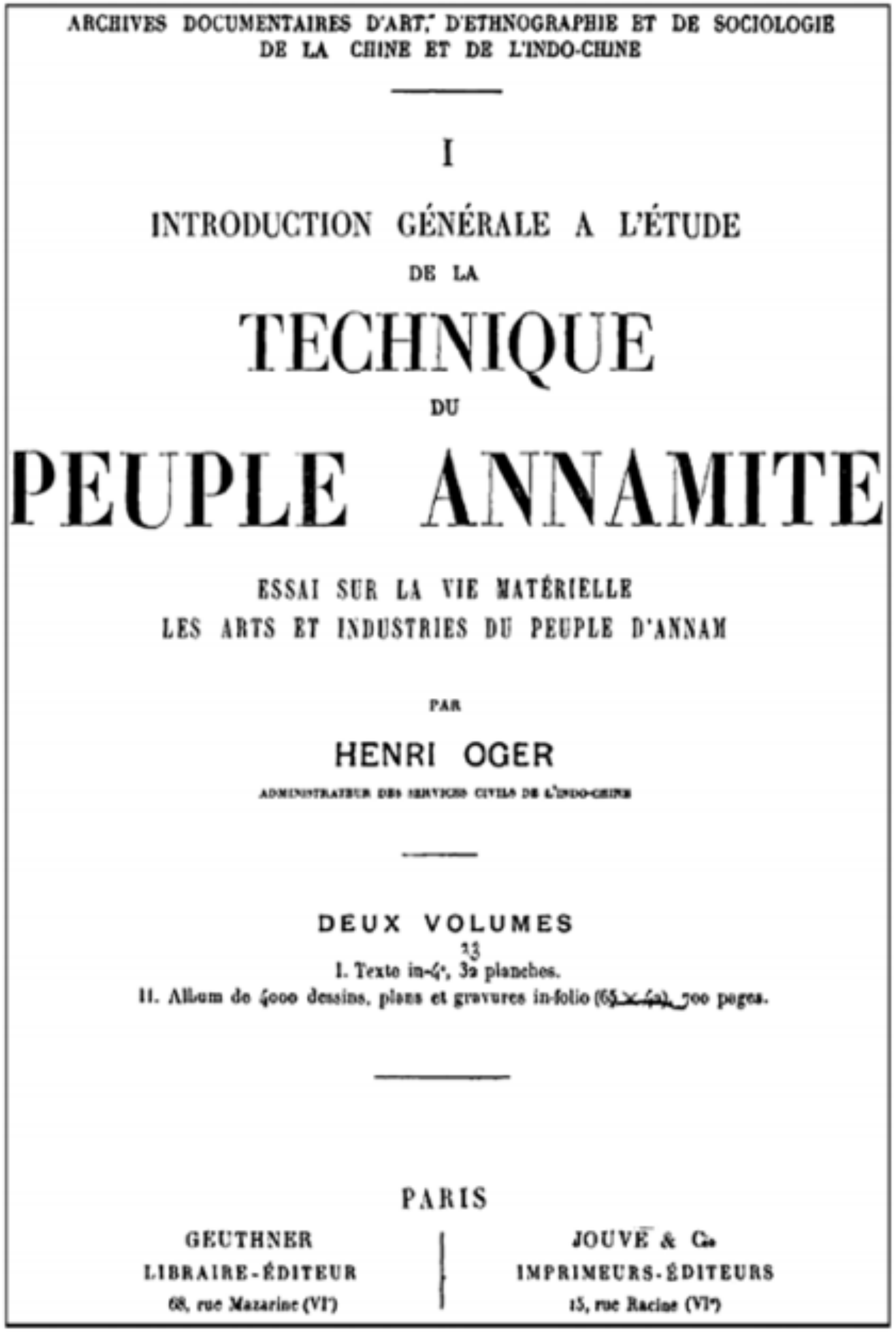
Case Study: Data Recovery Using Python

Harrison Dekker
UC Berkeley Libraries
@ucdatalab

summary

A 2009 edition of a rare 1909 ethnographic study of Vietnamese industry included a supplemental DVD with interactive software for browsing 4000+ sketches from the original publication. The researcher wanted to extract text data embedded in the application.



ARCHIVES DOCUMENTAIRES D'ART, D'ETHNOGRAPHIE ET DE SOCIOLOGIE
DE LA CHINE ET DE L'INDO-CHINE

I

INTRODUCTION GÉNÉRALE A L'ÉTUDE
DE LA
TECHNIQUE
DU
PEUPLE ANNAMITE

ESSAI SUR LA VIE MATÉRIELLE
LES ARTS ET INDUSTRIES DU PEUPLE D'ANNAM

PAR

HENRI OGER

ADMINISTRATEUR DES SERVICES CIVILS DE L'INDO-CHINE

DEUX VOLUMES

33
I. Texte in-4°, 32 planches.
II. Album de 4000 dessins, plans et gravures in-folio (65 × 42), 700 pages.

PARIS

GEUTHNER
LIBRAIRE-ÉDITEUR
68, rue Mazarine (VI°)

JOUVE & Cie
IMPRIMEURS-ÉDITEURS
15, rue Racine (VI°)

plan b

The application developer didn’t respond to the researcher’s request to share their data. Sought help from the Library Data Lab to programmatically extract the text.

