birth_date	yes	artist_info	date
26	Death Date	Death Date	death_date	yes	artist_info	date
27	Period Active	Period Active	period_active	yes	artist_info	
28	Century Active	Century Active	century_active	yes	artist_info	
29	Title	Title	title			
30	Title Modifier	Title Modifier	title_modifier			
31	Notes	Notes	notes			
32	Lot Notes	Lot Notes	lot_notes			
33	Handwritten Note	Handwritten Note	handwritten_note	yes		
34	Hand.Note Source	Handwritten Note Source	handwritten_note_source	yes		
35	Object Type	Object Type	object_type	yes		
36	Materials	Materials	materials			
37	Dimensions	Dimensions	dimensions			
38	Format	Format	format			
39	Inscription	Inscription	inscription			
40	Verbatim Seller	Verbatim Seller	verbatim_seller	yes		
41	Verb. Seller Mod	Verbatim Seller Modifier	verbatim_seller_modifier	yes		
42	Verb.Seller Mod.	Verbatim Seller Modifier	verbatim_seller_modifier	yes		
43	Seller Authority	Seller Authority	seller_authority	yes		name l location q questioned
44	Seller Auth.Mod.	Seller Authority Modifier	seller_authority_modifier	yes		
45	Transaction	Transaction	transaction			
46	Price	Price	price			price c currency d notes
47	Verbatim Buyer	Verbatim Buyer	verbatim_buyer	yes	buyer_info	
48	Verb. Buyer Mod.	Verbatim Buyer Modifier	verbatim_buyer_modifier	yes	buyer_info	
49	Buyer Authority	Buyer Authority	buyer_authority	yes	buyer_info	name l location q questioned
50	Buyer Auth. Mod.	Buyer Authority Modifier	buyer_authority_modifier	yes	buyer_info	
51	Previous Owner	Previous Owner	previous_owner	yes	prev_owner_info	
52	Prev.Own.Source	Previous Owner Source	previous_owner_source	yes	prev_owner_info	
53	Prev.Owner Auth.	Previous Owner Authority	previous_owner_authority	yes	prev_owner_info	name d date l location q questioned

Parsed raw text into MongoDB



Parsed raw text into MongoDB



The screenshot shows a MongoDB shell window with two tabs: 'mongod' and 'mongo'. The 'mongo' tab is active, displaying the result of a query. The query is `db.contents.findOne({lot_number:"0187",lugt_number:"2694"})`. The result is a JSON document representing a record from a French auction catalog.

```
> db.contents.findOne({lot_number:"0187",lugt_number:"2694"})
{
  "_id" : ObjectId("5486ff2da2234044dd9395b5"),
  "lot_number" : "0187",
  "artist_info" : [
    {
      "nationality" : "Dutch",
      "artist_brief_name" : "Rembrandt",
      "artist" : "REMBRANDT HARMENSZ. VAN RIJN",
      "verbatim_artist" : "Rembrandt"
    }
  ],
  "sale_location" : "Paris",
  "sale_begin_date" : "1777/05/15",
  "expert" : [
    "Basan (Pierre François)"
  ],
  "title" : "Vingt quatre idem, dont Agar répudiée, Notre Seigneur en Jardinier, &c.",
  "data_file" : "18th_cent_french_sales_contents_v2.txt",
  "country" : "France",
  "price" : "9 livres 1 sol 1d pour les lots nos.187 & 198",
  "lugt_number" : "2694",
  "commissaire_prieur" : [
    "Hayot de Longpré (Philippe)"
  ],
  "verbatim_seller" : [
    "De La Tour Daigues"
  ],
  "seller_authority" : [
    "Bruny de La Tour d'Aigues, Jean-Pierre-Alexandre de"
  ],
  "record_number" : "568415",
  "sale_end_date" : "et jours suivants",
  "object_type" : [
    "Dessin"
  ],
  "transaction" : "Inconnue",
  "catalogue_number" : "F-A458",
  "country_authority" : "France",
  "currency" : "livres",
  "price_decimal" : 9.05,
  "n_lots" : 2,
  "price_per_lot" : 4.525
}
```


Extract important information out of “price”

60 livres |d pour les lots 476 [a-b] & 695

Extract important information out of “price”

- Sometimes manual cleaning is best

The human brain is very good at noticing patterns and extracting the proper information out of text.

