Diffusion of Botanical Illustrations

Anh Duong Nguyen
Early Modern Group
Helsinki Digital Humanities Hackathon 2023

1 Introduction

This report concerns a side project extended from a group collective work 'Enlightening Illustrations: Analyzing the Role of Images in Early Modern Scientific Publications.' in the Helsinki Digital Humanities Hackathon 2023 (DHH23). Scientific illustrations are the main focus of the study, which emerged once techniques of graphical representation could applied to the panoply of natural philosophy (Ford, 2003). According to Ford (2003), botanical illustrations, in particular, have seen developments in using plant specimens to achieve realistic graphical representation and wide adoption of the new metal engraving technique. On the other hand, the practice of copying illustrations was arguably widespread, as producing such illustrations was labor-intensive (Ford, 2003) or copied elements were integrated by botanists into new customized illustrations that met their requirements (Nickelsen, 2006).

In this portfolio, I explore the reuse aspect of botanical illustrations and examine their diffusion based on the Diffusion of Innovation Theory.

2 Collecting botanical illustrations

During the group project, we identified a subset of botanical illustrations in books of Eighteenth Century Collections Online (ECCO), using customized Contrastive Language-Image Pretraining (CLIP) models. According to Vesalainen (2023), the botanical subset of illustrations contains roughly 13K botanical images. For this side project, I resampled a similar dataset according to results from two CLIP models, selecting images that were labeled by both models as botanical (see detailed implementation here).

Due to a technical issue with data retrieval, the final dataset used in the analysis contains 12119 images.

3 Search for image reuses - Using Resnet-50 image vectors

Resnet models are used in image classification, using deep learning techniques specifically Convolutional Neural Network (CNN). Within this architecture, low-level features (i.e. lines, edges...) can be learnt in the initial layers based on image inputs, and those features can be used to learn more abstract features (i.e. shapes, objects...) in deeper layers. Vesalainen (2023) and Suuronen (2023) used transfer learning from pre-trained open source models to extract features from the botanical illustration subset and compute cosine similarity metrics for image pairs - a "similarity score" with value between 0 and 1.

According model evaluation by Vesalainen (2023), the Resnet-50 model is better at detecting similarities in less similar images compared to the Resnet-18 model, though results are consistent when using similarity score threshold 0.94 to find image reuses. Peura (2023) continues close-reading evaluation and suggested Resnet-50 performed better and threshold 0.91 is enough to collect mostly correct identifications. I then set the threshold to 0.91 and collected a dataset of similar image pairs and their similarity score with over 4 million entries.

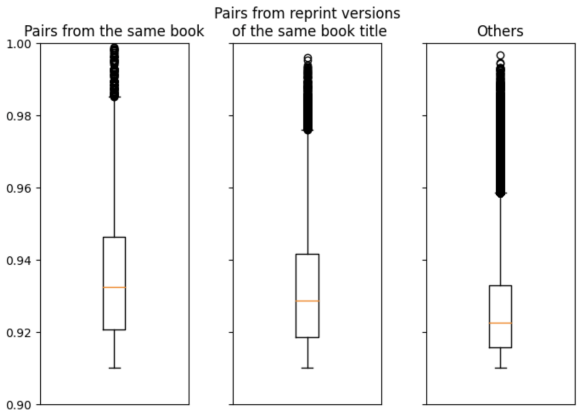
4 Illustration reuse analysis

In this section, illustration reuse detection is revisited to fit the purpose of downstream analysis. I experimented with several schemes with metadata of illustrated books, different thresholds, and a chain-like approach for detection.

Due to the focus on the diffusion of illustrations in downstream analysis, it is essential to detect how images were reused by unique book titles in chronological order. The image pair similarity result produced in the previous step, however, does not make a distinction between which picture in a given reuse

pair is reused and which is the reuse version. A simple sorting algorithm using cross-reference to publication years were performed to resolve this issue. Furthermore, there is no distinction between reuse pairs containing two images from different book titles, with those from the same book, from reprinted versions of a book title. Specifically, the similarity dataset contains over 172K reuse pairs (7.05 %) with illustrations from the same book, and over 468K reuse pairs (19.15 %) with illustrations from reprinted versions of the same book titles. The following Figure 1. shows how the cosine similarity scores differ with the three reuse pair types. Exact matches are likely to occur with pairs of images from the same book, and the median similarity score is slightly higher for reuse pairs from the same book or reprint versions of the same book title. For further analysis, only reuse pairs in category Others are used.

