Assessing the Impact of Image Resolution on OCR Transcription Accuracy

Toufik Boubehziz[1], Caroline Koudoro-Parfait[2], Gaël Lejeune[3]
(1) CNRS, Arts et Métiers institute of technology, 75013 Paris (France)
(2) Fachbereich III - Neuere Geschichte, Universität Trier, 54296, Trier, (Germany)
(3) STIH/CERES, Sorbonne Université, 28 rue Serpente, 75006 Paris (France)

***Abstract*—While higher resolution is often assumed to yield better Optical Character Recognition (OCR) accuracy, this comes at the cost of increased storage requirements and longer processing times. For digital libraries a question remains open : what is the optimal resolution in which documents should be stored. Obviously, one might expect that the highest resolution would be the best choice but chosing the best input quality has an impact on data storage and computing time and the real influence of image resolution (and size) on OCR and subsequent tasks seems to remain an open question. High-resolution images typically allow OCR engines to better distinguish character features, leading to improved recognition performance. Conversely, low-resolution images often result in increased character ambiguity, misclassifications, and noise, thereby reducing overall OCR reliability. These recognition errors not only compromise the immediate output quality but also propagate into downstream text processing tasks such as information retrieval, named entity recognition, and natural language understanding.**

In this paper we investigate the relationship between image resolution and OCR performance, with a focus on both character-level accuracy and the integrity of subsequent text processing pipelines. By analyzing OCR outputs across a range of resolutions and evaluating their impact on various post-recognition tasks, we seek to identify resolution thresholds that balance processing efficiency with textual fidelity. The findings have practical implications for document digitization workflows, especially in resource-constrained environments where high-resolution image storage and processing may be questionable.

I. Image resolution and OCR performance

Despite remarkable progress in Optical Character Recognition (OCR) algorithms over recent decades, the quality of input images continues to play a decisive role in determining the accuracy and reliability of the transcribed text. As digital libraries, archives, and heritage institutions increasingly depend on OCR for large-scale digitization of historical collections, a critical question arises: what is the optimal resolution at which documents should be stored and processed? While higher resolution is often assumed to yield better recognition accuracy, this comes at the cost of increased data storage requirements and longer processing times. Conversely, lower-resolution images should reduce these computational costs but seem likely to lead to misclassifications, character ambiguities, and transcription noise that propagate into downstream applications such as information retrieval, named entity recognition, and natural language understanding. This paper revisits a long-standing assumption in the field—that 300 dots per inch (dpi) represents the optimal resolution for OCR—and evaluates its validity in the context of modern OCR technologies and large-scale digital library workflows. We position our work from the perspective of using OCR for large-scale analysis on French-language corpora. In this study, the images submitted to OCR underwent no preprocessing, reflecting a realistic use case in which non-technical users may wish to apply OCR without extensive knowledge of image enhancement workflows. We do not assume that low-resolution OCR is a universal solution; rather, we investigate its potential within practical research contexts where computational cost, storage requirements, or the availability of high-quality scans may be constrained. At the same time, we fully acknowledge that for purposes such as high-quality physical printing or digital scholarly editions, retaining high-resolution images remains essential so that end users can enlarge and inspect them as needed. However, the choice of the resolution has a crucial influence on important matters like computing time for OCR or data storage size for large online datasets. By systematically analyzing OCR performance across a range of image resolutions and compression types, we seek to identify resolution thresholds that balance accuracy, storage efficiency, and computational performance. Our findings are particularly relevant for heritage institutions, where millions of pages of historical material—often scanned under varying conditions and at inconsistent quality levels—form the backbone of cultural and linguistic research. In Section II we elaborate on the context of this research, then in section III we present the experiments we designed. In section IV we present our results and propose a discussion and future perspectives in section V.