- 139k /188k records have a “price” entry

Computer as a tool to help do things more quickly and accurately, but you still have to wield it. It's still a craft, and it takes creativity to figure out how to put together the tools you have to do what you want.

Extract important information out of “price”

- OpenRefine for a lot of cleaning
Name normalization, data rearranging and simpler splitting or extraction of quantities (I have instructions for ArtNet data online)
- Here used Python and Natural Language Toolkit (NLTK)
Many functions for working with and analyzing human language data, (in my preferred coding language)

Note: I was lucky that the data set was this small. I could process each part of the data in seconds, so I didn't have to use a subset.

Finding patterns in text

(tokenization and tagging)

They refuse to permit us to obtain the refuse permit .

Finding patterns in text

(tokenization and tagging)

They	refuse	to	permit	us	to	obtain	the	refuse	permit	.
------	--------	----	--------	----	----	--------	-----	--------	--------	---

Finding patterns in text

(tokenization and tagging)

They	refuse	to	permit	us	to	obtain	the	refuse	permit	.
PRP	VBP	TO	VB	PRP	TO	VB	DT	NN	NN	.

Finding patterns in the price text (tokenization and tagging)

60 livres |d pour les lots 476 [a-b] & 695

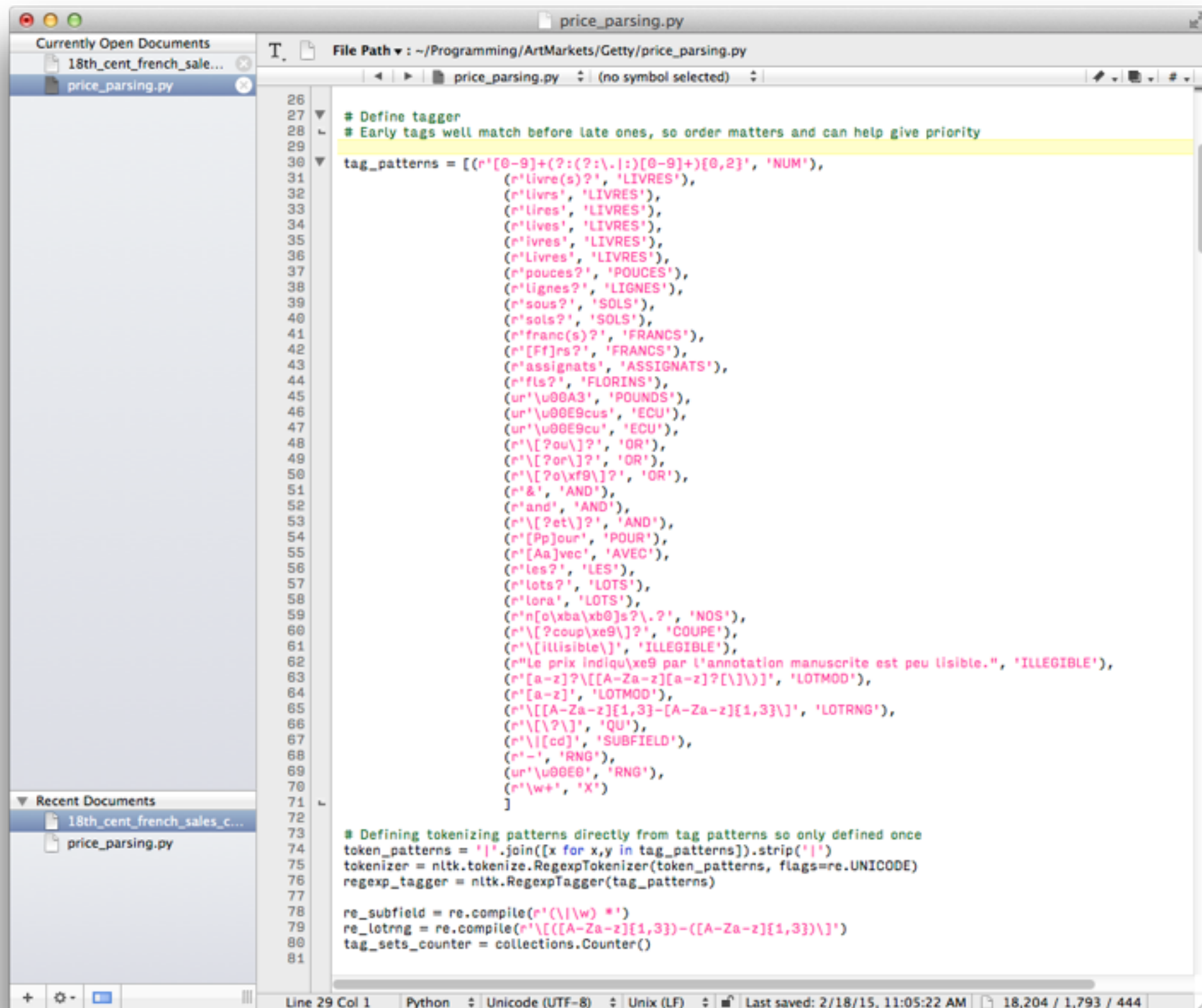
Finding patterns in the price text (tokenization and tagging)