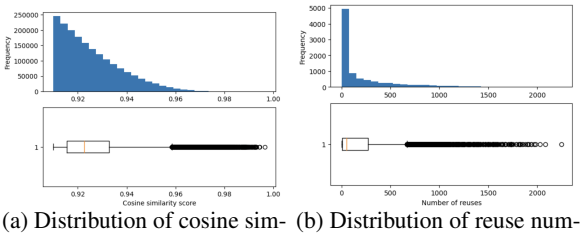
Figure 1: Cosine similarity scores of three reuse pair types



After filtering irrelevant reuse pairs, we are left with 1.8 million entries with a similarity score range of 0.91 - 0.996 (see Figure 2a). According to this new dataset, an image has roughly 204 reuses on average though there are many outliers with extensive reuses and half of the images have fewer than 49 reuses (see Figure 2b).

Another heuristic evaluation was done for qualified reuse pairs by selecting five random illustrations and some of their detected reuses for close reading. Figure 3 contains their top five most similar images (with a similarity score range of 0.947-0.98) while Figure 4 shows three random images detected as their reuses (with more diverse similarity scores). While the top five most similar reuses seem to capture similarities well, most of the randomly selected reuses do not share significant similarities with the query images.

Figure 2: Distribution of cosine similarity score and number of reuses (after removing irrelevant pairs)



(a) Distribution of cosine similarity score.

(b) Distribution of reuse number.

Figure 3: Five randomly selected query illustrations and their reuses - Top five most similar reuses

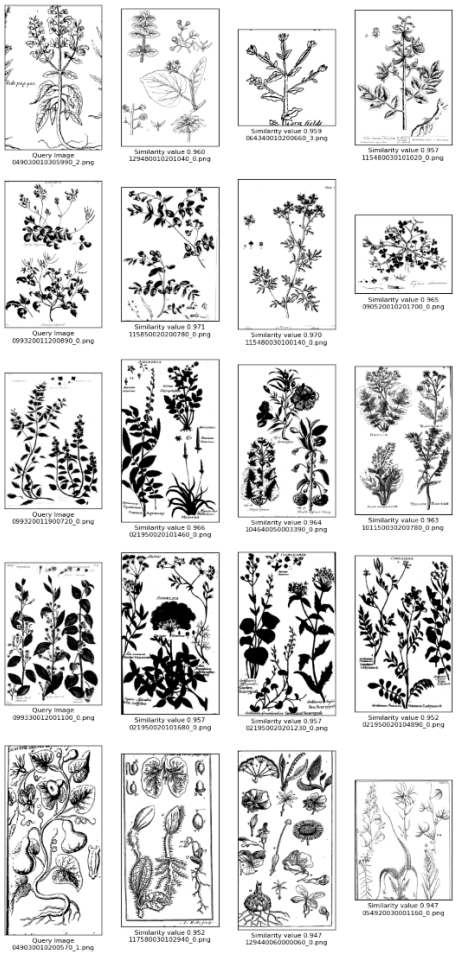
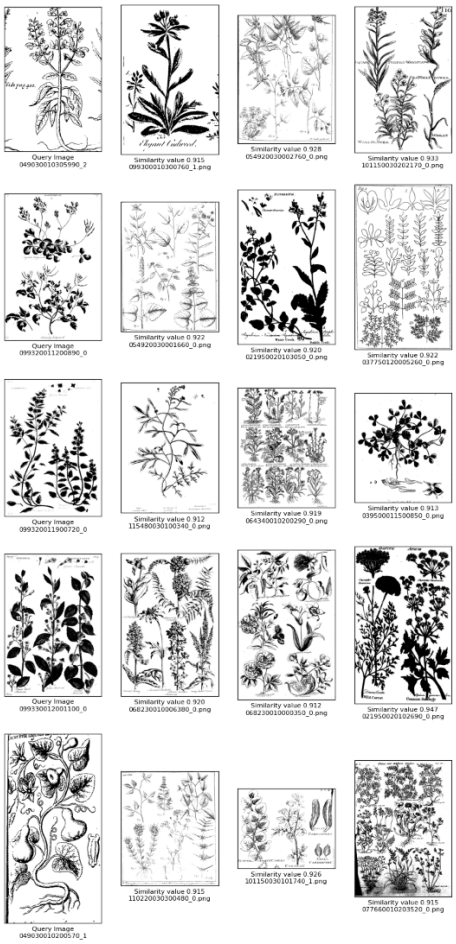