II. Background and Motivation

The digitization of textual heritage has expanded dramatically since the 1990s. National and research libraries have led large-scale digitization efforts, making millions of pages accessible. These initiatives have relied heavily on OCR technology to transform scanned page images into machine-readable text. Projects such as Germany's OCR-D Long-Term Archive Project[1] illustrate the scale and ambition of such efforts. However, the image quality across these collections is highly variable, and evaluation of OCR accuracy often remains opaque or inconsistently reported. User studies and surveys have revealed that OCR-generated texts frequently fail

[1]https://ola-hd.ocr-d.de/

to meet the quality needs of researchers. For instance, an OCR-D community survey[2] showed that while around 60% of users employ OCR transcriptions despite errors, 40% deem them unusable due to excessive noise. This reflects an enduring challenge: the gap between the potential of OCR technology and the practical limitations imposed by image quality, particularly for historical materials subject to degradation, variable typography, and inconsistent scanning standards. The widely cited "300 dpi rule" emerged from early industry practices and vendor recommendations in the 1990s, when hardware and OCR systems were calibrated for this resolution as shown by [2] and [3]. However, this value became a de facto standard more by convention than by evidence. Modern OCR systems such as Tesseract [3] [21], ABBYY FineReader[4] [23], and PaddleOCR[5] [4] employ advanced machine learning and deep learning models that may behave differently under varying image conditions. Yet, there has been little systematic study of how these contemporary systems perform below the canonical 300 dpi threshold. In the context of digital libraries, this issue is not merely theoretical. Storing documents at 300 dpi resolution or higher entails substantial storage and maintenance costs—potentially terabytes or petabytes of data—while many scanned images available online, especially those captured in the early 2000s, exist only at 72–90 dpi [5]. This makes it crucial to determine whether meaningful OCR results can be obtained from such low-quality inputs and to identify the lowest possible resolution that still produces usable transcriptions.

Historical literature has often reiterated the 300 dpi recommendation without fully re-examining its empirical foundations [25], [17], [8]. The early benchmark by [18] demonstrated that Tesseract's performance deteriorated significantly below 200 dpi, suggesting that 300 dpi provided a practical trade-off between legibility and file size. Subsequent studies, such as [12], found that OCR accuracy peaked at 300 dpi but could even decline at 600 dpi, indicating that excessively high resolutions may introduce diminishing returns or even degrade performance due to increased noise and pixel artifacts. Recent years have seen a resurgence of interest in re-evaluating this standard. Advances in deep learning have enabled OCR models to generalize across diverse visual conditions, prompting researchers to explore OCR performance at lower resolutions. For example, [8] demonstrated that OCR accuracy falls dramatically between 300 dpi and 72 dpi, but their study did not address intermediate thresholds. More recent neural OCR architectures, such as RNN-LSTM-based systems proposed by [24], and vision-based approaches inspired by human reading strategies [7], attempt to enhance recognition robustness even under severe resolution constraints. Despite these advances, the question of whether high-quality OCR can be achieved on low-resolution images remains unresolved, and ongoing research projects continue to explore strategies for improving input image quality [16].

[2]https://ocr-d.de/en/survey

[3]https://tesseract-ocr.github.io/tessdoc/

[4]https://help.abbyy.com/assets/en-us/finereader/16/Users_Guide.pdf

[5]https://www.paddleocr.ai/main/en/index.html

Moreover, studies have investigated the downstream consequences of OCR errors—such as their impact on information extraction [1], named entity recognition as [15], [9], vizualisation [14] or text classification [11], but only few works specifically addressed the relationship with image quality in French [6], [22], [13]. This study addresses this gap by coupling quantitative accuracy metrics with lexical analyses that capture how OCR-induced noise affects the linguistic composition of the text.

III. Methodology and Experimental Design

To explore the relationship between image resolution and OCR accuracy, this study employs a controlled experimental setup. The data source is the European Literary Text Collection (ELTeC)[20][6], an open-access corpus comprising novels in over twenty European languages. The French subset (ELTeC-fra) includes 100 novels published between the mid-19th and early 20th centuries, making it representative of typical digitized historical material. For the experiments, eleven French novels were selected to form a smaller benchmark corpus—small ELTeC-fra—with detailed metadata on resolution and image quality, each text was available in a verified ground-truth version (from ELTeC) and a scanned version obtained from the Gallica platform. The scanned pages exhibited various resolutions well below the canonical 300 dpi threshold. To simulate controlled degradation, these images were further downsampled to multiple resolution levels ranging from 30% to 100% of the original image quality. The Tesseract OCR engine was used for transcription, chosen for its stability, wide adoption in digital humanities research, and multilingual support. Our experiments allowed us to evaluate OCR performance using multiple quantitative and qualitative measures:

1) Character Error Rate (CER) and Word Error Rate (WER) – measuring the proportion of character- and word-level mismatches against the reference text.
2) Cosine Similarity – assessing textual similarity through vectorized bag-of-words representations or word embeddings.
3) Lexical Diversity Metrics – analyzing vocabulary richness via the Type-Token Ratio (TTR) and normalized TTR to detect noise-induced word inflation.
4) File Size and Processing Time – assessing trade-offs between image format (JPEG, PNG, BMP, RAW) and computational efficiency.

The evaluation proceeded in three phases:

1) Comparison of OCR accuracy across multiple resolutions (30–100%) to identify the degradation threshold.
2) Benchmarking against ground-truth ELTeC texts to quantify absolute transcription reliability.
3) Noise analysis through lexical statistics, examining how OCR errors distort vocabulary distributions.

[6]https://www.distant-reading.net/eltec/

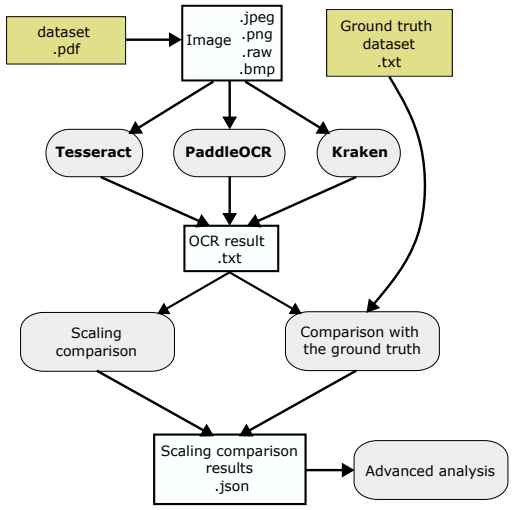


Fig. 1: Full analysis procedure diagram

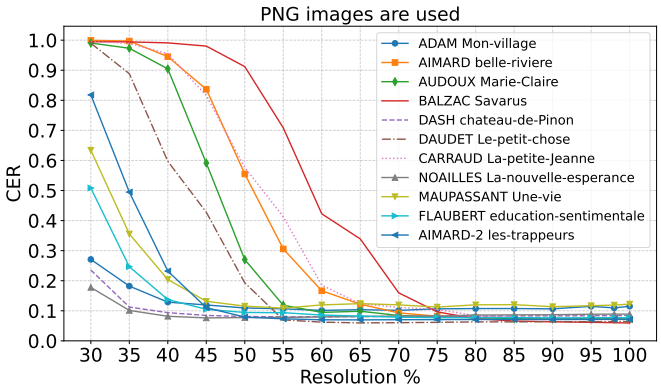
IV. What is the real influence of resolution on OCR quality ?

a) Resolution Thresholds and Accuracy Trends: The analysis of our results shows that the relationship between resolution and OCR accuracy is not completely linear. At the lowest resolution (circa 30% of the base image, equivalent to 21 dpi), OCR performance collapsed, with CER approaching 100% for about half of the documents. Accuracy, however, improved rapidly between 35% and 60% resolution (roughly 24–44 dpi), stabilizing beyond 55% (39 dpi). Above this threshold, gains in accuracy were minimal, suggesting that increasing resolution beyond 65% yields diminishing returns. We also found that document-specific factors—such as font uniformity, contrast, and paper texture—may mediate the effect of resolution.