Trick: *Number* of lots -> price per lot

60 livres |d pour les lots 476 [a-b] & 695

Matching text patterns and tagging types



The screenshot shows a code editor window titled 'price_parsing.py' with a file path of '~/.Programming/ArtMarkets/Getty/price_parsing.py'. The editor displays a Python script for defining a tagger and tokenizers. The script includes comments and a list of regular expressions for various text patterns, such as numbers, currency units, and specific words. The code is as follows:

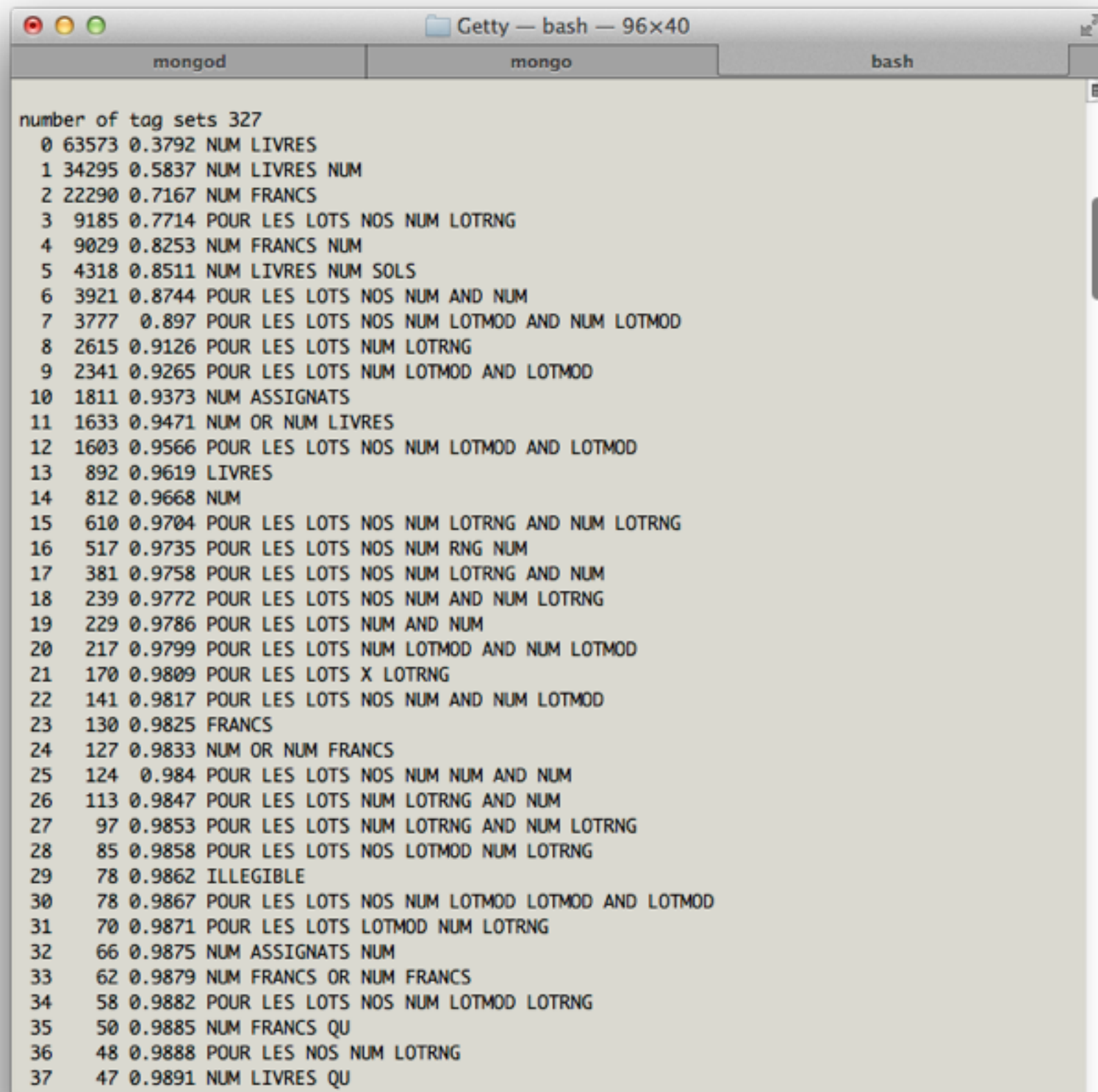
```
26
27 # Define tagger
28 # Early tags will match before late ones, so order matters and can help give priority
29
30 tag_patterns = [(r'[0-9]+(?:[.](?:[0-9]+)?)?[0,2]', 'NUM'),
31                 (r'livre(s)?', 'LIVRES'),
32                 (r'livrs', 'LIVRES'),
33                 (r'lres', 'LIVRES'),
34                 (r'lives', 'LIVRES'),
35                 (r'ivres', 'LIVRES'),
36                 (r'Livres', 'LIVRES'),
37                 (r'pouces?', 'POUCES'),
38                 (r'lignes?', 'LIGNES'),
39                 (r'sous?', 'SOLS'),
40                 (r'sols?', 'SOLS'),
41                 (r'franc(s)?', 'FRANCS'),
42                 (r'[Ff]rs?', 'FRANCS'),
43                 (r'assignats', 'ASSIGNATS'),
44                 (r'fls?', 'FLORINS'),
45                 (ur'\u00A3', 'POUNDS'),
46                 (ur'\u00E9cus', 'ECU'),
47                 (ur'\u00E9cu', 'ECU'),
48                 (r'\[?ou\]? ', 'OR'),
49                 (r'\[?or\]? ', 'OR'),
50                 (r'\[?o\xf9\]? ', 'OR'),