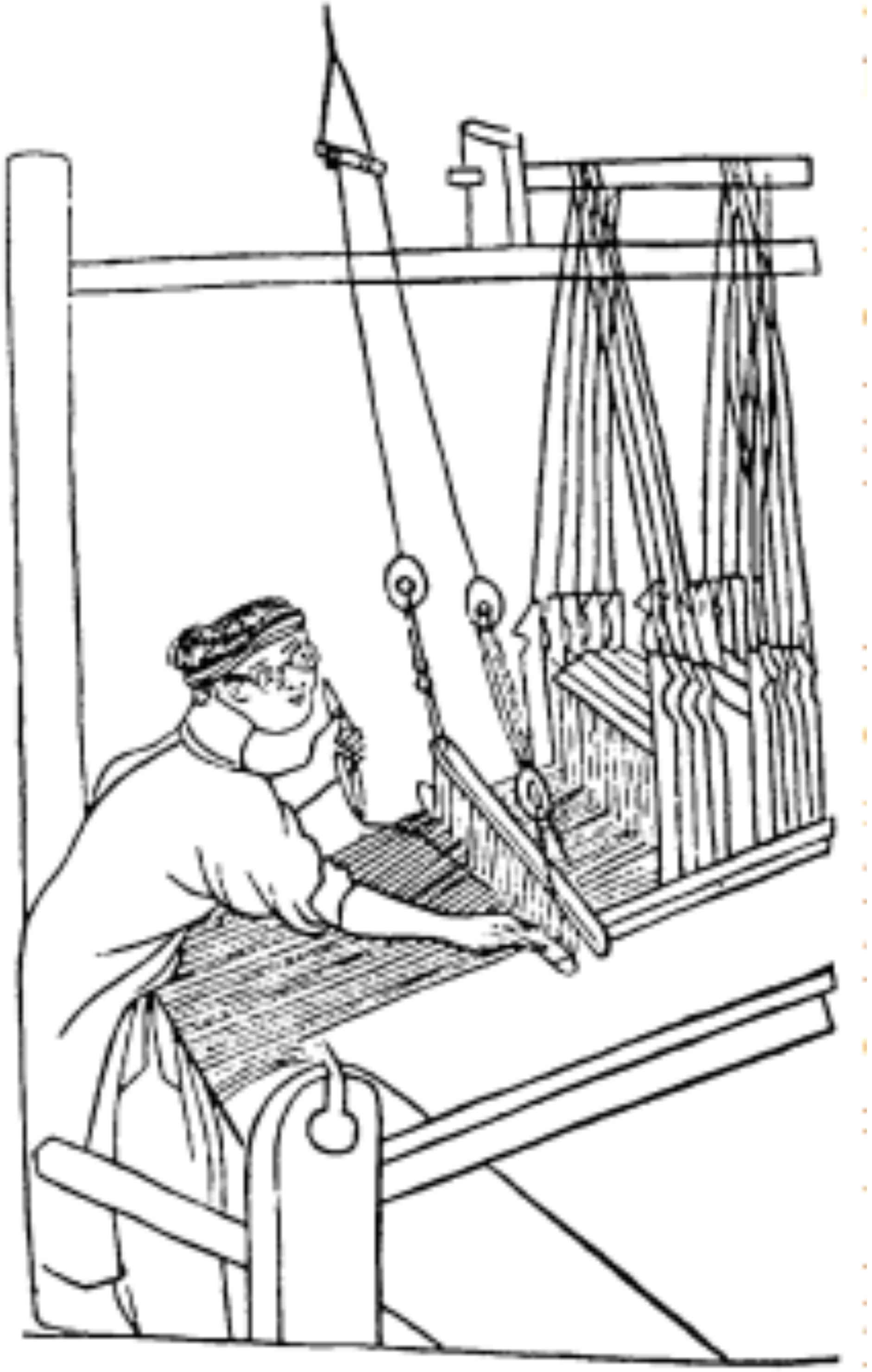
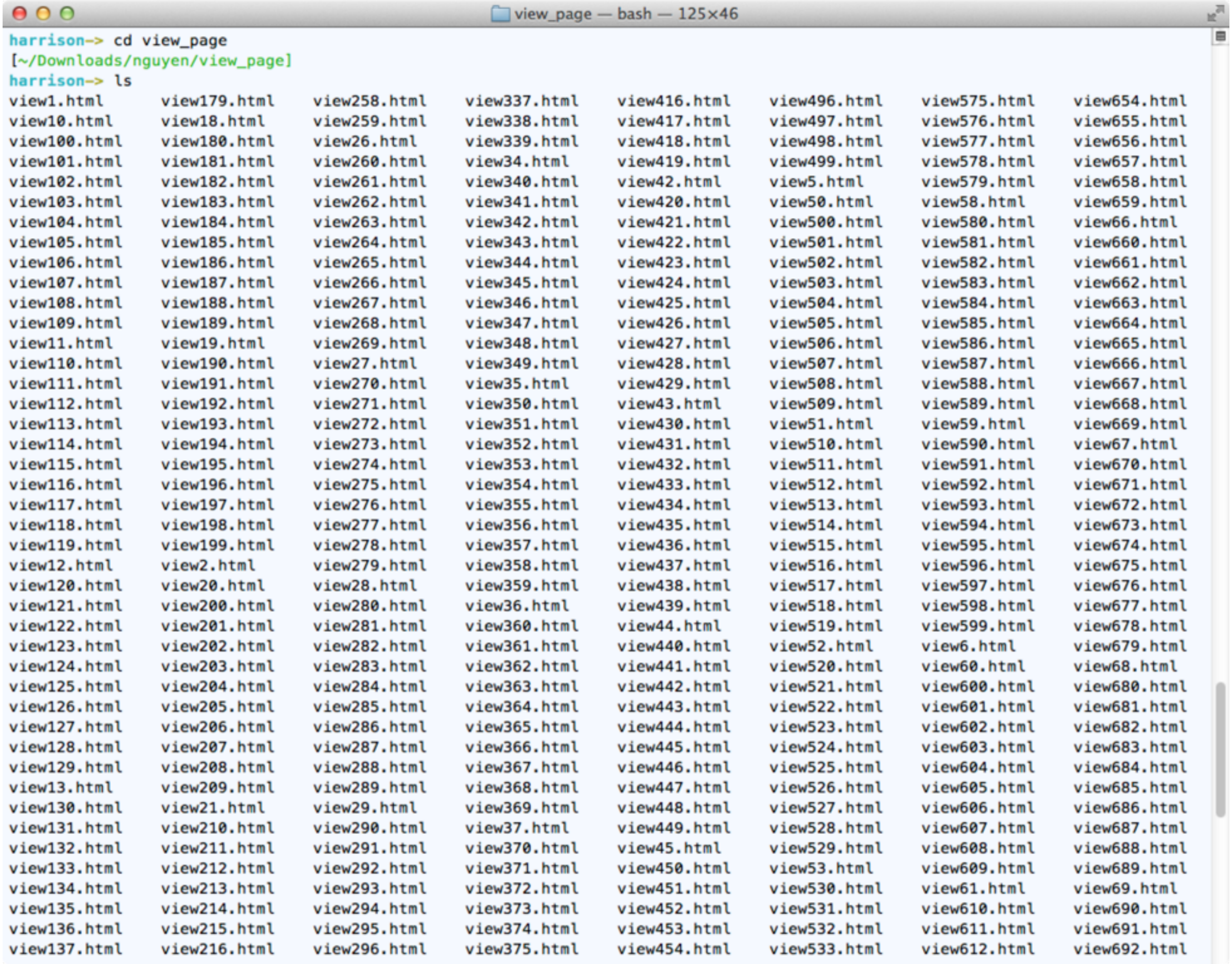
