Katherine Bergen
May 1, 2018
Transforming Texts

"A View Inside a Messy Workshop": Unraveling the Possibilities of Textual Analysis with Optical Character Recognition

> *"Car il n'y à chofe qui appor-*
> *te plus degaing à fon maiftre,que la terre bié cultiuee, & rai-*
> *fonnablement entretenue."*
>
> – Charles Estienne, La Maison Rustique (1564), as transcribed in plain text on the Internet Archive[1]

If you read French, you might be able to make out the general meaning of the quote above. If not, allow a rough translation: "For there is nothing which brings more profit to its master than earth which is well-cultivated and reasonably maintained." This is as true for the early modern French husbandman as it is for the modern digital humanist. There is a wealth of digitized text available freely online for the aspiring digital historian to make use of, but much of it remains in a "raw" state, as the quote above. This quote comes from the plain text version of Charles Estienne's *La Maison Rustique* (1564), which was transcribed using OCR, or optical character recognition. While this software allows for a quick transcription of images into text, that automatically-rendered text is only a rough approximation of the information viewable to the human eye.

This project began as an attempt to check the work of my annotation in the previous semester of the Making and Knowing Project. In order to understand the relationship between Ms. BnF Fr. 640 and the genre of the agricultural manual, I was to isolate the areas where the manuscript overlapped with *La Maison Rustique*. By understanding these topics of mutual interest, I could then discern how the manuscript adhered or did not adhere to this genre. I used a

[1] Estienne, Charles. "L'agriculture Et Maison Rustique De M. Charles Estienne Docteur En Medecine… " Internet Archive. January 01, 1564. Accessed May 01, 2019. https://archive.org/details/bub_gb_RqaadcZV2mcC

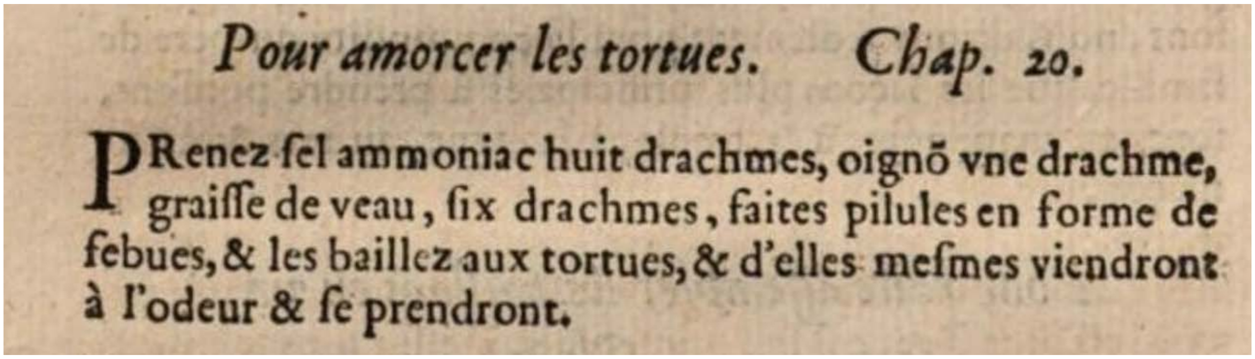
version of *La Maison Rustique* found on the Internet Archive[2]. This was the first available version that I found online, and I used its built-in search function to compare key vocabulary from the manuscript with the contents of Estienne's work. From the results of this search, I posited that where *La Maison Rustique* attempted to describe a whole system of natural phenomena that was interlinked and interconnected, the Ms. Fr. 640 seemed to aggregate piecemeal information about the natural world for the purposes of making a profitable product. Estienne was certainly concerned about profit as well, but much more so in terms of the household as an economic unit, rather than individual facets of nature that could be exploited for immediate gain.[3] This is quite a general statement, and I wondered whether a more "systematic" approach to the material would produce more revealing data. Instead of relying on the built-in search system to match key terms, what might be possible if I were to approach *La Maison Rustique* not as a static, linear document but as digital repository of textual data points?

What do I mean by digital repository? What I have learned from this semester's work is that documents are represented in the digital world in such a way as to make them as comfortable and familiar to human users as possible. The metaphor is extended from the desktop, to the folder, to the word processor with which I write this essay. Metaphor is used to smooth over the distances between a tap on the keyboard and the representation of a letter as a series of pixels on a screen. But it is possible, and indeed necessary, to do important work in the liminal spaces between the entry of information and its reception by a human audience. That information is recorded in a number of different forms, much as the Digital Critical Edition will include a

[2] Estienne, Charles. "L'agriculture Et Maison Rustique De M. Charles Estienne Docteur En Medecine… " Internet Archive. January 01, 1564. Accessed May 01, 2019. https://archive.org/details/lagricultureetma00esti
[3] See Katie Bergen, Fall 2018, "Annotation: "Dans La Maison Rustique": Cultivation in BnF Ms. Fr. 640 and the Genre Conventions of the Household Manual," https://docs.google.com/document/d/18erKiiynTri5JYu4IXJ4ph4h3a_PoPHz5nHpYoVjXPY.

number of different renderings of the information on any one folio. Different tools are useful in each different medium, so it is necessary to decouple the information a document contains from the way that information is constructed on a screen.