51                 (r'&', 'AND'),
52                 (r'and', 'AND'),
53                 (r'\[?et\]? ', 'AND'),
54                 (r'[Pp]our', 'POUR'),
55                 (r'[Aa]vec', 'AVEC'),
56                 (r'les?', 'LES'),
57                 (r'lots?', 'LOTS'),
58                 (r'lora', 'LOTS'),
59                 (r'n[o\xba\xbb]s?\.?', 'NOS'),
60                 (r'\[?coup\xe9\]? ', 'COUPE'),
61                 (r'\[?illisible\]', 'ILLEGIBLE'),
62                 (r'"Le prix indiqu\xe9 par l'annotation manuscrite est peu lisible.", 'ILLEGIBLE'),
63                 (r'[a-z]?[A-Za-z][a-z]?[A-Za-z]', 'LOTMOD'),
64                 (r'[a-z]', 'LOTMOD'),
65                 (r'\[[A-Za-z]{1,3}-[A-Za-z]{1,3}\]', 'LOTRNG'),
66                 (r'\[? \]', 'QU'),
67                 (r'\[?cd\]', 'SUBFIELD'),
68                 (r'-', 'RNG'),
69                 (ur'\u00E0', 'RNG'),
70                 (r'w+', 'X')
71 ]
72
73 # Defining tokenizing patterns directly from tag patterns so only defined once
74 token_patterns = '|'.join([x for x,y in tag_patterns]).strip('|')
75 tokenizer = nltk.tokenize.RegexpTokenizer(token_patterns, flags=re.UNICODE)
76 regexp_tagger = nltk.RegexpTagger(tag_patterns)
77
78 re_subfield = re.compile(r'(\[?cd\] *')
79 re_lotrng = re.compile(r'\[[A-Za-z]{1,3}-[A-Za-z]{1,3}\]')
80 tag_sets_counter = collections.Counter()
81
```

The status bar at the bottom indicates 'Line 29 Col 1', 'Python', 'Unicode (UTF-8)', 'Unix (LF)', and 'Last saved: 2/18/15, 11:05:22 AM'. The file size is shown as 18,204 / 1,793 / 444.