Figure 4: Five randomly selected query illustrations and their reuses - Three random reuses

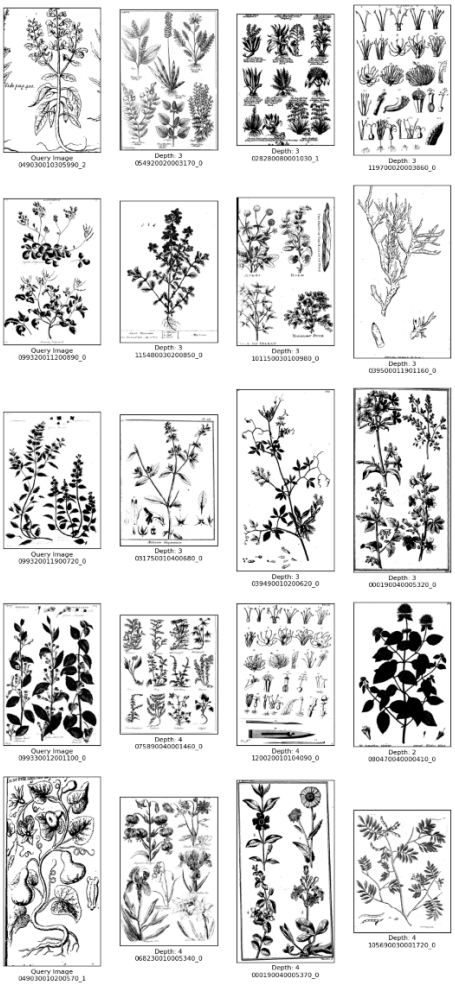


4.1 Side task 1: Chain of reuses

To approach the previously observed decline in reuse detection, I experimented with a chain-like approach to retrieve reuses using a higher threshold. Specifically, this approach pushes the threshold up to return only exact matches as reuses (depth 1), then collects exhaustively all reuses of reuses (depth 2 to n) which are also identified by the same high threshold.

The experiment was conducted on the same set of query illustrations in Figure 3 and Figure 4, while the threshold is set to 0.95 based on the similarity score of their top five most similar reuses. Figure 5 shows three random reuses detected for each query illustration using this algorithm. While the similarities detected by this approach are not consistent, we can see that the chain captures relevant similarities that were not visible when using a lower threshold (0.91) even when it gets deeper. This approach has its own limitations such as requiring much more computations and challenging to select and validate optimal threshold(s). Though I decided to use a medium threshold of 0.94 for further downstream analysis (as suggested by Vesalainen (2023)), there remains space to explore the potential of this approach.

Figure 5: Three random reuses detected for query illustrations by the chain of reuses algorithm



5 Examining the Diffusion of Innovation Theory with image reuses

5.1 Diffusion of innovations and botanical illustrations

The Diffusion of Innovation Theory was proposed by an American communication theorist and sociologist Everett M. Rogers in 1962. It provides a theoretical framework to study "diffusion as a process by which (1) an innovation (2) is communicated through certain channels (3) over time (4) among the members of a social system" (Rogers, 2014, p.). One important contribution of the framework is an S-shaped curve of adoption/diffusion to describe

the distribution of adopters along the time dimension. Figure 6 shows the curve derived by Rogers from data of the study by Ryan and Gross (1943) on hybrid seed corn in two Iowa communities, adapted by Valente (1993). The concept behind this model is "adoption rate" - the relative speed with which an innovation is adopted by members of a social system and the S-shaped curve results from plotting adopters on a cumulative frequency basis over time.