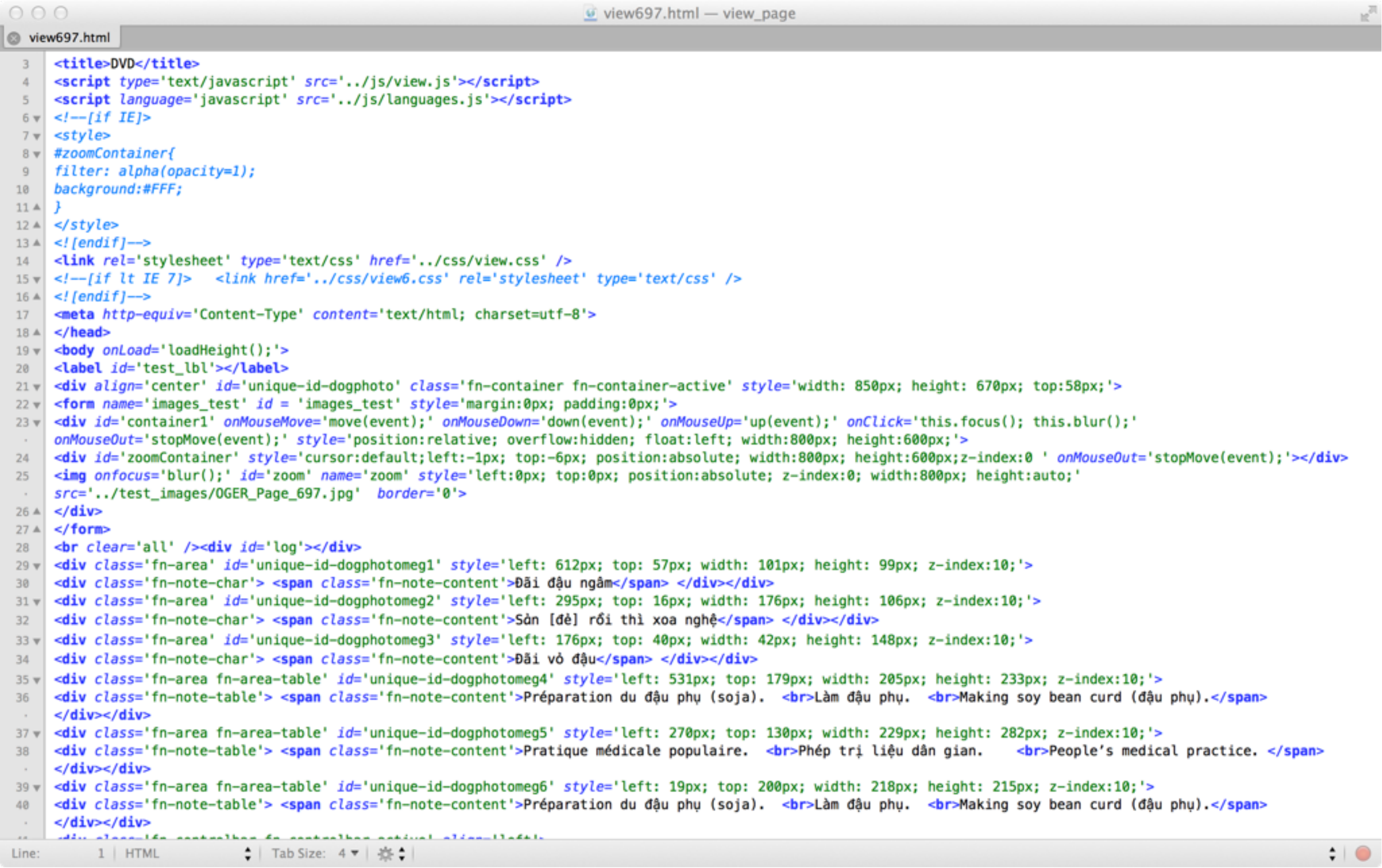


