Revitalizing Digitized Historical Collections: AI-driven Quality Control with Locally Trained Computational Models

Vincent Wai-Yip LUM
Digital Initiatives
The Chinese University of Hong Kong Library
vincentlum@cuhk.edu.hk

Siu Sun CHUNG, *Senior Member*, IEEE
Technology Support Services
The University of Hong Kong Libraries
sunchung@ieee.org

***Abstract*—This paper explores AI-driven approaches by applying methodological paradigms to Quality Control (QC) for digitized collections. We propose a practical framework – the first automated QC process covering pivotal operations and workflows in digitizing ancient and fragile collections. Using a dataset of digitized images (primarily from a Chinese Rare Book Collection, we explored supervised Deep and Machine Learning algorithms for defect detection and computational methods to measure the color variations using Jensen-Shannon (JS) divergence. Our Convolutional Neural Network (CNN) model demonstrated high accuracy for our target, making them favorable for detecting digital defects in this context. The experimental results indicate that automation can significantly reduce human labor in quality assurance, thereby facilitating further character recognition accuracy and seamless access to online academic collections.**

Keywords—AI, Quality Control, Machine Learning, Deep Learning, Computer Vision, Digitization, Process Automation, Library Collection, Human-in-the-loop

I. Introduction

A. Importance of Digitization

Historical collections have transcended their role as mere artifacts, transforming into digital repositories that preserve heritage and historical values. Their availability empowers researchers to unlock profound knowledge, providing an important insight into the multifaceted society and cultural circulation of that era and broadening the horizons of digital humanities research and its practical applications. The generation of extensive digitized historical collections has provided researchers across diverse domains with a powerful tool to advance their investigative endeavors.

Hirtle [1] discussed the impact of digitization on special collections in libraries through a qualitative analysis of the benefits and challenges posed by digitization. It examines the potential decrease in the use of paper originals due to the availability of digitized primary source resources on the internet, based on observations and anecdotal evidence. Hirtle's study reflects on the past decade's pioneering digitization projects to draw conclusions about the benefits and changes brought about by digitization in special collections.

Significant national funds have been allocated to universities for digital library research, leading to the implementation of various digital library and archives projects to preserve historical materials in digital formats [2]. These projects aim to provide greater opportunities and convenience for researchers to access important historical materials easily, with a focus on usability. The positive impact of digitization is often associated with increased access to historical materials, supporting the rationale for digital projects alongside enhanced preservation benefits. Digitization also aids in preserving aging materials that could otherwise face extinction [3], enhancing accessibility and searchability of library collections, and large-scale digitization of books has received significant attention in the past decade [4]. Given the fragile condition of historical items such as Ancient Chinese Rare Books, each physical access risks irreversible damage. Digitization serves as a critical preservation strategy, allowing fragile items to be accessed once for digitization, with their digital copies accessible perpetually. This endeavor seeks to unlock the classic and historical library materials seamlessly accessible online for the public good.

B. The role of QC in Digitization

Quality control plays a crucial role in various industries worldwide, ensuring the production of high-quality goods and services. From ancient times to the modern era of Industry 4.0 and artificial intelligence, quality control has evolved significantly [5]. It is critical for media digitization projects to ensure the digital files produced meet the organization's preservation specifications and mitigate the risk of errors. Effective quality control leads to higher-quality products, customer satisfaction, and competitiveness [6]. Quality control is crucial for a company's success as well, particularly in manufacturing. It ensures the production of high-quality products. The process, tools, and approaches are essential for maintaining quality standards in industry [7].

Quality control is crucial in the digitization process to ensure the digitized materials meet the required standards and produce high-quality, accurate, and valuable information and records [8]. The specific defects in a digital file can lead to re-digitization and additional rework for any corrections.

Institutions face various issues and challenges in sustaining quality control, which can lead to the ineffectiveness of the digitization process. It is important for organizations to be aware of these issues and find ways to handle and resolve them. The implementation of quality control can be done using both machines and manpower, and the differences between these approaches are discussed in [8]. Balancing these sometimes-conflicting imperatives will continue to pose a challenge for professionals in the GLAM (galleries, libraries, archives, and museums) sector.

Poor quality control can result in blurry pages, which adversely affect Optical Character Recognition reading [9]. OCR is an important element for digital scholarship, and it is the first step in research on text analysis. With the advancements in text-recognition technologies applied to digitized materials, image archives are facing increasing pressures from users to return accurate textual results during the digitization process and demand high-quality results from QC.