Figure 6: Rogers' diffusion curve

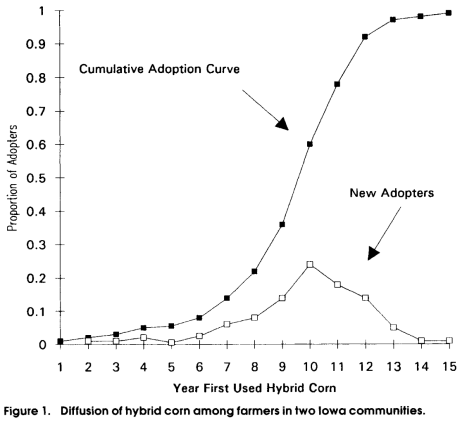


Figure 1. Diffusion of hybrid corn among farmers in two Iowa communities.

Furthermore, Rogers' framework also contains extensive conceptualization of major themes in prior diffusion research such as earliness of knowing about innovations, rate of adoption of different innovations in a social system, innovativeness, opinion leadership, diffusion networks, rate of adoption in different social systems, communication channel usage, and consequences of innovation. The perspective of diffusion is relevant to botanical illustrations in 18th-century scientific publications as Nickelsen (2006) notes copying as a central part of their production practices by draughtsmen. By adopting the diffusion perspective in studying historical scientific illustrations, scholars gained a powerful analytical tool to understand several facets of this phenomenon:

- What accounts for the earliness of reusing an illustration/illustrative element?
- How and why did different illustrations/illustrative elements get reused at different rates?
- As adoption or non-adoption of new illustrations/illustrative elements can be seen in illustrated works, what accounts for their innovativeness?

Table 1: Descriptive statistics

	cossim	year_1	year_2	work_1	work_2
count	246839.00	246591.00	246591.00	166.00	166.00
mean	0.95	1742.46	1771.76	1486.98	1486.98
std	0.01	25.22	19.48	6451.86	4593.41
min	0.94	1686.00	1686.00	1.00	1.00
25%	0.94	1715.00	1760.00	3.00	5.00
50%	0.95	1756.00	1769.00	38.50	50.00
75%	0.95	1759.00	1787.00	243.25	449.75
max	1.00	1800.00	1800.00	55019.00	32043.00

- Are there illustrated works with opinion leadership, or in other words, works that popularized certain illustrations/illustrative elements?
- How do illustrations/illustrative elements diffuse through a network of production and publications?
- Do illustrations/illustrative elements diffuse at different rates in different publication markets?
- How did draughtsmen and botanists learn about certain illustrations/illustrative elements and make decisions about whether or not to copy them?
- What are the consequences of adopting certain illustrations/illustrative elements?
  ... so on

In this analysis, I validate the S-shaped diffusion model using reuse data of botanical illustrations, discuss finding indications, and review further prospects of study using this diffusion approach.

5.2 Data

The dataset for analysis has 246839 entries/reuse pairs, with the cosine similarity score, and enriched with metadata such as their respective publication years and book titles. Descriptive statistics of the data can be found in Table 1 (cosine similarity: cossim, the publication year of reused illustrations: year_1 and of illustration reuses: year_2, the book title of reused illustrations (number of occurrences): work_1 and of illustration reuses: work_2).