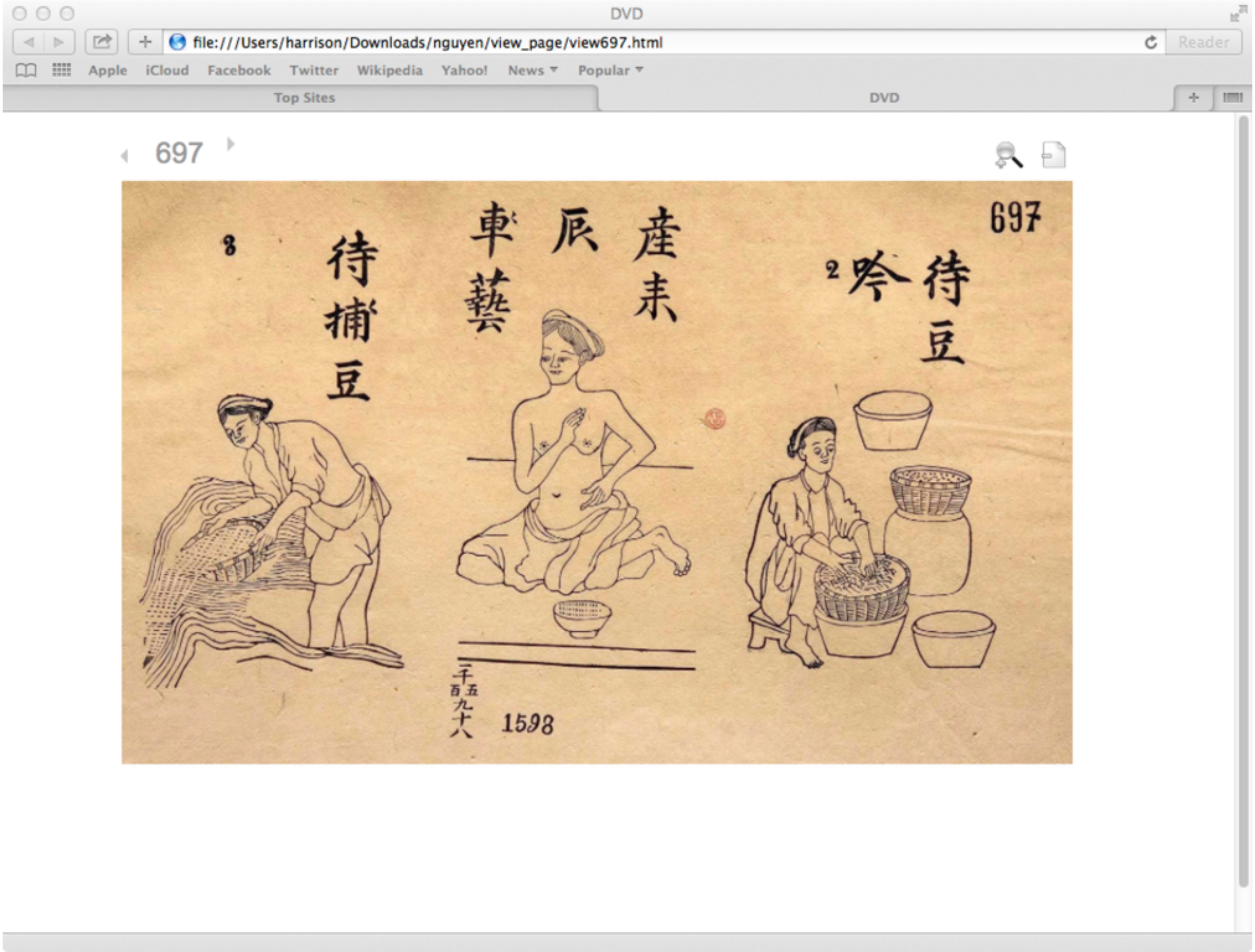
FIG. 19. – SILK LOOM



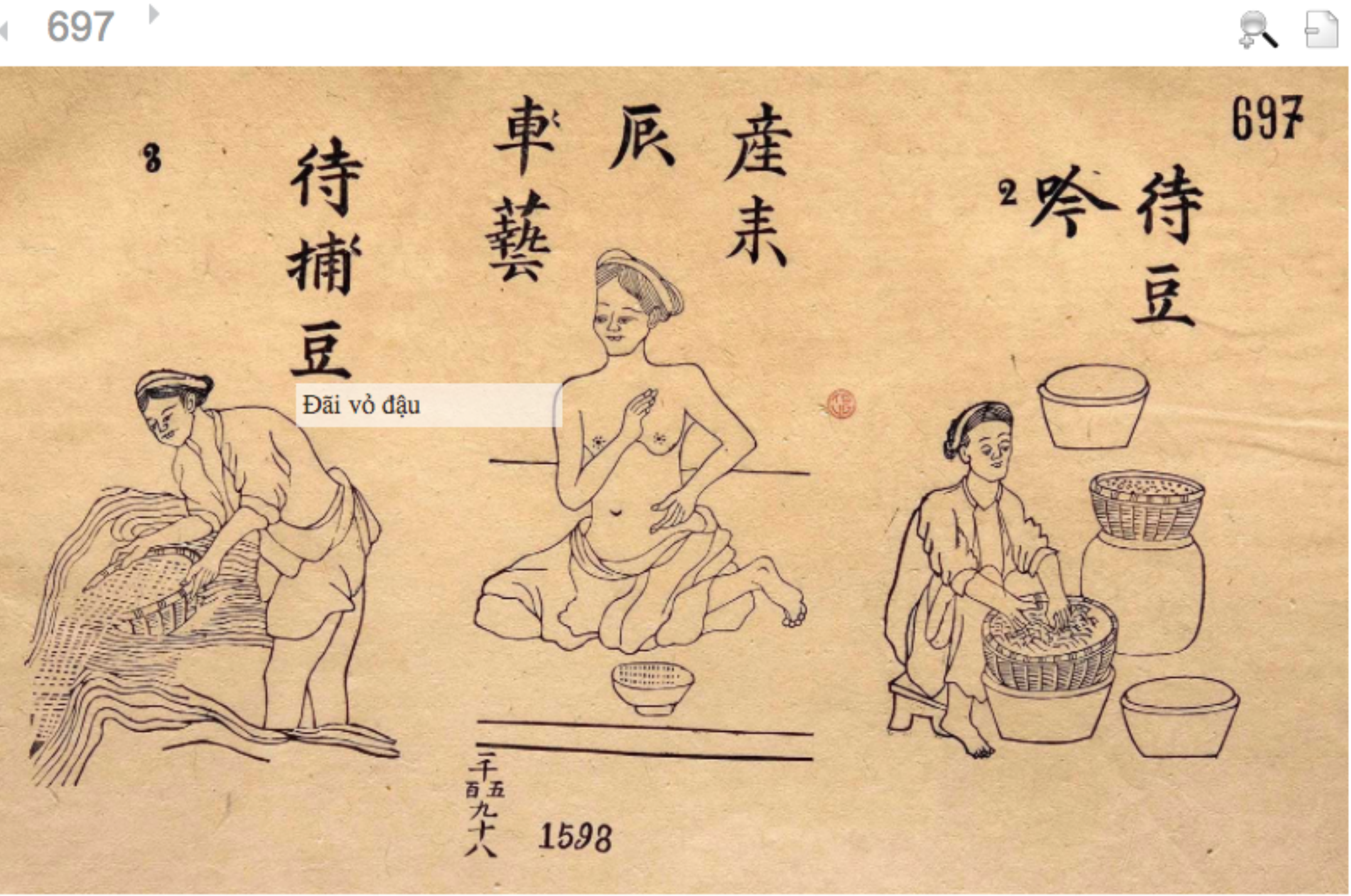
```
harrison-> cd view_page
[~/Downloads/nguyen/view_page]
harrison-> ls
view1.html      view179.html    view258.html    view337.html    view416.html    view496.html    view575.html    view654.html
view10.html     view18.html     view259.html    view338.html    view417.html    view497.html    view576.html    view655.html
view100.html    view180.html    view26.html     view339.html    view418.html    view498.html    view577.html    view656.html
view101.html    view181.html    view260.html    view34.html     view419.html    view499.html    view578.html    view657.html
view102.html    view182.html    view261.html    view340.html    view42.html     view5.html      view579.html    view658.html
view103.html    view183.html    view262.html    view341.html    view420.html    view50.html     view58.html     view659.html
view104.html    view184.html    view263.html    view342.html    view421.html    view500.html    view580.html    view66.html
view105.html    view185.html    view264.html    view343.html    view422.html    view501.html    view581.html    view660.html
view106.html    view186.html    view265.html    view344.html    view423.html    view502.html    view582.html    view661.html
view107.html    view187.html    view266.html    view345.html    view424.html    view503.html    view583.html    view662.html
view108.html    view188.html    view267.html    view346.html    view425.html    view504.html    view584.html    view663.html
view109.html    view189.html    view268.html    view347.html    view426.html    view505.html    view585.html    view664.html
view11.html     view19.html     view269.html    view348.html    view427.html    view506.html    view586.html    view665.html
view110.html    view190.html    view27.html     view349.html    view428.html    view507.html    view587.html    view666.html
view111.html    view191.html    view270.html    view35.html     view429.html    view508.html    view588.html    view667.html
view112.html    view192.html    view271.html    view350.html    view43.html     view509.html    view589.html    view668.html
view113.html    view193.html    view272.html    view351.html    view430.html    view51.html     view59.html     view669.html
view114.html    view194.html    view273.html    view352.html    view431.html    view510.html    view590.html    view67.html
view115.html    view195.html    view274.html    view353.html    view432.html    view511.html    view591.html    view670.html
view116.html    view196.html    view275.html    view354.html    view433.html    view512.html    view592.html    view671.html
view117.html    view197.html    view276.html    view355.html    view434.html    view513.html    view593.html    view672.html
view118.html    view198.html    view277.html    view356.html    view435.html    view514.html    view594.html    view673.html
view119.html    view199.html    view278.html    view357.html    view436.html    view515.html    view595.html    view674.html
view12.html     view2.html      view279.html    view358.html    view437.html    view516.html    view596.html    view675.html
view120.html    view20.html     view28.html     view359.html    view438.html    view517.html    view597.html    view676.html
view121.html    view200.html    view280.html    view36.html     view439.html    view518.html    view598.html    view677.html
view122.html    view201.html    view281.html    view360.html    view44.html     view519.html    view599.html    view678.html
view123.html    view202.html    view282.html    view361.html    view440.html    view52.html     view6.html      view679.html
view124.html    view203.html    view283.html    view362.html    view441.html    view520.html    view60.html     view68.html
view125.html    view204.html    view284.html    view363.html    view442.html    view521.html    view600.html    view680.html
view126.html    view205.html    view285.html    view364.html    view443.html    view522.html    view601.html    view681.html
view127.html    view206.html    view286.html    view365.html    view444.html    view523.html    view602.html    view682.html
view128.html    view207.html    view287.html    view366.html    view445.html    view524.html    view603.html    view683.html