Regular Expressions text pattern matching language + NLTK RegEx tagger

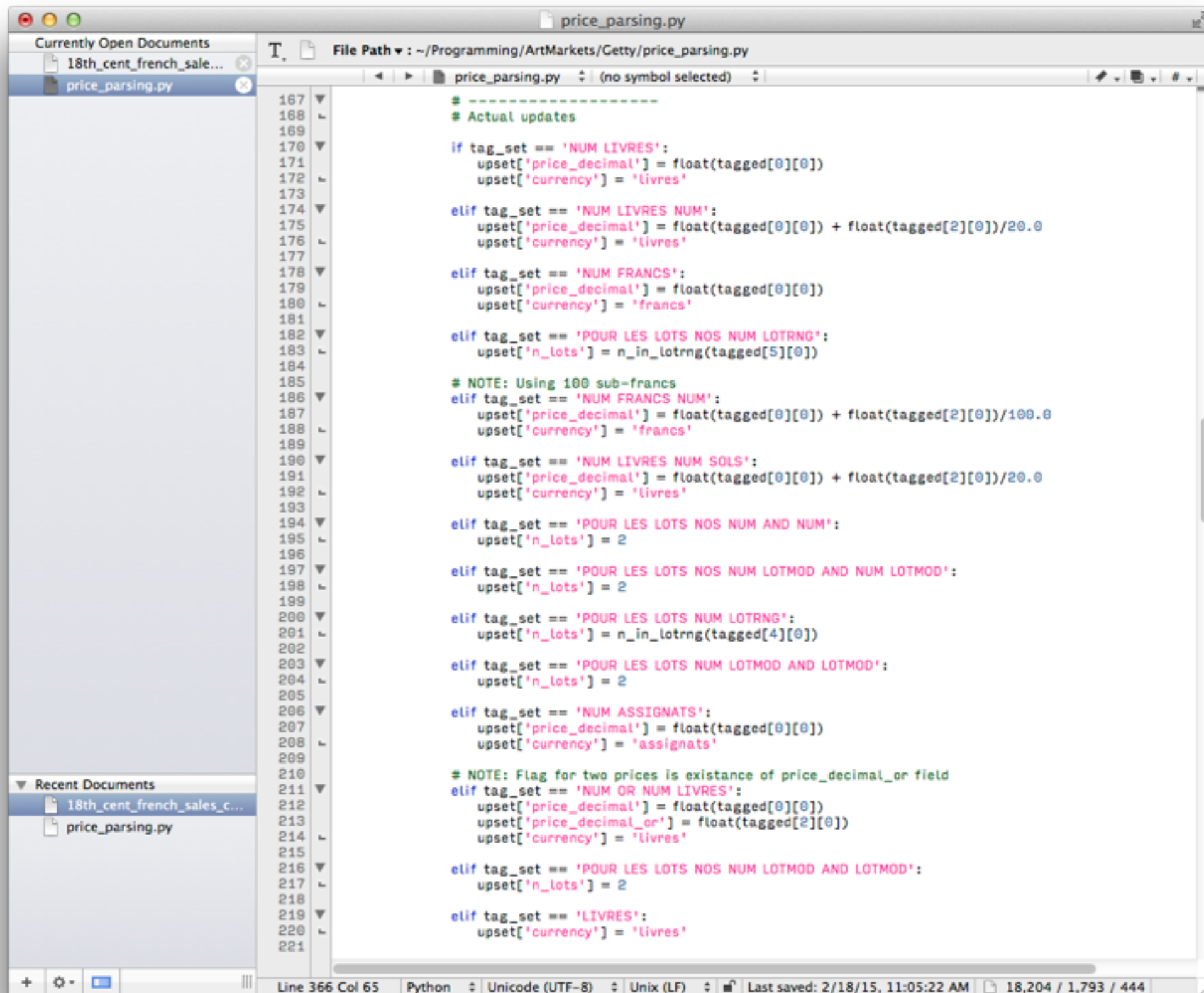
```
tag_patterns = [(r'[0-9]+(?:[0-9]+){0,2}', 'NUM'),
                 (r'livre(s)?', 'LIVRES'),
                 (r'livrs', 'LIVRES'),
                 (r'lires', 'LIVRES'),
                 (r'lives', 'LIVRES'),
                 (r'ivres', 'LIVRES'),
                 (r'Livres', 'LIVRES'),
                 (r'pouces?', 'POUCES'),
                 (r'lignes?', 'LIGNES'),
                 (r'sous?', 'SOLS'),
                 (r'sols?', 'SOLS'),
                 (r'franc(s)?', 'FRANCS'),
                 (r'[Ff]rs?', 'FRANCS'),
                 (r'assignats', 'ASSIGNATS'),
                 (r'fls?', 'FLORINS'),
                 (ur'\u00A3', 'POUNDS'),
                 (ur'\u00E9cus', 'ECU'),
                 (ur'\u00E9cu', 'ECU'),
                 (r'\[?ou\]', 'OR'),
                 (r'\[?or\]', 'OR'),
                 (r'\[?o\x{9}\]', 'OR'),
                 (r'&', 'AND'),
                 (r'and', 'AND'),
                 (r'\[?et\]', 'AND'),
                 ...
                 (r'\w+', 'X')]
```