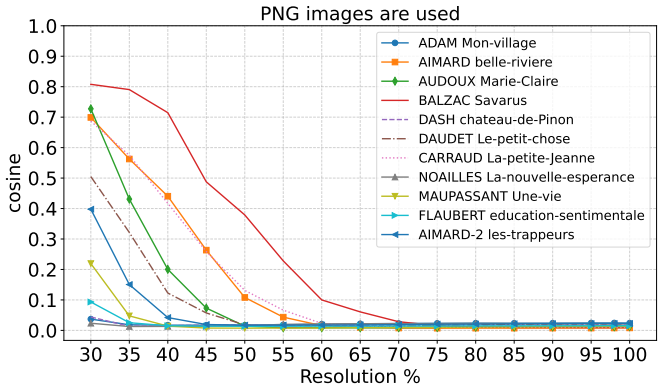
b) Influence of Compression Format: Image compression type had a smaller but measurable impact. Lossless formats such as PNG preserved finer details, resulting in slightly lower error rates at weak resolutions (below 50%) as illustrated in the figures 2a and 2b . However, at higher resolutions ($\geq$65%), performance across PNG, BMP, RAW, and JPEG converged. JPEG, despite being lossy, offered the best balance of speed and storage, with conversion times around 25% shorter and file sizes up to 50% smaller. This suggests that in large-scale digitization workflows, JPEG can be safely used without significant OCR degradation, provided that the resolution exceeds approximately 40% of original one.

c) Comparative Analysis with Ground Truth: When comparing OCR outputs at varying resolutions against the ground-truth ELTeC texts, three main performance patterns emerged from Figure 2a):

- Stable performance group: some documents maintained high accuracy (CER<0.1) even at 30–40% resolution (such as ADAM, DASH and NOAILLES .
- Progressive improvement group: other works displayed poor results below 45% but converged to A CER under 0.15 above 70% resolution (AIMARD 2, FLAUBERT and MAUPASSANT)
- Intermediate group: Most novels followed a sigmoid-shaped curve, showing rapid improvement between 40–60% before plateauing also (AIMARD, CARRAUD and DAUDET).



(a) Evaluation with Character Error Rate



(b) Evaluation with cosine distance on words

Fig. 2: Impact of PNG image quality on OCR transcription accuracy, compared to ground truth OCR transcriptions.

We have the same trends in Figure 2b, with a different shape in the curves since the cosine distance seems to underestimate OCR noise as compared to CER[13]. This variability highlights that OCR quality at low resolutions cannot be predicted solely by dpi but depends on document condition, typographic complexity, and image contrast.

d) Lexical Diversity and Noise Propagation: The lexical analysis provided a linguistic perspective on OCR degradation. As resolution decreased, OCR-generated texts exhibited inflated vocabulary sizes due to random character substitutions and word fragmentation, producing numerous hapax legomena (unique one-off tokens). The normalized TTR at 30% resolution was up to nine times higher than the ground truth, confirming severe lexical noise. At resolutions between 75% and 100%, normalized TTR values converged toward 1.0–1.5, indicating near-faithful reproduction of the true vocabulary. Examination of frequent word lists further revealed

that low-resolution transcriptions introduced non-linguistic tokens (“hi”, “ji”, “1”) and distorted common function words (“a”, “dé”,).These distortions appear to have limited impact on downstream NLP tasks such as NER or part-of-speech tagging, which primarily rely on broader lexical coherence. Nonetheless, when resolution exceeded 60%, OCR outputs were sufficiently clean for such tasks, demonstrating that usable linguistic data can still be extracted from sub-100 dpi sources.

V. Discussion and conclusion

The empirical findings challenge the universality of the 300 dpi standard. For the documents tested, a resolution as low as 39 dpi was sufficient to achieve accurate OCR transcriptions. Below that, performance degradation was exponential, while above 55–60%, additional resolution yielded marginal benefits relative to the exponential increase in file size and processing time. We believe these findings have significant implications for digital library infrastructure. If OCR engines can perform reliably at around 40 dpi, vast amounts of historical material previously deemed “too low-quality” may become accessible without the need for costly rescanning, which would support greener computing practices, as advocated by various information retrieval researchers, such as [19]. Similarly, institutions can reduce storage overhead and data transmission costs by retaining medium-resolution copies without sacrificing textual fidelity. From a computational standpoint, JPEG compression offers the most efficient balance between accuracy, file size, and processing time. Thus, it emerges as the most efficient option for large-scale digitization projects. Although lossless formats preserve marginally higher accuracy at extreme low resolutions, the trade-offs in storage and speed outweigh their benefits for most applications. This is especially relevant for online repositories, where bandwidth and load times are critical. Beyond the technical implications, the study also underscores a broader methodological point: OCR quality assessment should extend beyond pixel-level metrics to include linguistic indicators such as lexical richness and vocabulary stability. These measures not only capture character-level noise but also quantify how transcription errors distort linguistic structure, which is crucial for downstream text analysis in digital humanities and computational linguistics.