The accuracy of the digitized items can also contribute to the quality of Metadata, which goes hand in hand with *Metadata Quality Control in Digital Repositories and Collections* [10].

C. Challenges in QC

Visual checks are immensely time-consuming. McFarland notes the difficulty of creating a time-effective quality control workflow with visual checks [11], discussing "quality problems that took us a great deal of time to isolate and correct." Developing countries, such as Nigeria, encountered challenges such as inadequate funding for purchasing and maintaining modern digital equipment [3], hindering the digitization process. To minimize the time cost, [12] suggested that visually, "reviewing a reasonably-sized sample of images allows us to find some of these problems, but more crucially, allows us to identify recurring errors so that we can find ways to prevent them." However, in real cases, it is difficult to identify the recurring pattern of defects in digitization, so the sampling approach is risky despite the time-saving. The cost and benefit of quality control checks in large-scale digitization are discussed in [9], which provides some insights into making informed decisions about allocating time and resources and developing digital workflows that efficiently publish large quantities of materials online while continuing to meet user needs to the best of their ability. Reference [13] discussed the bottlenecks in quality control for digitization projects and proposed conducting cost-effective quality control such as combining automated and manual Quality Assurance (QA).

Different QC methods were discussed in [14], including:

1. X-Bar Chart

Randomly selected products are tested for the given attributes the chart is tracking. A common form of a quality control chart is the X-bar chart, where the x-axis on the graph tracks the samples tested, and the y-axis tracks the degree to which the variance of the tested attribute is acceptable.

2. Taguchi Method

The Taguchi method gauges quality as a calculation of loss to society associated with a product. In particular, loss in a product is defined by variations and deviations in its function as well as detrimental side effects that result from the product.

Loss from variation in function is a comparison of how much each unit of the product differs in the way it operates. The greater that variance, the more significant the loss in function and quality. This could be represented as a monetary figure denoting how usage has been impacted by defects in the product.

3. 100% Inspection Method

This 100% inspection method is a quality control process involving looking at and assessing all product parts. This type of quality control is done to rule out flaws in products.

We can see that performing all the visual checks as gone through in [15] is not so trivial given the complexity of each QC checking, and it does create some challenges here. We attempted to break down those barriers with the introduction of AI so that the top known defect cases of our institution can be checked with automation instead of human inspection.

II. Motivations

Our library collections house a wealth of unique and invaluable resources and archival materials that serve as crucial sources for teaching and research endeavors. The users demand a low level of flaws in our digital collections, and we perform 100% inspection during our QC process. Given the large volume of materials to be digitized annually, an automation framework is essential for optimizing resources and budget. This study proposes a holistic framework, representing the first automated QC process that includes pivotal operations and workflows in the aforementioned realms. We explored supervised Machine Learning and Deep Learning algorithms and computational methods for defect detection. The experimental results indicate that automation can significantly reduce human labor in quality assurance, thereby facilitating the seamless availability of online collections. Ultimately, the open collections empower researchers to uncover profound insights within historical and cultural contexts, enhancing the accuracy of Optical Character Recognition (OCR) due to higher-quality digital images, thereby benefiting text-mining processes and broadening the landscape of digital humanities research [16].

III. Related Work

Automated Quality Control in smart factories ensures products meet defined criteria using deep learning. It plays a crucial role in maintaining product standards and customer satisfaction in Industry 4.0 [17].

This process is typically performed by human experts; unfortunately, experts make mistakes frequently as the task is tedious and tiring, even for well-trained operators. Another alternative is to equip a computer-aided system with the ability to inspect and recognize objects automatically [18]. The use of computation together with evaluated algorithms allows us to have a capable tool for automatic inspection of product quality. Thus, automation becomes necessary for inspection and recognition of objects in order to guarantee the quality of a product. The paper [19] discussed the issue of automatic inspection of the quality of an object using Computer Vision (CV). Two use cases were explored with a computer vision system. The first case was to detect fabric color defects in the textile industry, whilst the second was to detect the defects in golden apples. This study concluded that in order to obtain a better performance of the system, the acquisition of knowledge from human experts and the techniques to represent it in terms of numerical information is mandatory.