Tag patterns and counts

A terminal window titled 'Getty — bash — 96x40' with three tabs: 'mongod', 'mongo', and 'bash'. The 'mongo' tab is active. The terminal displays a list of 37 tag patterns, each preceded by an index number from 0 to 37. Each pattern consists of a count, a frequency, and a tag description. The patterns are sorted by frequency in descending order.

```
number of tag sets 327
0 63573 0.3792 NUM LIVRES
1 34295 0.5837 NUM LIVRES NUM
2 22290 0.7167 NUM FRANCS
3 9185 0.7714 POUR LES LOTS NOS NUM LOTRNG
4 9029 0.8253 NUM FRANCS NUM
5 4318 0.8511 NUM LIVRES NUM SOLS
6 3921 0.8744 POUR LES LOTS NOS NUM AND NUM
7 3777 0.897 POUR LES LOTS NOS NUM LOTMOD AND NUM LOTMOD
8 2615 0.9126 POUR LES LOTS NUM LOTRNG
9 2341 0.9265 POUR LES LOTS NUM LOTMOD AND LOTMOD
10 1811 0.9373 NUM ASSIGNATS
11 1633 0.9471 NUM OR NUM LIVRES
12 1603 0.9566 POUR LES LOTS NOS NUM LOTMOD AND LOTMOD
13 892 0.9619 LIVRES
14 812 0.9668 NUM
15 610 0.9704 POUR LES LOTS NOS NUM LOTRNG AND NUM LOTRNG
16 517 0.9735 POUR LES LOTS NOS NUM RNG NUM
17 381 0.9758 POUR LES LOTS NOS NUM LOTRNG AND NUM
18 239 0.9772 POUR LES LOTS NOS NUM AND NUM LOTRNG
19 229 0.9786 POUR LES LOTS NUM AND NUM
20 217 0.9799 POUR LES LOTS NUM LOTMOD AND NUM LOTMOD
21 170 0.9809 POUR LES LOTS X LOTRNG
22 141 0.9817 POUR LES LOTS NOS NUM AND NUM LOTMOD
23 130 0.9825 FRANCS
24 127 0.9833 NUM OR NUM FRANCS
25 124 0.984 POUR LES LOTS NOS NUM NUM AND NUM
26 113 0.9847 POUR LES LOTS NUM LOTRNG AND NUM
27 97 0.9853 POUR LES LOTS NUM LOTRNG AND NUM LOTRNG
28 85 0.9858 POUR LES LOTS NOS LOTMOD NUM LOTRNG
29 78 0.9862 ILLEGIBLE
30 78 0.9867 POUR LES LOTS NOS NUM LOTMOD LOTMOD AND LOTMOD
31 70 0.9871 POUR LES LOTS LOTMOD NUM LOTRNG
32 66 0.9875 NUM ASSIGNATS NUM
33 62 0.9879 NUM FRANCS OR NUM FRANCS
34 58 0.9882 POUR LES LOTS NOS NUM LOTMOD LOTRNG
35 50 0.9885 NUM FRANCS QU
36 48 0.9888 POUR LES NOS NUM LOTRNG
37 47 0.9891 NUM LIVRES QU
```