This study contributes to OCR and digital preservation research in two main ways. First, it provides a data-driven reassessment of the “300 dpi rule,” demonstrating that satisfactory OCR results can be achieved at substantially lower resolutions. Second, it offers a systematic framework combining quantitative accuracy metrics with lexical analysis to evaluate the real-world usability of OCR outputs. Key conclusions include:

1) Resolution Threshold: Reliable OCR transcription can be achieved at approximately 39 dpi for standard document dimensions (14×21 inches). Below this level, error rates rise sharply.
2) Compression Effects: JPEG compression has a negligible impact on OCR accuracy beyond 40% resolution but offers significant storage and time savings.
3) Noise Characteristics: OCR degradation inflates lexical diversity metrics; monitoring the Type-Token Ratio can serve as an indicator of transcription quality.
4) Downstream Usability: Texts produced from images above 60% of the original resolution (around 44 dpi) are suitable for linguistic and information retrieval tasks despite minor imperfections.

These findings open promising avenues for future work. Extending this analysis to a larger, multilingual corpus would allow for statistical validation across typographic styles, languages, and scripts (e.g., Gothic, cursive, Cyrillic, or non-Latin). Additionally, integrating OCR error modeling with post-OCR correction algorithms could further enhance text usability at ultra-low resolutions. Further research could also investigate how OCR errors specifically affect diacritics in French, and whether degradation patterns differ from general character loss. One way would be to examine whether the observed challenges stem from letter recognition in general or specifically from diacritics, and by assessing if the situation is comparable for low-resource languages [10] lacking reliable spelling-correction tools. Finally, the study advocates for greater transparency and reproducibility in OCR quality reporting within digital library infrastructures, ensuring that both technical and linguistic dimensions of OCR output are systematically evaluated.

Acknowledgment

This work received support from the French government, managed by the National Research Agency (ANR), under the France 2030 program, reference ANR-23-IACL-0007.

References

[1] Bazzo, G.T., Lorentz, G.A., Suarez Vargas, D., Moreira, V.P.: Assessing the impact of ocr errors in information retrieval. In: European Conference on Information Retrieval. pp. 102–109. Springer (2020)
[2] Blando, L.R., Kanai, J., Nartker, T.A.: Prediction of ocr accuracy using simple image features. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR95). pp. 319–322 (1995)
[3] Cannon, M., Hochberg, J., Kelly, P.: Quality assessment and restoration of typewritten document images. International Journal on Document Analysis and Recognition (IJDAR) **2**, 80–89 (1999). https://doi.org/10.1007/s100320050039
[4] Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: Paddleocr 3.0 technical report. arXiv preprint arXiv:2507.05595 (2025)
[5] Einsele, F., Hennebert, J., Ingold, R.: Towards identification of very low resolution, anti-alaised characters. In: 2007 9th International Symposium on Signal Processing and Its Applications. pp. 1–4 (2007). https://doi.org/10.1109/ISSPA.2007.4555324
[6] Gabay, S., Clérice, T., Reul, C.: OCR17: Ground Truth and Models for 17th c. French Prints (and hopefully more) (May 2020), https://hal.archives-ouvertes.fr/hal-02577236, working paper or preprint
[7] Gilbey, J.D., Schönlieb, C.: An end-to-end optical character recognition approach for ultra-low-resolution printed text images. CoRR **abs/2105.04515** (2021), https://arxiv.org/abs/2105.04515
[8] Habeeb, I., Azmi, S., Mohd Yusof, S.A., Ahmad, F.: Improving optical character recognition process for low resolution images. International Journal of Advancements in Computing Technology(IJACT) **6**, 13–21 (05 2014)