5.3 Method

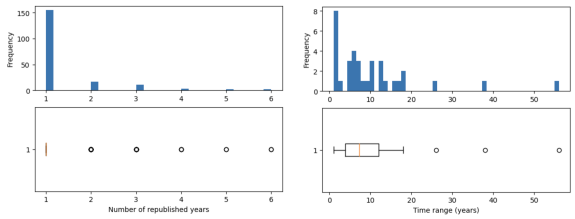
As a property of a specific innovation within a social system, Roger introduced the concept of "rate of adoption" - the relative speed with which an innovation is adopted by members of a social system. The S-shaped curve results from plotting the

cumulative number of adopters. The number of adopters is usually plotted as a proportion to the total number of members in the social system in question, as in Figure 5. To account for these metrics, I first consider an adoption happens when an illustration/illustrative element is reused in a book title. This is a limited view of adoption: it takes into account the fact that an illustration/illustrative element can be adopted by a book title A at time t_1 though A first appeared in the book market at time t_0, while missing information such as the illustration might have not been reused in a reprint version of A at a later time point t_2. On the other hand, I define the social system as the botanical book market archived by ECCO, and the total members in this system as the number of book titles in the book market over time.

5.3.1 Side task 2: Botanical book market over time

To concisely grasp the botanical book market, I reviewed the recorded botanical book market and compared two approaches to measuring potential adopters. Firstly, the dataset recorded only one publication year for 155 over 190 botanical works, only 2 works reached the maximum 6 different years of publication/republication. For long-seller works (those with two or more publication/republication years), the average time between two recorded publications is 10 years, though the time gap is less than 7.5 years for over half of the long-sellers. Figure 6 shows the distribution for the number of publication/republication years and the average time gap between recorded publications.

Figure 7: Publication of botanical works

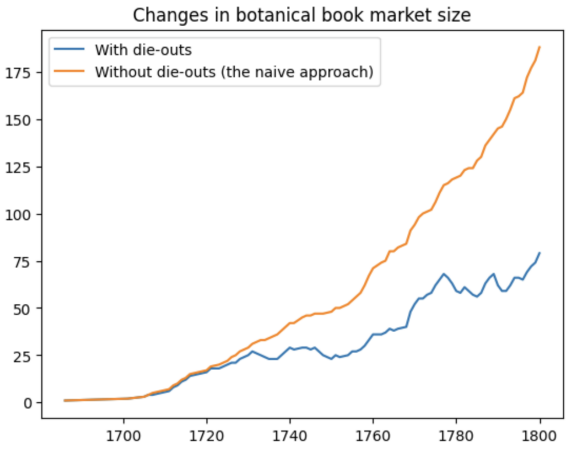


(a) The number of publication/republication years

(b) The average time gap between recorded publications

The naive approach to measuring the total members of the system is a market size growth, or in other words, the number of new works published each year. However, the dataset suggests that most illustrated botanical works are likely to die out from the market after a certain time, for example only three over 190 works got republished after more than 20 years from their last (re)publication. It is therefore unrealistic to expect works to adopt new illustrations after a long time since their last publication, since the works themselves might potentially never get republished. The alternative approach takes into account this aspect of the book market, reducing the market growth with the number of die-outs to achieve a more realistic measurement for the total members of the book market (Figure 7).

Figure 8: Number of total members of the botanical book market over time



5.4 Results

Overall, the curves for the accumulative share of adopters of illustrations appear not to perfectly fit the S-curve model, though resemblance can be spotted in several examples in Figure 9 and Figure 10. Most illustrations went through several diffusion waves as indicated by the adoption rate/new adopters line, especially those with diverse illustrative elements. We might expect different elements in a collage-like illustration had their reuse moments in different time. Some random examples in Figure 9 have either very few reuses (the first image) or got published late during the studied time period (the two last images), resulting in non typical curves. Top performers were mostly published before 1760s, many of them was identified to be reused in total over 30% of botanical works at their peaks. Their cumulative share of adopters curves, therefore, display higher resemblance to Roger's S-shaped diffusion curve.