Take, for example, this entry from *La Maison Rustique*. Here it is as an image:



Pour amorcer les tortues. Chap. 20.

PRenez ſel ammoniac huit drachmes, oignõ vne drachme, graiſſe de veau, ſix drachmes, faites pilules en forme de febues, & les baillez aux tortues, & d'elles meſmes viendront à l'odeur & ſe prendront.

This image is a faithful reproduction of the graphic qualities of the original page: the format of the text is reproduced, as are the images around it. It is possible to see how the original printer of this page used the format of the text and its arrangement into blocks as a semantic tool to express the different parts of the entry: title, chapter number, text.

Conversely, here is that same information rendered in text through OCR, or optical character recognition:

Pour amorcer les tortues. Chap. 20.
|)Renez fel ammoniac huit drachmes, oignô vne drachme,
■* graiffe de veau , fix drachmes , faites pilules en forme de febues,& les baillez aux tortues, cV d'elles mcfmcs viendront à Todcur & fe prendront»

This text is the product of an algorithm attempting to match data from the image to textual characters. As you can tell, that matching process is only partially effective. The capital letter "P" has been rendered as a pipe character [|] and a close-parenthesis [)]. The indentation caused by the initial dropping onto the second line of text is rendered by the OCR as a black box and an asterisk. Text rendered in this way forms the database which the built-in search program from the Internet Archive website uses to match the keywords I type. As such, there are many instances of false positives and false negatives that distort the results of any such search.

It is clear, then, that the information represented on the screen to a human eye is different to the information available to a computer algorithm. As a historian, there were several questions I wished to be able to answer from the digital text of *La Maison Rustique*.[4] In order to do so, I would need to bring the human-readable and the machine-readable texts into some sort of agreement. To extend the metaphor, I would need to "prepare the earth", or "clean" my data, in order for any sort of fertile inquiry into the text to take place.

Were I to approach "cleaning" the entirety of the text, as a non-expert I would take the text to a university resource like a digital humanities lab in order to get their help. This exercise was not intended to replicate the work that a digital humanities librarian would do, but to give me an idea of how to ask for what I needed. What makes a data set "clean"? How would I want to be able to use it in its "clean" state? What are the best practices for achieving that state? Knowing how to ask for the help I would need proved an ample challenge.

From my list of historical questions, I knew I would want to be able to search with accuracy for key terms in the text. The more complex questions I wanted to answer would involve not just words in isolation, but phrases, contexts, and even idioms. To do this, I would want to use languages like Python or Shell to run tests on the text that would give me an accurate idea of the frequency of words in the text, and their relationship to each other. Therefore, the

[4] My original historical questions can be found here: https://github.com/tcatapano/LMR1564/blob/master/historical_questions.md