view129.html    view208.html    view288.html    view367.html    view446.html    view525.html    view604.html    view684.html
view13.html     view209.html    view289.html    view368.html    view447.html    view526.html    view605.html    view685.html
view130.html    view21.html     view29.html     view369.html    view448.html    view527.html    view606.html    view686.html
view131.html    view210.html    view290.html    view37.html     view449.html    view528.html    view607.html    view687.html
view132.html    view211.html    view291.html    view370.html    view45.html     view529.html    view608.html    view688.html
view133.html    view212.html    view292.html    view371.html    view450.html    view53.html     view609.html    view689.html
view134.html    view213.html    view293.html    view372.html    view451.html    view530.html    view61.html     view69.html
view135.html    view214.html    view294.html    view373.html    view452.html    view531.html    view610.html    view690.html
view136.html    view215.html    view295.html    view374.html    view453.html    view532.html    view611.html    view691.html
view137.html    view216.html    view296.html    view375.html    view454.html    view533.html    view612.html    view692.html
```



```
<title>DVD</title>
<script type='text/javascript' src='../js/view.js'></script>
<script language='javascript' src='../js/languages.js'></script>
<!--[if IE]>
<style>
#zoomContainer{
filter: alpha(opacity=1);
background:#FFF;
}
</style>
<![endif]-->
<link rel='stylesheet' type='text/css' href='../css/view.css' />
<!--[if lt IE 7]>   <link href='../css/view6.css' rel='stylesheet' type='text/css' />
<![endif]-->
<meta http-equiv='Content-Type' content='text/html; charset=utf-8'>
</head>
<body onLoad='loadHeight();'>
<label id='test_lbl'></label>
<div align='center' id='unique-id-dogphoto' class='fn-container fn-container-active' style='width: 850px; height: 670px; top:58px;'>
<form name='images_test' id = 'images_test' style='margin:0px; padding:0px;'>
<div id='container1' onMouseMove='move(event);' onMouseDown='down(event);' onMouseUp='up(event);' onClick='this.focus(); this.blur();'
onMouseOut='stopMove(event);' style='position:relative; overflow:hidden; float:left; width:800px; height:600px;'>
<div id='zoomContainer' style='cursor:default;left:-1px; top:-6px; position:absolute; width:800px; height:600px;z-index:0 ' onMouseOut='stopMove(event);'></div>
<img onfocus='blur();' id='zoom' name='zoom' style='left:0px; top:0px; position:absolute; z-index:0; width:800px; height:auto;'
src='../test_images/OGER_Page_697.jpg'  border='0'>
</div>
</form>
<br clear='all' /><div id='log'></div>
<div class='fn-area' id='unique-id-dogphotomeg1' style='left: 612px; top: 57px; width: 101px; height: 99px; z-index:10;'>
<div class='fn-note-char'> <span class='fn-note-content'>Đãi đậu ngâm</span> </div></div>
<div class='fn-area' id='unique-id-dogphotomeg2' style='left: 295px; top: 16px; width: 176px; height: 106px; z-index:10;'>
<div class='fn-note-char'> <span class='fn-note-content'>Sản [đẻ] rồi thì xoa nghệ</span> </div></div>
<div class='fn-area' id='unique-id-dogphotomeg3' style='left: 176px; top: 40px; width: 42px; height: 148px; z-index:10;'>
<div class='fn-note-char'> <span class='fn-note-content'>Đãi vỏ đậu</span> </div></div>
<div class='fn-area fn-area-table' id='unique-id-dogphotomeg4' style='left: 531px; top: 179px; width: 205px; height: 233px; z-index:10;'>
<div class='fn-note-table'> <span class='fn-note-content'>Préparation du đậu phụ (soja).  <br>Làm đậu phụ.  <br>Making soy bean curd (đậu phụ).</span>
</div></div>
<div class='fn-area fn-area-table' id='unique-id-dogphotomeg5' style='left: 270px; top: 130px; width: 229px; height: 282px; z-index:10;'>
<div class='fn-note-table'> <span class='fn-note-content'>Pratique médicale populaire.  <br>Phép trị liệu dân gian.    <br>People's medical practice. </span>
</div></div>
<div class='fn-area fn-area-table' id='unique-id-dogphotomeg6' style='left: 19px; top: 200px; width: 218px; height: 215px; z-index:10;'>
<div class='fn-note-table'> <span class='fn-note-content'>Préparation du đậu phụ (soja).  <br>Làm đậu phụ.  <br>Making soy bean curd (đậu phụ).</span>
</div></div>
<div class='fn-controlbar fn-controlbar-active' align='left'>
```