Digital pathology (DP) refers to the digitization of tissue slides, which involves the clinical diagnostic workflows and demands a tool for rapidly performing quality control to not only identify and delineate artifacts but also discover cohort-level outliers (e.g., slides stained darker or lighter than others in the cohort). Reference [20] employed a combination of image metrics (e.g., color histograms, brightness, contrast), features (e.g., edge detectors), and supervised classifiers (e.g., pen detection) to identify artifact-free regions on digitized slides. This referenced project, *HistoQC*, could provide an automated, quantifiable means of quality assurance for identifying artifacts, and this can improve both the repeatability and robustness of DP workflows. We also applied similar techniques to our QC process to identify digitization defects with a combination of image metrics in Computer Vision and supervised classifiers (folded line detection).

Statistical process control (SPC) strategies were suggested by [21] for QC on digital imaging in the cultural heritage community. SPC program naturally involves monitoring selected output parameters. These could be sampling rate (dpi), resolution, noise, and tonal/color fidelity. By way of a suitable target artifact and analysis software, each of these variables can be measured, and their values can be compared to pre-established numerical aims and error bounds. Corrective action is taken when trends approach or exceed these bounds. Despite the benefits of SPC in industry, which are recognizable in terms of quality, efficiency, and economy, most defect cases resulting from digitization would be difficult to analyze by quantitative methods with statistical models like our top known defect cases, which we would go through next. Institutions invest tremendous amounts of time, money, and human capital to perform the QC process after the digitization of their collections. Automation was proposed to relieve the cost of the QC process. An automatic method for quality assurance of scanned content based on computer vision was presented in [22]. A visual dictionary derived from local image descriptors enables efficient perceptual image fingerprinting in order to compare scanned book pages and detect duplicated pages. Nevertheless, this study can only check the page duplication with computer vision.

IV. Methodologies

In this paper, we present a novel AI approach with Computer Vision and Machine Learning and Deep Learning algorithms for building a smart Quality Control (QC) system for digital images after the process of digitization for the library materials. To build our knowledge base towards those problematic cases for the learning system, we have performed some systematic reviews on the past reported defect cases that required rework or re-digitization for corrective actions. Finally, we have compiled a list of the top known defect cases that we have identified in the past, and it spreads around 90% of our total defect cases that require further follow-up. The six primary visual defects of the digitized images in our quality assessments are listed as follows:

1. **Skewness**: page misalignment
2. **Blurriness**: out-of-focus or blurred content
3. **Folded line appearance**: folds or creases on the page
4. **Color tone deviation**: abnormal color balance
5. **Invalid image cropping**: cut-off or excessive boarders
6. **Transparency issue**: showing reverse side of the page

A. Feature Extraction: Image data analysis and preprocessing

The process begins with loading and converting the image to a suitable format for analysis. A key aspect of this analysis involves identifying the dominant color, which is presumed to be the paper's original color. Hence, we apply color tolerance to extract ink elements from the image by setting colors close to the paper's hue to white.

Techniques such as noise reduction and contrast enhancement helped isolate the causes of blurriness and prepare the images for Machine Learning models.

B. Defect detection with Computer Vision and Machine/Deep Learning

To tackle the six key issues illustrated above, we utilize Computer Vision and Machine Learning and Deep Learning approaches to perform AI-based QC:

1. **Skew Detection**: A *Convolutional Neural Network* (CNN) based deep learning model detects the skew angle of scanned pages. Through multiple (80, in each evaluating set) epochs training, the model learns to identify misaligned text lines, classifying the degree of skewness for readability. In the hidden layers of the CNN model, the central area of the page is scanned with 96x96 pixels convolution layers to initiate model learning, with three hidden layers and two additional fully connected layers as deep learning CNN infrastructure. Training data was divided into 0.75:0.25 (train : validation) train split ratio.

2. **Blurriness Detection**: Detecting blurriness is crucial for preserving the quality of historical materials. The method includes image cropping to remove irrelevant borders, followed by edge detection using the *Canny algorithm*.