Extracting price and lots from tag patterns (52)



```
167 # -----
168 # Actual updates
169
170 if tag_set == 'NUM LIVRES':
171     upset['price_decimal'] = float(tagged[0][0])
172     upset['currency'] = 'livres'
173
174 elif tag_set == 'NUM LIVRES NUM':
175     upset['price_decimal'] = float(tagged[0][0]) + float(tagged[2][0])/20.0
176     upset['currency'] = 'livres'
177
178 elif tag_set == 'NUM FRANCS':
179     upset['price_decimal'] = float(tagged[0][0])
180     upset['currency'] = 'francs'
181
182 elif tag_set == 'POUR LES LOTS NOS NUM LOTRNG':
183     upset['n_lots'] = n_in_lotrng(tagged[5][0])
184
185 # NOTE: Using 100 sub-francs
186 elif tag_set == 'NUM FRANCS NUM':
187     upset['price_decimal'] = float(tagged[0][0]) + float(tagged[2][0])/100.0
188     upset['currency'] = 'francs'
189
190 elif tag_set == 'NUM LIVRES NUM SOLS':
191     upset['price_decimal'] = float(tagged[0][0]) + float(tagged[2][0])/20.0
192     upset['currency'] = 'livres'
193
194 elif tag_set == 'POUR LES LOTS NOS NUM AND NUM':
195     upset['n_lots'] = 2
196
197 elif tag_set == 'POUR LES LOTS NOS NUM LOTMOD AND NUM LOTMOD':
198     upset['n_lots'] = 2
199
200 elif tag_set == 'POUR LES LOTS NUM LOTRNG':
201     upset['n_lots'] = n_in_lotrng(tagged[4][0])
202
203 elif tag_set == 'POUR LES LOTS NUM LOTMOD AND LOTMOD':
204     upset['n_lots'] = 2
205
206 elif tag_set == 'NUM ASSIGNATS':
207     upset['price_decimal'] = float(tagged[0][0])
208     upset['currency'] = 'assignats'
209
210 # NOTE: Flag for two prices is existence of price_decimal_or field
211 elif tag_set == 'NUM OR NUM LIVRES':
212     upset['price_decimal'] = float(tagged[0][0])
213     upset['price_decimal_or'] = float(tagged[2][0])
214     upset['currency'] = 'livres'
215
216 elif tag_set == 'POUR LES LOTS NOS NUM LOTMOD AND LOTMOD':
217     upset['n_lots'] = 2
218
219 elif tag_set == 'LIVRES':
220     upset['currency'] = 'livres'
221
```

Line 366 Col 65 Python Unicode (UTF-8) Unix (LF) Last saved: 2/18/15, 11:05:22 AM 18,204 / 1,793 / 444

Price per lot and currency

138149 price_per_lot
138401 currency

	AR	AT	AW	AX
1	price_decimal	currency	n_lots	price_per_lot
2801	33	livres		33
2802	46.05	livres		46.05
2803	31.05	livres		31.05
2804	8.95	livres		8.95
2805	3.25	livres		3.25
2806	8.95	livres		8.95
2807	4.5	livres		4.5
2808	19	livres		19
2809	9.05	livres		9.05
2810	9	livres	5	1.8
2811	9	livres	5	1.8
2812	9	livres	5	1.8
2813	9	livres	5	1.8
2814	6.2	livres		6.2
2815	9	livres	2	4.5
2816	8.95	livres	3	2.983333333
2817	8.95	livres	3	2.983333333
2818	8.95	livres	3	2.983333333
2819	5	livres	2	2.5
2820	4	livres	5	0.8
2821	4	livres	5	0.8
2822	4	livres	5	0.8
2823	4	livres	5	0.8
2824	3.05	livres	2	1.525
2825	3.05	livres	2	1.525
2826	3.05	livres	2	1.525
2827	3.05	livres	2	1.525
2828	4.25	livres	2	2.125
2829	4.25	livres	2	2.125
2830	72	livres		72
2831	3.5	livres	2	1.75
2832	3.5	livres	2	1.75
2833	4	livres	5	0.8
2834	9	livres		9
2835	26.5	livres		26.5
2836	5.95	livres	5	1.19
2837	5.95	livres	5	1.19