DVD
file:///Users/harrison/Downloads/nguyen/view_page/view697.html
Reader
Apple
iCloud
Facebook
Twitter
Wikipedia
Yahoo!
News
Popular
Top Sites
DVD
697
697
2
待豆
吟
產未
辰
車‘藝
3
待捕‘豆
一千五百九十八
1598



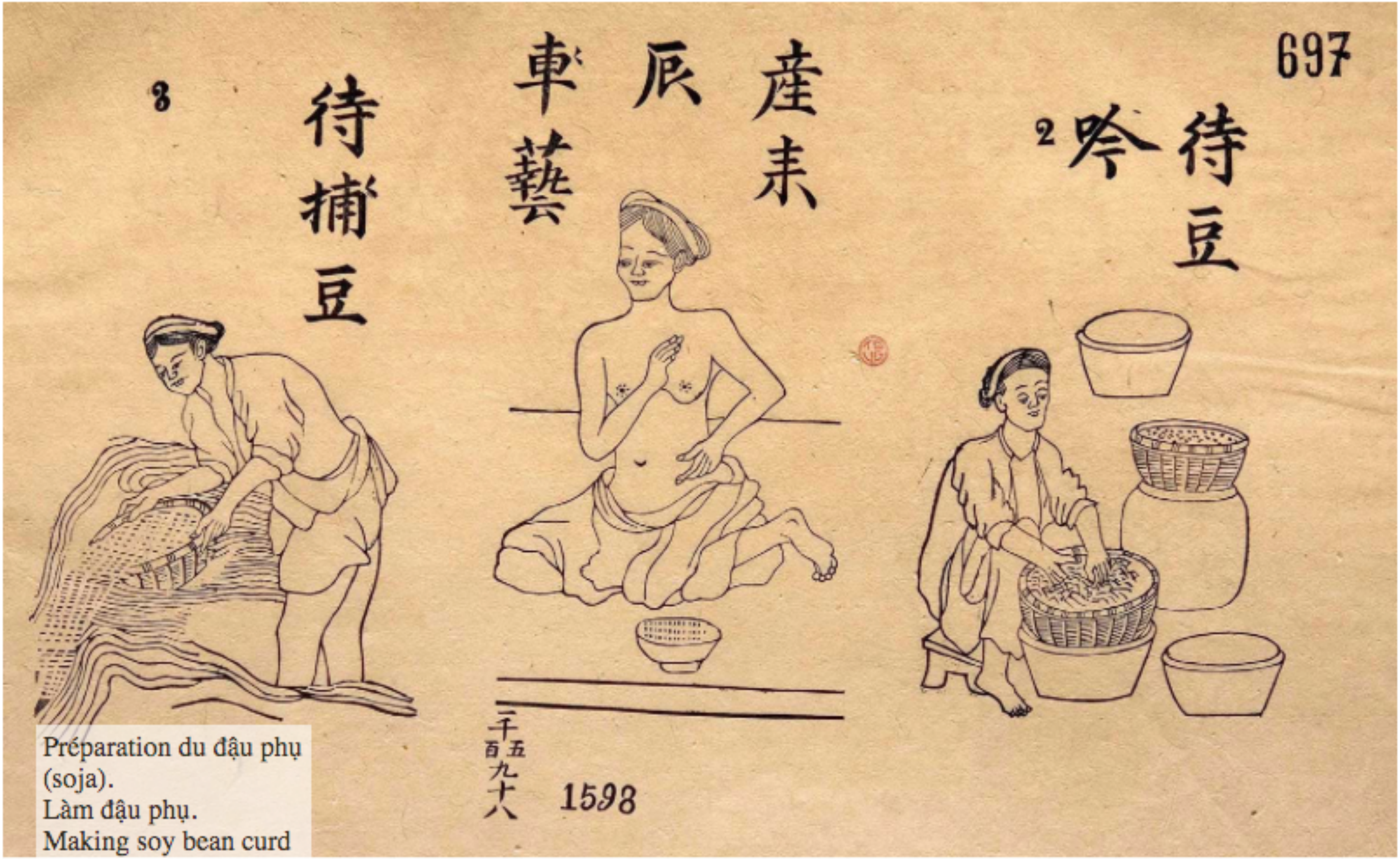
2 待豆吟

產耒辰車藝

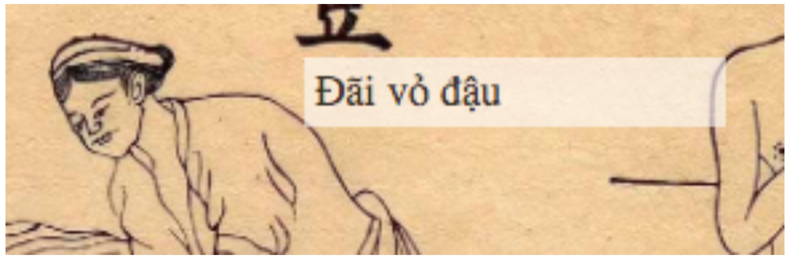
3 待捕豆

Đãi vỏ đậu

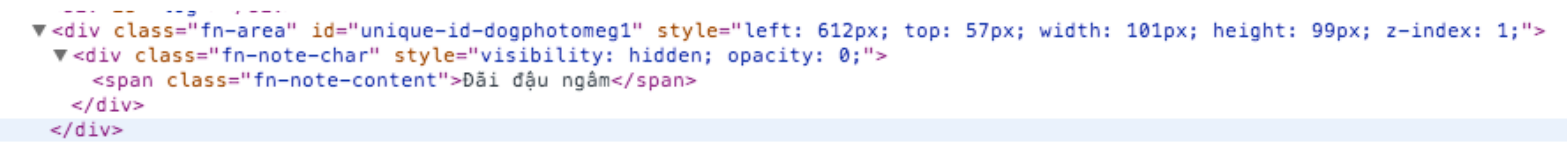
一千五百九十八 1598



Préparation du đậu phụ (soja).
Làm đậu phụ.
Making soy bean curd (đậu phụ).



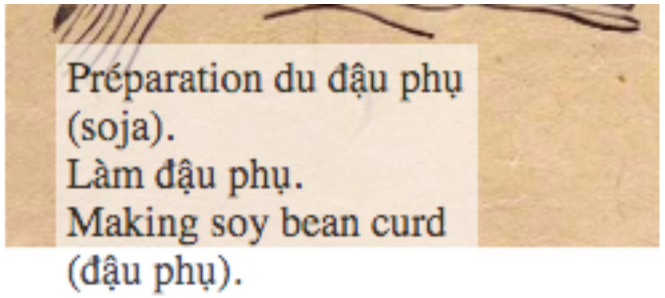
Đãi vỏ đậu



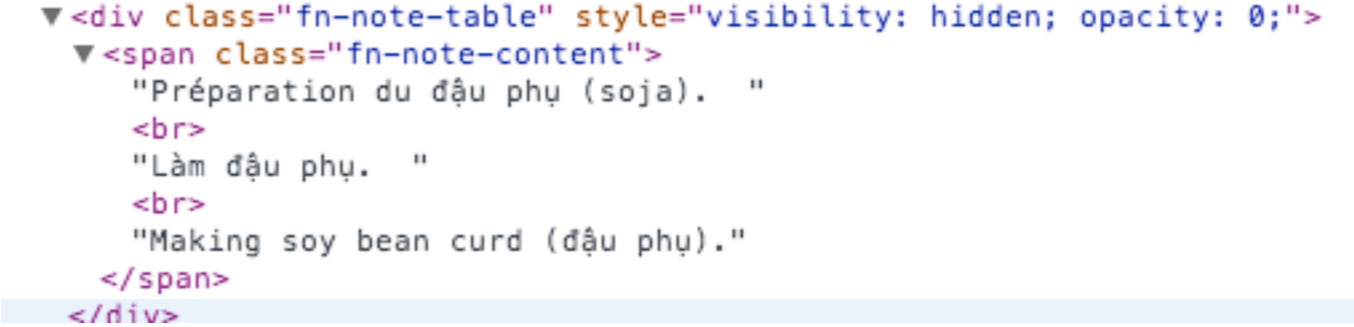
```
▼<div class="fn-area" id="unique-id-dogphotomeg1" style="left: 612px; top: 57px; width: 101px; height: 99px; z-index: 1;">
  ▼<div class="fn-note-char" style="visibility: hidden; opacity: 0;">
     <span class="fn-note-content">Đãi đậu ngâm</span>
   </div>
 </div>
```

objective

Researcher wanted to compare the Vietnamese artist’s descriptive inscription with the words of the French Historian. Reading the raw html was too difficult. Problem compounded by absence of any explicit linkage between the sections of text.