Figure 9: Five randomly selected illustrations and their diffusion curves (in chronological order)

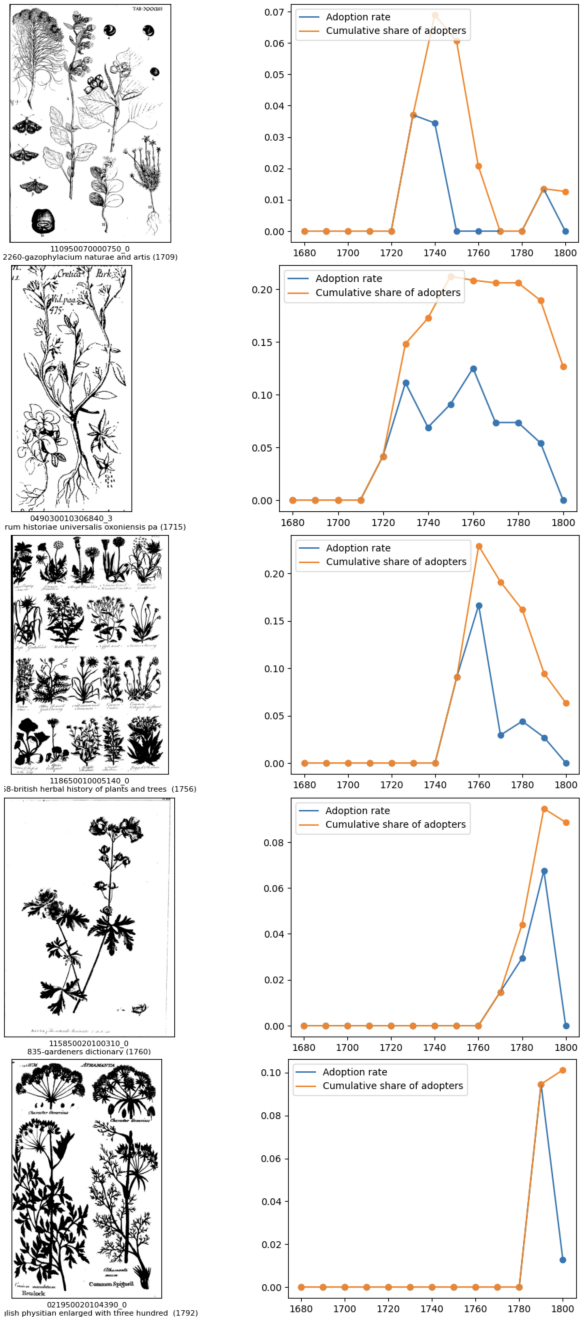
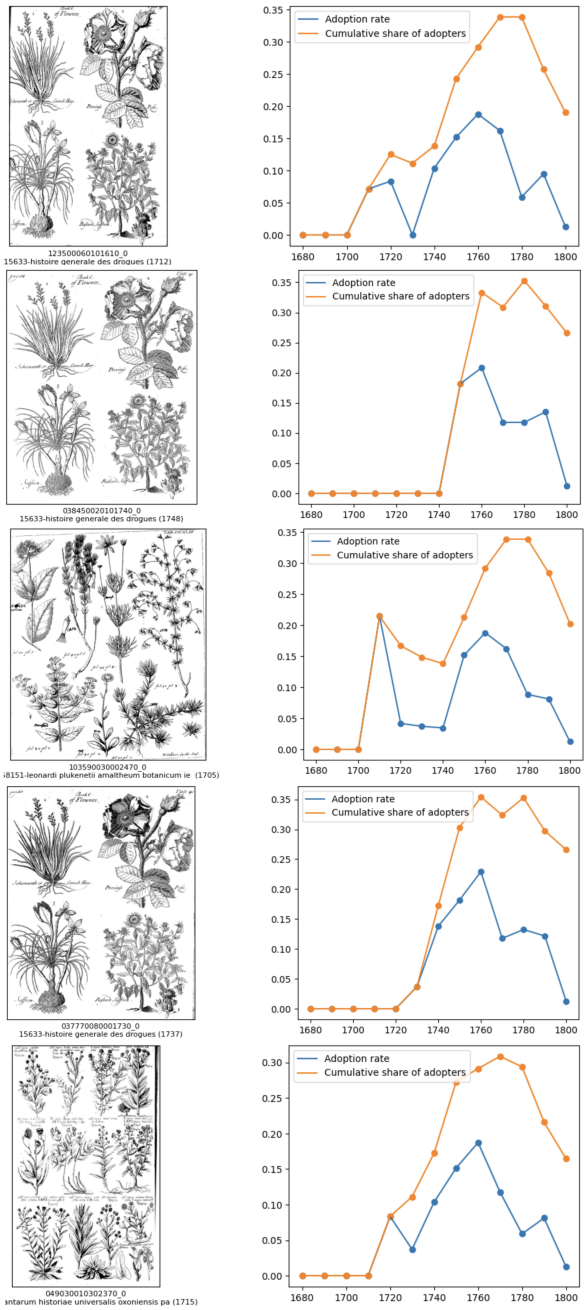