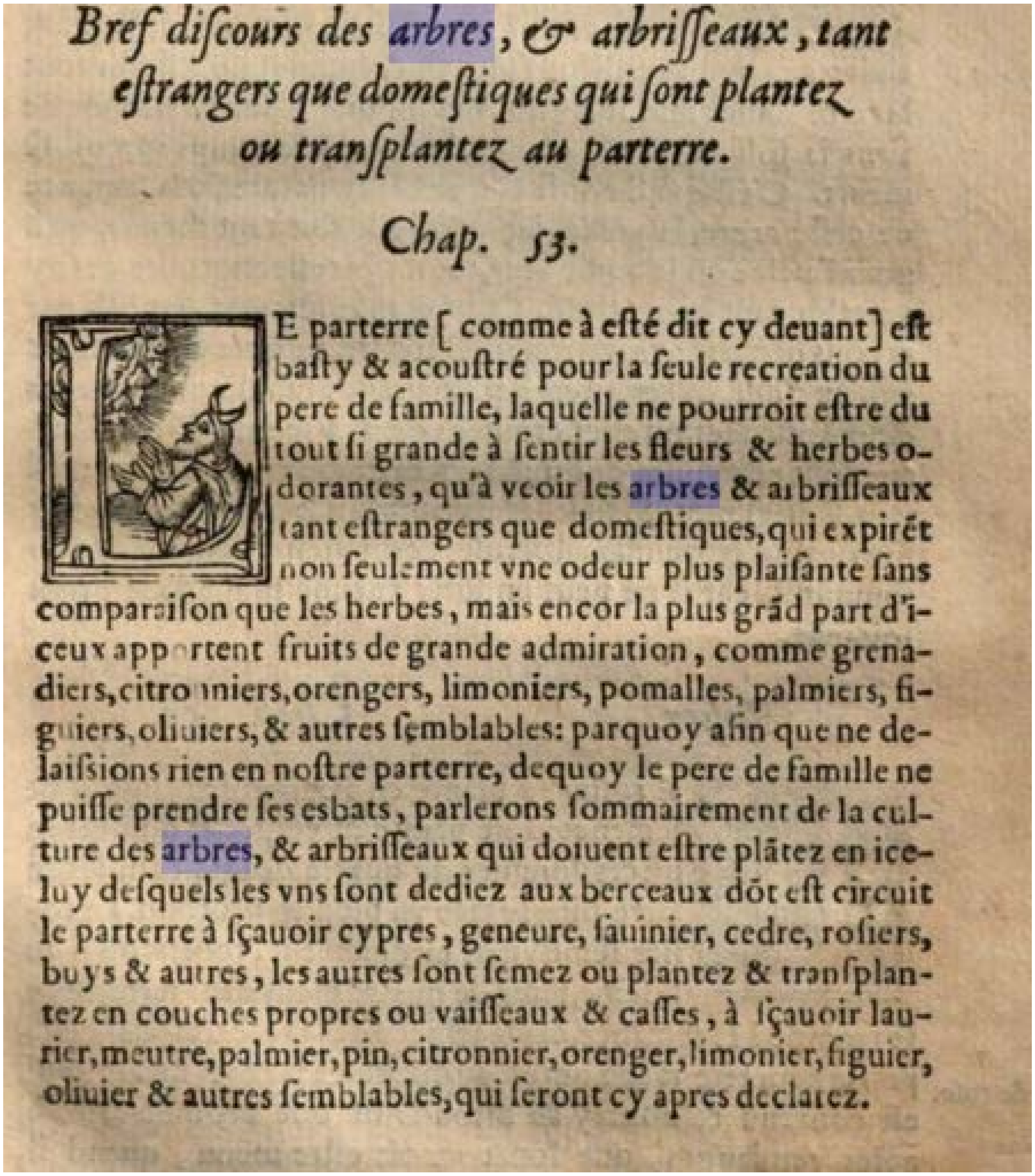
“clean” text would need to be relatively free of erroneous characters, misspellings, and variations in spelling (i.e. “to-day” vs. “today”).

There are several ways to approach the “cleaning” of OCR text in order to be able to manipulate it using languages like Python or Shell. Correcting for OCR errors in the entire body of text would be too lengthy a project for one semester, so I began by choosing a “chunk” of text, a chapter, in order to begin my assessment for how to proceed. Estienne divided his work into “books”, or volumes, which were broken up thematically. Book One deals with the house and the people who live there, Book Six deals with woodlands and hunting, and Books Two through Five deal with the meadows, gardens, pools, and pasturelands that form the permeable boundary between the wild and the domestic. Within each book, entries are divided into “chapters”. Unlike in a modern book, where a chapter can be a hundred pages long, these chapters are anywhere between 150 to 1,000 words. They are essentially miniature instructional essays, each on a single topic, arranged thematically within the book. These single topics can be precise, such as the example on turtles above, or they can be more generalized and extensive, such as a chapter introduced below on trees and their fertilization.

Terry Catapano was kind enough to seek out five different OCR text files of *La Maison Rustique*, hosted either at the Internet Archive or on Google Books. A few were different editions of the text, which went through several French-language printings in the sixteenth and seventeenth centuries. The remaining two were both the correct edition of the text, and they were both created using the same software, ABBYY. However, the version I ended up using, found here, was created using ABBYY 11.0, while the other version, found here, was created using

ABBYY 8.0.[5] Those three interstitial upgrades certainly made a difference in the readability of the text.

Here is the chapter I chose to work with, rendered as an image from the ABBYY 8.0 version:



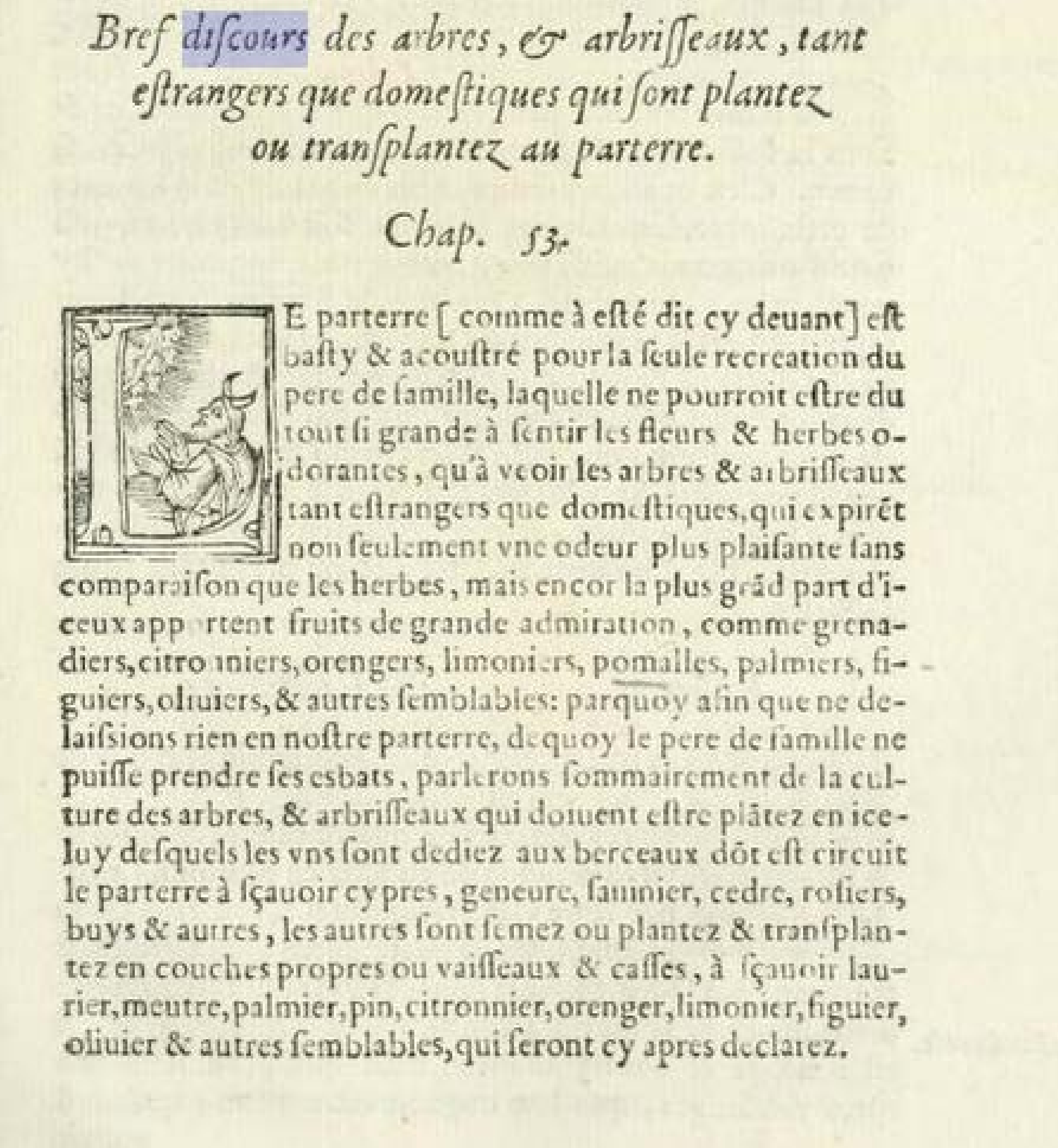
Bref diſcours des arbres, & arbriſſeaux, tant eſtrangers que domeſtiques qui ſont plantez ou tranſplantez au parterre.

Chap. 53.

LE parterre [comme à eſté dit cy deuant] eſt baſty & acouſtré pour la ſeule recreation du pere de famille, laquelle ne pourroit eſtre du tout ſi grande à ſentir les fleurs & herbes odorantes, qu'à veoir les arbres & arbriſſeaux tant eſtrangers que domeſtiques, qui expirẽt non ſeulement vne odeur plus plaiſante ſans comparaiſon que les herbes, mais encor la plus grãd part d'iceux apportent fruits de grande admiration, comme grenadiers, citronniers, orengers, limoniers, pomalles, palmiers, figuiers, oliuiers, & autres ſemblables: parquoy afin que ne delaiſsions rien en noſtre parterre, dequoy le pere de famille ne puiſſe prendre ſes esbats, parlerons ſommairement de la culture des arbres, & arbriſſeaux qui doiuent eſtre plãtez en iceluy deſquels les vns ſont dediez aux berceaux dõt eſt circuit le parterre à ſçauoir cypres, geneure, ſauinier, cedre, roſiers, buys & autres, les autres ſont ſemez ou plantez & tranſplantez en couches propres ou vaiſſeaux & caſſes, à ſçauoir laurier, meutre, palmier, pin, citronnier, orenger, limonier, figuier, oliuier & autres ſemblables, qui ſeront cy apres declarez.

[5] Estienne, Charles. "L'agriculture Et Maison Rustique De M. Charles Estienne Docteur En Medecine… " Internet Archive. January 01, 1564. Accessed May 01, 2019. https://archive.org/details/bub_gb_RqaadcZV2mcC (ABBYY 11.0); Estienne, Charles. "L'agriculture Et Maison Rustique De M. Charles Estienne Docteur En Medecine… " Internet Archive. January 01, 1564. Accessed May 01, 2019. https://archive.org/details/lagricultureetma00esti (ABBYY 8.0)

Here it is in the 11.0 version:



Bref diſcours des arbres, & arbriſſeaux, tant eſtrangers que domeſtiques qui ſont plantez ou tranſplantez au parterre.