The gradient (*Gx, Gy*) at any pixel location (*x, y*) in the image is calculated using partial derivatives in the *x* and *y* directions:

$$G_x = \frac{\partial I}{\partial x}, \quad G_y = \frac{\partial I}{\partial y}$$

The magnitude of the gradient (*G*) is then computed as:

$$G = \sqrt{G_x^2 + G_y^2}$$

Our solution also uses *Laplacian variance* and *Fast Fourier Transform* (FFT) to detect motion blur and sharpness:

For a given image *I* (*x, y*), the *Laplacian* (ΔI) is computed as:

$$\Delta I(x, y) = \frac{\partial^2 I(x, y)}{\partial x^2} + \frac{\partial^2 I(x, y)}{\partial y^2}$$

The variance of the *Laplacian*, *V ar(*ΔI*)*, is used as a measure of sharpness:

$$Var(\Delta I) = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} (\Delta I(x, y) - \mu_{\Delta I})^2$$

Where *M* and *N* are the dimensions of the image, and $\mu_{\Delta I}$ is the mean of the *Laplacian*.

Using this blur metric, we quantify the degree of blurriness in each image and compare *the computed results* with human QC judgments as the ground truth. Thus, we can obtain the blurriness threshold to quantify the blurriness level, which requires further follow-ups.

Folded Page Detection: Our solution is designed to identify fold lines in digitized book pages, particularly from historical collections. The dataset consists of two primary labeled folders: one containing images of pages with visible folds (as samples of ground truth) and the other with images that are free of folds. These folders serve as the foundation for training the model. To ensure the model is exposed to a wide variety of data while maintaining a balanced representation between the fold/non-fold categories, data augmentation is employed—each image is resized, transformed, and rotated to produce eight additional variants—ensuring the CNN sees a wide variety of fold and non-fold examples.

Before feeding the data into the CNN, pixel values are normalized ([0, 1] to [-1, 1] range). This normalization step ensures that the model performs efficiently across different lighting and contrast variations in the dataset.

While data augmentation improves the model's training accuracy by providing more balanced data, excessive augmentation could lead to inaccuracies in labeling, making the model less reliable. Therefore, we adopted a moderate augmentation approach to strike a balance between accuracy and the quality of data.

Similar to the skew detection model, a five-layer CNN is built in a 0.75:0.25 train-validation split ratio. Also, class-weighting is applied in the loss function to address class imbalance. The model contains over 500,000 parameters. Training progressed through multiple epochs, with initial training losses observed at various checkpoints during each epoch, see Table 1 and Fig. 1.

Table 1. The Accuracy, AUC and Loss across Epoch

Epoch	[80 / 259] Loss	[160 / 259] Loss	[240 / 259] Loss	[259 / 259] Loss	AUC	Loss
1	0.759063	3.104599	3.444041	1.035591	0.653846	1.273647
2	0.684241	0.846577	0.689463	0.749385	0.480769	2.182197
3	0.655190	0.663523	0.645279	0.575642	0.461538	2.129458
4	0.536806	0.639113	0.513243	0.602721	0.519231	1.949687
5	0.524882	0.489347	0.481206	0.601150	0.576923	1.789975
6	0.070386	0.028039	0.000000	0.000000	0.981000	0.759910

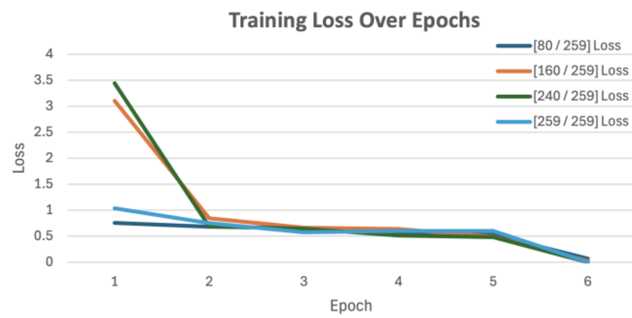

Fig. 1. Training Loss Over Epochs

3. **Color Deviation**: The computer vision method adopted Jensen-Shannon (JS) divergence [23] to compare image color distributions and detect outlier images based on their color profiles. JS divergence is bounded symmetrizing as in [24].

The primary goal is to measure how much the color distribution of individual images deviates from the overall color distribution of all images in the particular book or collection.

4. **Cropping Issues**: For supervised machine learning, a manual image labeling routine is conducted. Images are annotated by allocating specific coordinates that are saved for reference by the classifier we later trained using an Adam optimizer with weight decay. The loss function used is *CrossEntropyLoss*. After each epoch, the model is evaluated on a validation set, computing metrics such as *accuracy*, *Area Under the Curve* (AUC), and *loss* to assess performance.

5. **Transparency Issue:** The computer vision algorithm detects ink bleeding through pages by comparing weak and strong edges using edge detection techniques and Laplacian variance to classify transparency issues.