Figure 10: Five images from the top 10 most reused illustrations and their diffusion curves



The images are from different books (can be from reprint versions of on book title), and are sorted based on their popularity

While Roger did not model any decline in his diffusion curve, we see a common declining trend in the diffusion curve of top performers from 1780s. Given that the book market did not grow substantially compared to earlier decades in terms of book titles, the decline is unlikely a result of the diversifying book market, rather occurs as adopters dying out. Another distinguished trend in the results, differing from what suggested by the S-shaped curve model is the disrupted slopes. This is especially evident in the first and third example of Figure 10. Roger argues that while most innovations have an S-shaped rate of adoption, they vary in the slope of this S curve where the slope signals more rapid adoption as stiffer it gets. As illustrations with few elements (i.e. second and fourth example of Figure 8) have relatively stable slopes, it is possible that different elements in one illustration have different diffusion patterns at different time points.

6 Final thoughts

In this portfolio, I reviewed the progress of data exploration and enrichment by Early Modern Group in DHH23 and extend it to a case study on diffusion of botanical illustrations. The ECCO archive has gone through several phases of enrichment: book genre classification, image segmentation, illustration type classification, and image similarity detection; two of which were revisited. In particular, I focus on image similarity detection as it is fundamental to the case study, exploring an under-performing area of the detection model, and experimented an alternative approach to existing similar image detection scheme. More specifically, detection performance appears less efficient with image pairs that are not from the same book or book title, indicating stylistic similarities largely account for similarity metrics. The accuracy remarkably declines when measuring similarity of two illustrations that are less stylistically similar.

Illustration reuses among botanical books in eighteenth century has provided a novel data set to examining Roger's Innovation Diffusion Theory. Results are mixed regarding the relevance of S-shaped diffusion curve proposed in the theory, as illustrations were reused and thus diffused in patterns that do not necessarily fit Roger's model. Main divergences of the data from the model include an declining trend from 1780s onwards and a few disrupted slopes in diffusion pattern. Approaches to operationalizing innovation (illustrations with certain segmentation from archived books), the studied social system and its members (botanical book market and unique book titles) and adoption (machine-detected illustration reuse) all exert direct impact analysis results, and are critical elements to consider when interpreting observed divergences. Apart from data quality issues such as missing publications from the archive and errors in data enrichment, limitations of the theoretical framework should also taken into account such as a static representation of social system.

Future study can explore alternative approaches and evaluating them on image similarity detection. Such foundational work is believed to contribute tremendously to data enrichment quality and enables diffusion study with higher precision. Furthermore, other questions in diffusion study (see Section 5.1 for more) can be investigated further using both quantitative and qualitative methods.

References

Brian J. Ford. 2003. Scientific Illustration in the Eighteenth Century. In Roy Porter, editor, *The Cambridge History of Science*, 1 edition, pages 561–583. Cambridge University Press.

Kärin Nickelsen. 2006. Draughtsmen, botanists and nature: constructing eighteenth-century botanical illustrations. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 37(1):1–25.

Telma Peura. 2023. Re-exploring reuse in scientific publications.

Everett M. Rogers. 2014. *Diffusion of innovations, 5th edition*. Free Press. OCLC: 893102250.

Aleksi Suuronen. 2023. Digital Humanities Hackathon 2023: Detecting Similar Images.

Thomas W. Valente. 1993. Diffusion of Innovations and Policy Decision-Making. *Journal of Communication*, 43(1):30–45.

Ari Vesalainen. 2023. Image similarity for the 18th-century illustrations.