Chap. 53.

LE parterre [comme à eſté dit cy deuant] eſt baſty & acouſtré pour la ſeule recreation du pere de famille, laquelle ne pourroit eſtre du tout ſi grande à ſentir les fleurs & herbes odorantes, qu'à veoir les arbres & arbriſſeaux tant eſtrangers que domeſtiques, qui expirēt non ſeulement vne odeur plus plaiſante ſans comparaiſon que les herbes, mais encor la plus grād part d'iceux apportent fruits de grande admiration, comme grenadiers, citronniers, orengers, limoniers, pomalles, palmiers, figuiers, oliuiers, & autres ſemblables: parquoy afin que ne delaiſsions rien en noſtre parterre, dequoy le pere de famille ne puiſſe prendre ſes esbats, parlerons ſommairement de la culture des arbres, & arbriſſeaux qui doiuent eſtre plātez en iceluy deſquels les vns ſont dediez aux berceaux dōt eſt circuit le parterre à ſçauoir cypres, geneure, ſaunier, cedre, roſiers, buys & autres, les autres ſont ſemez ou plantez & tranſplantez en couches propres ou vaiſſeaux & caſſes, à ſçauoir laurier, meutre, palmier, pin, citronnier, orenger, limonier, figuier, oliuier & autres ſemblables, qui ſeront cy apres declarez.

These two images are graphic depictions of the same text. The 8.0 version is perhaps of slightly worse quality - the ink is lighter and printing pressure has been applied unevenly. Still, these are clearly products of the same edition and imprint, down to the identical woodcut that begins the entry.

Now, compare those two clear images with the following text. Here is the OCR-generated text

from 8.0:

```
Bref difcohrs des arbres ^ ^ arhrijjeotux ytant

ejîrangers que dome^iques qui font plantez^

ou tranJpUntez^ au parterre,

Chap. S3f



^



1E parterre [ comme à efté dit cy deuant] eft
oafly & acourtré pour la feule récréation du
I père de (amille, laquelle ne pourroit cftre du
H ont (i grande à ftntirks fleurs & herbes o-
y^j(,vm'tàox2iV\iQ'i^c\\xï veoir les arbres & aibriflcaux
«2^^Jjj tant étrangers que domt.fl:iques,qui expirét
: — =^ij non feulement vne odeur plus plaifanie ians



comparcifonque les herbes, maisencor la plus g" âd part d'i-
ceuxapp rtent fruits de grande admiration , comme gtcna-
dierSjcitro miers,orengcrs, limoni.rs, pomalles, palmiers, fî-
guiers,oliuiers,& autres femblables: parquoy artn que ne de-
laifsions rien en noftre parterre, d^quoy le pcre de famille ne
puifle prendre fcscsbats, parkrons fommaircmenr df la cul-
ture des arbres, & arbriiïeaux qui doiuent tllrc plâtez en ice-
luy defquels les vns font dcdiez aux berceaux dôttft circuit
le parterre à f<jauoir cyprès, gcneurc, fauinier, cedrc, rolicrs,
buys& autres, les autres lont fcmez ou plantez & tranl'plan-
tezcn couches propres ou vaifleaux îS.' ca(fcs,à Tçuinir lau-
rier, meut re, palmier, pin, citronnier, orengcr, limonier, figuier,
oiiuier & autres femblablcs,qui feront cy après dcclaiez.
```

And here from 11.0:

Katherine Bergen
May 1, 2018
Transforming Texts
Bref dif ours des arbres, £7* arbrijfeaux 9 tant
cflrangers que domefiiques qu i font plante^
ou tranfplantez^au parterre.

Cbap. jj.

E parterre [comme à efté* dit cy deuant] eft
bafty & acouftré pour la feule récréation du
perc de famille, laquelle ne pourroit eftre du
tout (i grande à fentir les fleurs & herbes o-
dorantes , qu'à vcoir les arbres & aibriffeaux
tant cflrangers que domeftiques,qni expirét
ou feulement vne odeur plus plaiiante fins

comparrifon que les herbes , mais encor la plus grad part d'i-
ceuxapp rtent fruits de grande admiration, comme grena-
dicis,citro iniers,orengcrs, limoniers pomalles, palmicis» fi-
guiers,oliuiers,& autres femblablcs: parquoy afin que ne de-
laifsions rien en noftre parterre, dequoy le perc de famille ne
puifle* prendre fesesbats, parlerons fommairemenr de la cul-
ture des arbres, & arbrifTeaux qui doiuent eftre placez en ice-
luy defquels les vns font dédiez aux berceaux dôt eft circuit
le parterre à fçauoir cyprès , geneure, lâuinier, cedre, roiiers,
buys & autres , les autres font femez ou plantez & tranfplan-
tez en couches propres ou vai(Tcaux & cafTes , à fçauoir lau-
riir,mcutre, palmier, pin, citronnier,orenger,limonicr,figuicr,
oliuier & autres femblables,qui feront cy après dcclaicz.