C. Dataset preparation

1) Dataset Limitation

Problematic (True Defect) cases are very limited as GLAM only kept the re-worked files since the original files with defects were not preserved, leading to the following observations in our workflow:

1. **Limited Variety of Training Data**: If the images in the training set do not encompass a wide range of defect types, the model could become biased toward the majority class, leading to poor performance in detecting minority class and failing to generalize to unseen data.

2. **Imbalanced Dataset**: If the number of non-defect images far exceeds the true defect images (or vice versa), the model may become biased toward the dominant class, leading to poor performance in the minority class.
3. **Insufficient Data**: The size of the dataset might be too small to train the model effectively, especially when using a complex neural network like CNN. This can result in the model memorizing training examples instead of learning general features.

2) Training set

Our dataset is collected with 908 labeled images and tested among 34,284 images in the partial sample of digitized images from the Chinese Rare Book Digital Collection [25]. As the materials from this collection are fragile and historical, they were more likely to contain defects than contemporary collections. All these images had undergone manual QC previously, so we could use the existing QC report as ground truth labels for ML and provide a validating foundation for our models.

V. Results

In our comprehensive quality control process, the combined accuracy for detecting defects across six categories— (1) skew, (2) blur, (3) fold, (4) invalid cropping issues, (5) color deviation, and (6) transparency—stands at approximately 85%. However, setting thresholds for defect detection presents a challenge. When thresholds are set too lenient (low), we observe high recall rates, meaning the system identifies most defective cases. While this ensures sensitivity to potential issues, it comes at the cost of specificity, resulting in a significant number of false positives (FP). These incorrectly flagged cases slow down our workflow, requiring additional manual verification and reducing overall efficiency.

Conversely, setting thresholds too high to prioritize precision reduces FP but risks missing genuine defects, which is particularly problematic in the digitization of rare and fragile materials. The detection of subtle defects, such as ink transparency or misaligned cropping, is crucial for ensuring the quality and usability of the scanned collection.

Balancing *precision* and *recall* is essential in academic library settings, where preserving the integrity of rare collections is paramount. A fine-tuned approach, with well-calibrated thresholds, is required to spot true defects while maintaining workflow efficiency. Thus, ongoing refinement of thresholds and model performance is necessary to meet both operational demands and preservation standards.

VI. Conclusions and Future Work

The presented tool suite detects the aforementioned defects in the digitized images and presents them for additional manual analysis and corrective re-digitization. A systematic overview of evaluation data and performance characteristics like accuracy is delivered. The results of the analysis confirm our hypothesis that an automated approach can detect errors with reliable quality, thus making quality control for large digitization projects a feasible and affordable process. We still keep humans in the loop to assess the performance of AI; we consider it essential to apply responsible AI to visual checking on our precious collections.

Acknowledgments

We are also grateful to Mr. Joshua Yan Ho SEN and Mr. Jeremy Tin Yau IP for their technical support on Python coding. This project would not have been possible without their help and insights.

References

[1] P. B. Hirtle, "The impact of digitization on special collections in libraries," *Libr. Cult.*, vol. 37, no. 1, pp. 42–52, Winter 2002, doi: 10.1353/LAC.2002.0012.

[2] D. Sinn, "Impact of digital archival collections on historical research," *J. Assoc. Inf. Sci. Technol.*, vol. 63, no. 8, pp. 1521–1537, July 2012, doi: 10.1002/ASI.22650

[3] K. O. Jagboro, B. O. Omotayo, and W. A. Aboyade, "Digitization of library collection in developing countries: The Hezekiah Oluwasanmi Library experience," *Lib. Philos. Pract.*, 2012.

[4] M. Conway, "Archival quality and long-term preservation: A research framework for validating the usefulness of digital surrogates," *Archival Science*, vol. 11, no. 3, pp. 293–309, 2011.

[5] L. D. Kounis, Ed., *Quality Control - An Anthology of Cases*. IntechOpen, 2023. doi: 10.5772/intechopen.97961.

[6] A. Dhasmana, "Quality control and management in manufacturing industries: A critical analysis," *Turkish Online J. Qual. Inquiry*, 2023. doi: 10.52783/tojqi.v11i3.10012.

[7] M. Zacharias, "The importance of quality control for the success of a company," *Asian J. Logistics Manag.*, vol. 1, no. 2, pp. 99–106, 2022. doi: 10.14710/ajlm.2022.16787.