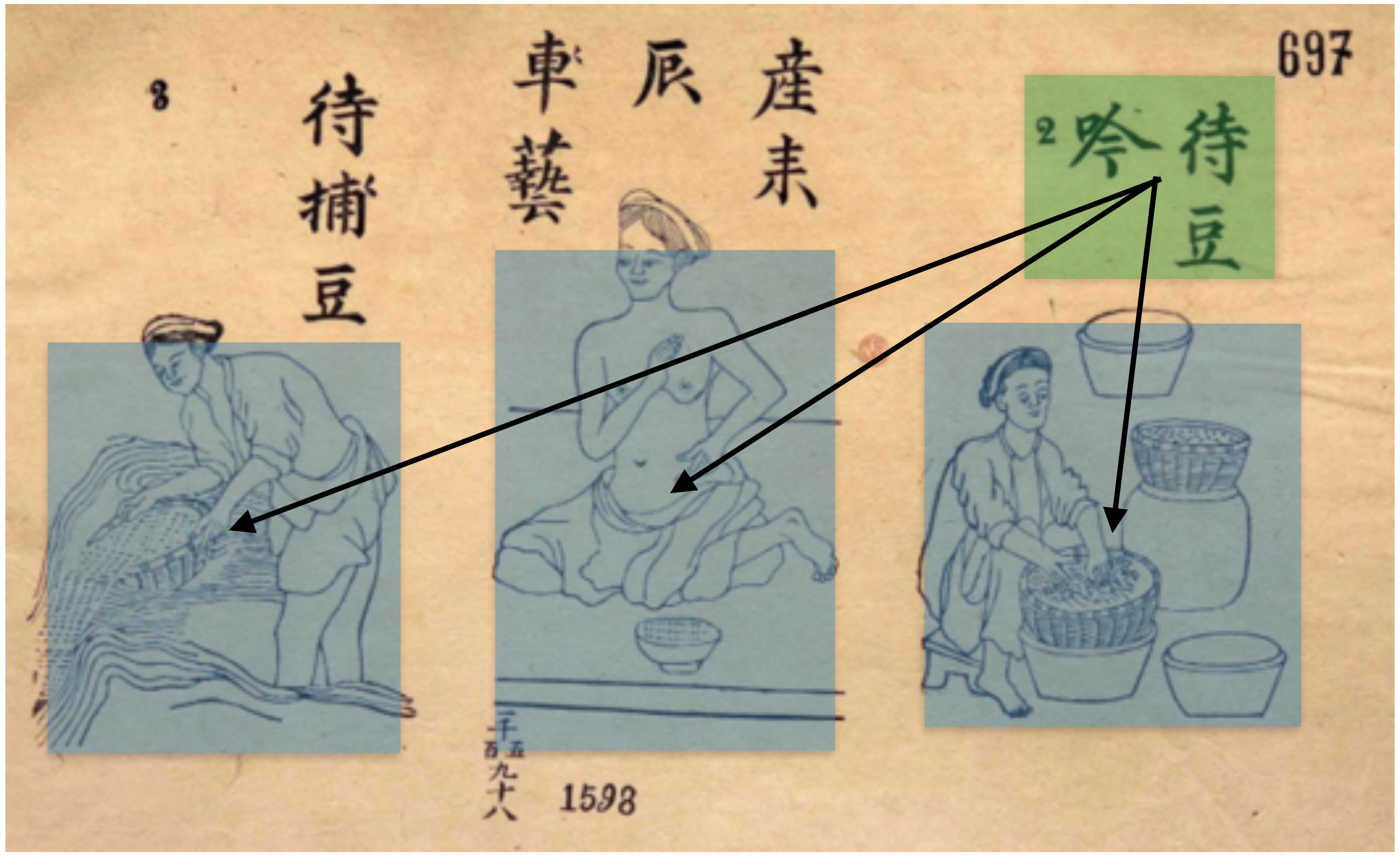
Préparation du đậu phụ (soja).
Làm đậu phụ.
Making soy bean curd (đậu phụ).

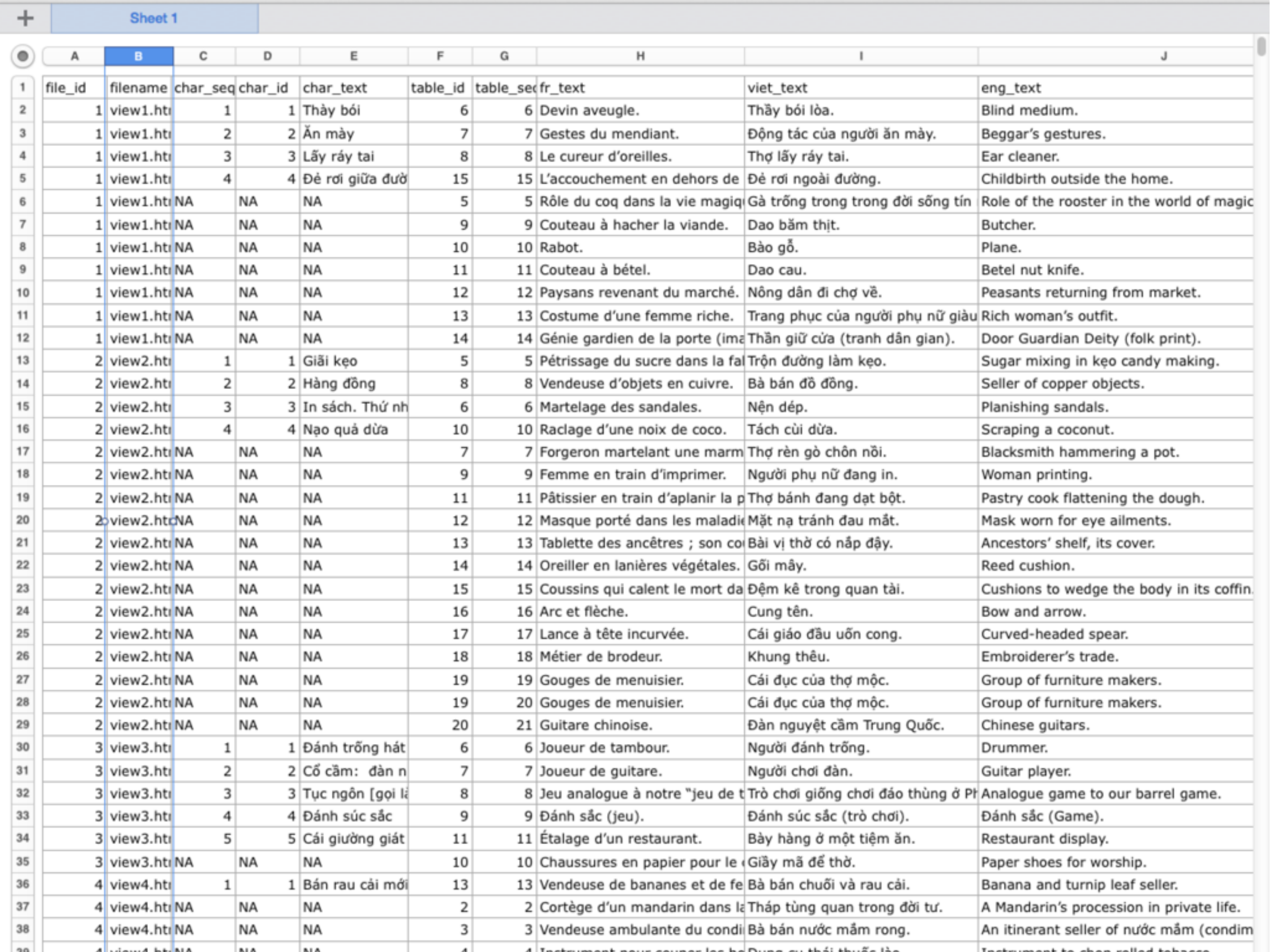


```
▼<div class="fn-note-table" style="visibility: hidden; opacity: 0;">
  ▼<span class="fn-note-content">
     "Préparation du đậu phụ (soja).  "
     <br>
     "Làm đậu phụ.  "
     <br>
     "Making soy bean curd (đậu phụ)."
   </span>
 </div>
```

solution

Make a best guess based on the measured distance between the two “boxes”. Save matched text in spreadsheet to allow easy proofing and analysis.