Right away the differences become quite clear. While both “translations” are imperfect, certain assumptions made by each version of the algorithm become clear. Although both tend to interpolate “c”s for “e”s, the 8.0 version is less accurate when attempting to parse letters that exist in combination. For example, the word “arbrisseaux” is rendered as “arbriiïeaux” by the 8.0 version, and as “aibriffeaux” by the 11.0 version. Both are obviously incorrect, but where the 8.0 version interpolates an “i” and an “ï” for the medial s, the 11.0 version reverts them to “ff”. This is still incorrect, but more consistent, and therefore easier to “clean” using a Python algorithm. Similarly, the 8.0 version creates “domt.fl:iques” for “domestiques”, where the 11.0 version uses “domeftiques”. It is clear that the upgrades to ABBYY software have significantly improved the legibility of the text.

With that said, there are particular drawbacks in the OCR text which become apparent on further study. There are many places where the letter “e” is read as “c”, similarly the medial s is often read as “f”. The letter “t” is often read as “r”, “i”, or “l”, and the letter “h” as “b”. During the handpress period, type sorts were constructed for certain letters to appear in ligature, as “æ”, and this also often confounds the algorithm.

French orthography is another consideration when looking at “cleaning” a text. For the most part, accents seem to come through fairly well in the OCR text. Sometimes they come through where they do not belong, as in “lâuinier” for *lavinier*. Some accent marks are antiquated, which opens another set of questions in determining what a “clean” text would be. The suspension mark above words like “expirēt” is a holdover from scribal abbreviations or siglia which were used in manuscript production in Europe. These suspension marks indicate a missing letter, much as the modern circumflex (^) indicates a missing “s” in modern French (e.g. “hôtel”). In this case, the suspension mark often signals a missing “n” or “m”, much as it did in the manuscript age. Therefore, words like “expirēt”, or *expirent*, become “expirét” in OCR. This is a convention that invites further study. Other words, like “orenger” (*oranger* in modern French) are not contracted to "orēger". I wonder if this is because “expirēt” contracts the conjugated suffix of the root “expirer”, and the root is preserved whether it is contracted or not. To contract “orenger” and have it be comprehensible, the reader would need to be presumed to understand what an orenger was when spelled out fully. For a project like mine, which has dealt at least initially with nouns rather than verbs, it might be that a discussion of how to “clean” contractions is, for the time being, moot. Still, it's important to recognize that these contractions could cause false orthographical distinctions ("expirét" vs *expirent*).

At this point, it occurred to me that I might need to think around my problem of "clean" text. Which parts of the text would need to be cleaned, and which could I perhaps leave aside? It was suggested to me that I should create my own transcription of the image in order to establish a "gold standard". Below is my transcription of Chapter 53, Book One of *La Maison Rustique*:

Bref discours des arbres, & arbrisseaux, tant estrangers que domestiques qui sont plantez ou transplantez au parterre.

Chap. 53.

Le parterre [comme à esté dit cy deuant] est basty & acoustré pour la seule recreation du pere de famille, laquelle ne pourroit estre du tout si grande à sentir les fleurs & herbes o-dorantes, q'à veoir les arbres & arbrisseaux tant estrangers que domestiques, qui expirẽt non seulement vne odeur plus plaisant sans comparaison que les herbes, mais encor la plus grãd part d'i-ceux apportent fruits de grande admiration, comme grenadiers, citronniers, orengers, limoniers, pomalles, palmiers, figuiers, oliuiers, & autres semblables: parquoy afin que ne delaissions rien en nostre parterre, dequoy le pere de famille ne puisse prendre ses esbats, parlerons sommairement de la culture des arbres, & arbrisseaux qui doiuent estre plãtez en ice-luy desquels les vns sont dediez aux berceaux dõt est circuit le parterre à sçauoir cypres, geneure, sauinier, cedre, rosiers, buys & autres, les autres sont semez ou plantez & transplantez en couches propres ou vaisseaux & casses, à sçauoir lau-rier, meutre, palmier, pin, citronnier, orenger, limonier, figuier, oliuier & autres semblables, qui seront cy apres declarez.

Immediately, issues and caveats presented themselves. Should hyphens in words like "d'i-ceux" be maintained? How should contractions (as in the word "grãd") be rendered – in their full form, or as they are? Should u/v distinctions be maintained, or normalized (as in the word "sçauoir", or *savoir*). It would be my inclination to normalize and modernize the entire text, in effect translating it from sixteenth-century French to modern French. Given that spelling was phonetic and non-standardized, there could be a multiplicity of spellings and contractions that might conform inconsistently to conventional use.