[8] N. F. Roslan, S. F. M. Yatin, N. A. Azman, and N. S. C. Soh, "Digitization of records and archives: Quality control process," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 8, no. 9, 2018. doi: 10.6007/ijarbss/v8-i9/4586.

[9] J. Chapman and S. Leonard, "Cost and benefit of quality control visual checks in large-scale digitization of archival manuscripts," *Library Hi Tech*, vol. 31, no. 3, pp. 405–418, 2013. doi: 10.1108/LHT-01-2013-0002.

[10] J.-R. Park and Y. Tosaka, "Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms," *Cataloging Classif. Quart.*, vol. 48, no. 8, pp. 696–715, 2010. doi: 10.1080/01639374.2010.508711.

[11] Library of Congress and Ameritech National Digital Library. (1996-1999). *Lessons Learned: Workflow and Project Management* [Online]. Available: http://lcweb2.loc.gov/ammem/award/lessons/workflow.html. Accessed: Aug. 1, 2024.

[12] J. Riley and K. Whitsel, "Practical quality control procedures for digital imaging projects," *OCLC Syst. Services: Int. Digit. Library Perspect.*, vol. 21, no. 1, pp. 40–48, 2005.

[13] X. Lin, "Quality assurance in high volume document digitization: A survey," in *Proc. Second Int. Conf. Document Image Anal. Libraries (DIAL'06)*, IEEE, 2006. doi: 10.1109/DIAL.2006.33.

[14] A. Hayes, "Digitization quality control workflow," *Investopedia*. [Online]. Available: https://www.investopedia.com/terms/q/quality-control.asp.

[15] SHN, *Guide to Quality Control and Quality Checklist*, 2016. [Online]. Available: https://sustainableheritagenetwork.org/system/files/atoms/file/GuidetoQualityControlandQualityChecklists.pdf.

[16] Fenlon, K., Frazier, E., & Muñoz, T., Digital Humanities. In D. Baker & L. Ellis (Eds.), *Encyclopedia of Libraries, Librarianship, and Information Science* (First Edition) pp. 501-510. Academic Press, 2025. doi: 10.1016/B978-0-323-95689-5.00140-1

[17] S. Mandapaka et al., "Application of automated quality control in smart factories—A deep learning-based approach," in *2023 IEEE Int. Syst.*

Conf. (SysCon), Vancouver, BC, Canada, 2023, pp. 1–8. doi: 10.1109/SysCon53073.2023.10131100.

[18] Ahsanul Akib, Kamruddin Nur, Suman Saha, Jannatul Ferdous Srabonee, and M. F. Mridha, "Computer Vision-Based IoT Architecture for Post COVID-19 Preventive Measures," *Journal of Advances in Information Technology*, Vol. 14, No. 1, pp. 7-19, February 2023.

[19] O. O. Vergara-Villegas et al., "Automatic product quality inspection using computer vision systems," in *Lean Manufacturing in the Developing World*, Springer, 2014, pp. 135–156. doi: 10.1007/978-3-319-04951-9_7.

[20] A. Janowczyk et al., "HistoQC: An open-source quality control tool for digital pathology slides," *JCO Clin. Cancer Inform.*, no. 3, pp. 1–7, 2019. doi: 10.1200/CCI.18.00157.

[21] M. Stelmach and D. Williams, "When good scanning goes bad: A case for enabling statistical process control in image digitizing workflows," in *Archiving Conf.*, vol. 3, no. 1, pp. 237–243, 2006. doi: 10.2352/issn.2168-3204.2006.3.1.art00054.

[22] R. Graf et al., "Duplicate detection approaches for quality assurance of document image collections," in *Proc. Fifth Int. Conf. Manag. Emergent Digit. EcoSyst. (MEDES '13)*, ACM, 2013. doi: 10.1145/2536146.2536157.

[23] J. Lin, "Divergence measures based on the Shannon entropy," in *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145-151, Jan. 1991, doi: 10.1109/18.61115.

[24] F. Nielsen, “On the Jensen–Shannon symmetrization of distances relying on abstract means,” *Entropy*, vol. 21, no. 5, pp. 485, 2019, doi: 10.3390/e21050485

[25] *Rare Book Digital Collection*, CUHK Digital Repository. [Online]. Available: https://repository.lib.cuhk.edu.hk/en/collection/chi-rarebook.