+ Sheet 1

	A	B	C	D	E	F	G	H	I	J
1	file_id	filename	char_seq	char_id	char_text	table_id	table_sec	fr_text	viet_text	eng_text
2	1	view1.ht	1	1	Thày bói	6	6	Devin aveugle.	Thầy bói lòa.	Blind medium.
3	1	view1.ht	2	2	Ăn mày	7	7	Gestes du mendiant.	Động tác của người ăn mày.	Beggar's gestures.
4	1	view1.ht	3	3	Lấy ráy tai	8	8	Le cureur d'oreilles.	Thợ lấy ráy tai.	Ear cleaner.
5	1	view1.ht	4	4	Đẻ rơi giữa đườ	15	15	L'accouchement en dehors de	Đẻ rơi ngoài đường.	Childbirth outside the home.
6	1	view1.ht	NA	NA	NA	5	5	Rôle du coq dans la vie magiq	Gà trống trong trong đời sống tín	Role of the rooster in the world of magic
7	1	view1.ht	NA	NA	NA	9	9	Couteau à hacher la viande.	Dao băm thịt.	Butcher.
8	1	view1.ht	NA	NA	NA	10	10	Rabot.	Bào gỗ.	Plane.
9	1	view1.ht	NA	NA	NA	11	11	Couteau à bétel.	Dao cau.	Betel nut knife.
10	1	view1.ht	NA	NA	NA	12	12	Paysans revenant du marché.	Nông dân đi chợ về.	Peasants returning from market.
11	1	view1.ht	NA	NA	NA	13	13	Costume d'une femme riche.	Trang phục của người phụ nữ giàu	Rich woman's outfit.
12	1	view1.ht	NA	NA	NA	14	14	Génie gardien de la porte (ima	Thần giữ cửa (tranh dân gian).	Door Guardian Deity (folk print).
13	2	view2.ht	1	1	Giãi kẹo	5	5	Pétrissage du sucre dans la fal	Trộn đường làm kẹo.	Sugar mixing in kẹo candy making.
14	2	view2.ht	2	2	Hàng đồng	8	8	Vendeuse d'objets en cuivre.	Bà bán đồ đồng.	Seller of copper objects.
15	2	view2.ht	3	3	In sách. Thứ nh	6	6	Martelage des sandales.	Nện dép.	Planishing sandals.
16	2	view2.ht	4	4	Nạo quả dừa	10	10	Raclage d'une noix de coco.	Tách cùi dừa.	Scraping a coconut.
17	2	view2.ht	NA	NA	NA	7	7	Forgeron martelant une marm	Thợ rèn gò chôn nồi.	Blacksmith hammering a pot.
18	2	view2.ht	NA	NA	NA	9	9	Femme en train d'imprimer.	Người phụ nữ đang in.	Woman printing.
19	2	view2.ht	NA	NA	NA	11	11	Pâtissier en train d'aplanir la p	Thợ bánh đang dạt bột.	Pastry cook flattening the dough.
20	2	view2.ht	NA	NA	NA	12	12	Masque porté dans les maladi	Mặt nạ tránh đau mắt.	Mask worn for eye ailments.
21	2	view2.ht	NA	NA	NA	13	13	Tablette des ancêtres ; son co	Bài vị thờ có nắp đậy.	Ancestors' shelf, its cover.
22	2	view2.ht	NA	NA	NA	14	14	Oreiller en lanières végétales.	Gối mây.	Reed cushion.
23	2	view2.ht	NA	NA	NA	15	15	Coussins qui calent le mort da	Đệm kê trong quan tài.	Cushions to wedge the body in its coffin.
24	2	view2.ht	NA	NA	NA	16	16	Arc et flèche.	Cung tên.	Bow and arrow.
25	2	view2.ht	NA	NA	NA	17	17	Lance à tête incurvée.	Cái giáo đầu uốn cong.	Curved-headed spear.
26	2	view2.ht	NA	NA	NA	18	18	Métier de brodeur.	Khung thêu.	Embroiderer's trade.
27	2	view2.ht	NA	NA	NA	19	19	Gouges de menuisier.	Cái đục của thợ mộc.	Group of furniture makers.
28	2	view2.ht	NA	NA	NA	19	20	Gouges de menuisier.	Cái đục của thợ mộc.	Group of furniture makers.
29	2	view2.ht	NA	NA	NA	20	21	Guitare chinoise.	Đàn nguyệt cầm Trung Quốc.	Chinese guitars.
30	3	view3.ht	1	1	Đánh trống hát	6	6	Joueur de tambour.	Người đánh trống.	Drummer.
31	3	view3.ht	2	2	Cổ cầm: đàn n	7	7	Joueur de guitare.	Người chơi đàn.	Guitar player.
32	3	view3.ht	3	3	Tục ngôn [gọi là	8	8	Jeu analogue à notre "jeu de t	Trò chơi giống chơi đáo thùng ở Ph	Analogue game to our barrel game.
33	3	view3.ht	4	4	Đánh súc sắc	9	9	Đánh sắc (jeu).	Đánh súc sắc (trò chơi).	Đánh sắc (Game).
34	3	view3.ht	5	5	Cái giường giát	11	11	Étalage d'un restaurant.	Bày hàng ở một tiệm ăn.	Restaurant display.
35	3	view3.ht	NA	NA	NA	10	10	Chaussures en papier pour le	Giầy mã để thờ.	Paper shoes for worship.
36	4	view4.ht	1	1	Bán rau cải mới	13	13	Vendeuse de bananes et de fe	Bà bán chuối và rau cải.	Banana and turnip leaf seller.
37	4	view4.ht	NA	NA	NA	2	2	Cortège d'un mandarin dans la	Tháp tùng quan trong đời tư.	A Mandarin's procession in private life.
38	4	view4.ht	NA	NA	NA	3	3	Vendeuse ambulante du condi	Bà bán nước mắm rong.	An itinerant seller of nước mắm (condim
39	4	view4.ht	NA	NA	NA	4	4	Instrument pour couper les ho	Dụng cụ thái thuốc lào.	Instrument to chop rolled tobacco.

why python?

- easy to process large numbers of files and large amounts of data
- great tools for parsing html into more manageable formats
- powerful methods for pattern matching and text extraction
- easy to implement math expressions
- inherently readable and reusable
- open source with an enormous user and developer community, across many disciplines. Easy to find help.
- code runs on Linux, Windows, Mac