With this issue in mind, it occurred to me that for the time being, my proposed historical inquiry was a simple one: which key words from Ms. Fr. 640 would correspond to topics in *La Maison Rustique*? The snippet of code below is designed to sort the 100 most common word types by the number of times they appear in the text.

```
      # open file and read contents into a list of lines

      with open('arbresgoldstand.txt', 'r') as f:
          lines = f.read().splitlines()


      from string import punctuation
      from collections import Counter

      tokens = []

      for line in lines: #for each line in the file lines
          for word in line.split(): #for each word in a line (string),
.split on white space (you should run from here and see what it does)
              tokens.append(word.strip(punctuation).lower()) #put stuff
into the tokens list. take words, take out the punctuation, make it
lowercase

      #display 100 most common types
      types = Counter(tokens) #types are the tokens list put through a
Counter function (? or just tool or whatever)
      types.most_common(100) #i can't believe most common is its own
method.
```

The output was as follows:

```
[('', 12),                    ('ne', 3),                   ('estre', 2),
 ('les', 5),                  ('en', 3),                   ('grande', 2),
 ('que', 4),                  ('des', 2),                  ('herbes', 2),
 ('qui', 4),                  ('tant', 2),                 ('plus', 2),
 ('parterre', 4),             ('estrangers', 2),           ('semblables', 2),
 ('à', 4),                    ('domestiques', 2),          ('sçauoir', 2),
 ('de', 4),                   ('plantez', 2),              ('bref', 1),
 ('autres', 4),               ('transplantez', 2),         ('discours', 1),
 ('arbres', 3),               ('comme', 2),                ('au', 1),
 ('arbrisseaux', 3),          ('cy', 2),                   ('chap', 1),
 ('sont', 3),                 ('est', 2),                  ('53', 1),
 ('ou', 3),                   ('du', 2),                   ('esté', 1),
 ('le', 3),                   ('pere', 2),                 ('dit', 1),
 ('la', 3),                   ('famille', 2),              ('deuant', 1),
 ('basty', 1),
 ('acoustré', 1),
 ('pour', 1),
 ('seule', 1),
 ('recreation', 1),
 ('laquelle', 1),
 ('pourroit', 1),
 ('tout', 1),
 ('si', 1),
 ('sentir', 1),
 ('fleurs', 1),
 ('o-dorantes', 1),
 ("q'à", 1),
 ('veoir', 1),
 ('expirẽt', 1),
 ('non', 1),
 ('seulement', 1),
 ('vne', 1),
 ('odeur', 1),
 ('plaisant', 1),
 ('sans', 1),
 ('comparaison', 1),
 ('mais', 1),
 ('encor', 1),
 ('grãd', 1),
 ('part', 1),
 ("d'i-ceux", 1),
 ('apportent', 1),
 ('fruits', 1),
 ('admiration', 1),
 ('grenadiers', 1),
 ('citronniers', 1),
 ('orengers', 1),
 ('limoniers', 1),
 ('pomalles', 1),
 ('palmiers', 1),
 ('figuiers', 1),
 ('oliuiers', 1),
 ('parquoy', 1),
 ('afin', 1),
 ('delaissions', 1),
 ('rien', 1),
 ('nostre', 1),
 ('dequoy', 1),
 ('puisse', 1),
 ('prendre', 1),
 ('ses', 1),
 ('esbats', 1),
 ('parlerons', 1),
 ('sommairement', 1),
 ('culture', 1),
 ('doiuent', 1),
 ('plãtez', 1),
 ('ice-luy', 1),
 ('desquels', 1),
 ('vns', 1),
 ('dediez', 1),
 ('aux', 1)]
```

This output shows that the words used most frequently are often pronouns and modifiers that are three characters or less in length. Until I need the full text for more sophisticated analysis, it might be worthwhile to focus only on longer words. Most descriptive nouns are longer than three characters, and this might be a simple way to sort useful vocabulary.

Still, I found myself looking for templates to help guide my next steps. It had become clear that “cleaning” text generated with OCR was not the simple proposition it had at first seemed. I had begun this project envisioning the “cleaning” step as a quick detour before beginning what I saw as the heart of the intellectual endeavor. Preparing this kind of raw data for scholarly use is a widespread project in the digital humanities, and it seemed to me that there must be some way to leap over the intellectual and procedural difficulties I was having. I found several useful resources among the many available online and on GitHub which helped me think through what a “clean” data set really is.

I began by searching for resources on cleaning early modern text generated using OCR. I found projects like EMOP, the Early Modern OCR Project at Texas A&M University led by Dr.

Laura Mandell, and the work of scholars like Ted Underwood and Konstantin Baierer.[6] EMOP's website demonstrates how holistic rendering text through OCR must be: Dr. Mandell optimizes the quality of each image, trains the OCR engine to recognize characters and fonts, and finally takes the post-production step of checking the results of the OCR for errors. My approach in its entirety was reduced to Dr. Mandell's third step. For EMOP, it seems the old saying rings true, that an ounce of prevention is worth a pound of cure. Many of the issues evident in my raw data from the Internet Archive might be avoided using a better-trained OCR engine custom-built for *La Maison Rustique*.

Similarly, Ted Underwood's Github repository proved to be an invaluable resource. He has provided detailed templates for attempting to clean up early modern text using OCR. However, his most useful piece of advice was in his readme on the DataMunging page of the repository. As he writes, "…[L]et's be frank: very little of this is plug-and-play. It's a view inside a messy workshop. Maybe, at best, it's a collection of resources you could cannibalize to build your own workflow." There is no best practice for cleaning early modern text, because each project presents unique challenges. Like an artisan in an early modern workshop, sometimes it is necessary to experiment and to build one's own tools for the job at hand.

Dr. Underwood goes on to suggest that the most useful part of his repository might be his lexicographical rules, the logical framework for any decision-making a cleaning algorithm might need to do when confronted with raw data. These rule sets are vast lists of possible switches and changes from the entirety of the corpus of texts he was using.[7] Each set of terms, from place

[6] The Early Modern OCR Project: https://emop.tamu.edu/ ; Ted Underwood. "Tedunderwood/DataMunging." GitHub. https://github.com/tedunderwood/DataMunging November 4, 2017. Accessed May 01, 2019. ; Konstantin Bairer. "Kba/awesome-ocr." GitHub. February 26, 2019. Accessed May 01, 2019. https://github.com/kba/awesome-ocr.

[7] See: https://github.com/tedunderwood/DataMunging/blob/master/rulesets/HyphenRules.txt

names to hyphenates, had been painstakingly harvested from a survey of Dr. Underwood's corpora in order to generate the correct terminology for his code to check against. The problem of the medial "s" would be solved by a compiled list of Ambiguous Pairs like "few" and "sew" which could be used to check against context within a sentence. The automation which produced the desired result of "clean" text was dependent on a framework of detailed close-reading of the text.

A final useful resource confirmed that cleaning OCR text requires specific, intensive detail work. In a short white paper given at the Midstates Conference for Undergraduate Research in Computer Science and Mathematics in 2015, William Rial and Sofia Visa laid out their process for attempting to engineer a workable OCR engine for early modern French text.[8] Much of their work, like EMOP, focused on the training of the engine itself. Their post-processing workflow (3.3) provides a useful logical framework for creating code to "clean" OCR text. Through a series of if/else propositions, they manage to sort for contractions, errors, and spelling. However, as in Dr. Underwood's case, they admit that "…to build this dictionary of errors we visually inspected and analyzed the OCR output of 30 pages and created a dictionary of all wrongly recognized words." Any systematized approach to cleaning OCR text must begin with a human-generated dataset of errors and corrections, and this set must be large enough to encompass a representative sample of the errors in the entire corpus.

These conclusions may seem obvious to the seasoned digital humanist. Even Charles Estienne would have known that preparing the soil for use is one of the most critical, difficult and time-consuming duties of a good husbandman. Just as soil does not become farmable

[8] William Franklin Rial and Sofia Visa, "A Framework for Using Tesseract to Transcribe Early Modern Texts Having Non-standard Fonts" (presentation, Midstates Conference for Undergraduate Research in Computer Science and Mathematics at Bowling Green State University, 2015.) https://www.cs.bgsu.edu/MCURCSM/proceedings/A-1.pdf

without effort and toil, digital tools are not miraculous workarounds with the magical ability to transform data. The digitization of text may make it more readily manipulable, but that manipulation still requires painstaking human effort. The digital environment is designed to lull its users into a false sense of ease and inevitability, but the work of a digital humanist requires a critical eye and the expectation of unforeseen complications. With that critical eye in view, it occurs to me that an algorithmic "cleaning" of this raw data may not be what is required to answer the questions I have. Were the corpus of data to extend to several editions of the work, or a library's worth of books, it would certainly be necessary to automate its transcription. For one work amounting to around 200 pages of text, such transcription might be more suited to old-fashioned hand-eye coordination between a human reader and her keyboard. There are transcription crowdsourcing tools available online which might aid in this work, and the amount of preparation, iteration, and postprocessing would be significantly smaller.[9] Future historians (including myself) may want to use the text of the DCE in comparison with other digitized early modern texts. We may count ourselves lucky that there is such a large amount of digitized information available openly on the internet. Still, this exercise granted me the opportunity to analyze deeply what that "information" really is, and how it is bound by the constraints of its medium.

[9] See: http://scripto.org/, http://t-pen.org/TPEN/, https://fromthepage.com/. Even large-scale projects have elected to use human transcription over or alongside automation, see: https://www.shakespearesworld.org.