[9] Hamdi, A., Pontes, E.L., Sidere, N., Coustaty, M., Doucet, A.: In-depth analysis of the impact of OCR errors on named entity recognition and linking. Nat. Lang. Eng. **29**(2), 425–448 (2023). https://doi.org/10.1017/S1351324922000110, https://doi.org/10.1017/S1351324922000110

[10] Jayatilleke, N., de Silva, N.: Zero-shot ocr accuracy of low-resourced languages: A comparative analysis on sinhala and tamil (2025), https://arxiv.org/abs/2507.18264

[11] Jiang, M., Hu, Y., Worthey, G., Dubnicek, R., Underwood, T., Downie, J.: Impact of ocr quality on bert embeddings in the domain classification of book excerpts. CEUR Workshop Proceedings **2989**, 266–279 (2021), publisher Copyright: © 2021 Copyright for this paper by its authors.; 2021 Conference on Computational Humanities Research, CHR 2021 ; Conference date: 17-11-2021 Through 19-11-2021

[12] Kanungo, T., Marton, G.A., Bulbul, O.: Omnipage vs. sakhr: Paired model evaluation of two arabic ocr products. In: Proceedings of the SPIE Conference on Document Recognition and Retrieval VI. vol. 3651, pp. 109–120. SPIE (1999). https://doi.org/10.1117/12.347449, https://www.kanungo.com/pubs/spie99-ocreval.pdf

[13] Koudoro-Parfait, C.: Exploring 19th-Century Literary Space using AI : Evaluation and Analysis of Spatial Named Entity Recognition Tools (in French). Theses, Sorbonne Université (Jan 2025), https://theses.hal.science/tel-05042915

[14] Koudoro-Parfait, C., Hernandez, M., Lejeune, G., Dupont, Y.: Epiméthée : a workflow from OCR to spatial mapping. In: 19th International Conference on Document Analysis and Recognition (ICDAR). p. to appear. , Wuhan, China (2025)

[15] Koudoro-Parfait, C., Lejeune, G., Roe, G.: Spatial Named Entity Recognition in Literary Texts: What is the Influence of OCR Noise? In: Moncla, L., Brando, C., McDonough, K. (eds.) GeoHumanitiesSIGSPATIAL 2021: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities, Beijing, China, November 2 - 5, 2021. pp. 13–21. ACM (2021). https://doi.org/10.1145/3486187.3490206, https://doi.org/10.1145/3486187.3490206

[16] Lima, F., Silva, E.: Enhancing text recognition in ocr systems through image processing with bsrgan. In: Anais do XXI Simpósio Brasileiro de Sistemas de Informação. pp. 497–505. SBC, Porto Alegre, RS, Brasil (2025). https://doi.org/10.5753/sbsi.2025.246551, https://sol.sbc.org.br/index.php/sbsi/article/view/34366

[17] Pandey, R.K., Vignesh, K., Ramakrishnan, A.G., B, C.: Binary document image super resolution for improved readability and ocr performance (2018), https://arxiv.org/abs/1812.02475

[18] Rice, S.V., Jenkins, F.R., Nartker, T.A.: The fourth annual test of ocr accuracy. Tech. rep., Technical Report 95 (1995)

[19] Scells, H., Zhuang, S., Zuccon, G.: Reduce, reuse, recycle: Green information retrieval research. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2825–2837. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531766, https://doi.org/10.1145/3477495.3531766

[20] Schöch, C., Erjavec, T., Patras, R., Santos, D.: Creating the european literary text collection (eltec): Challenges and perspectives. Modern Languages (2021)

[21] Smith, R.: An overview of the Tesseract OCR engine. In: Ninth international conference on document analysis and recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007), https://dl.acm.org/doi/10.5555/1304596.1304846

[22] Tanguy, J.B.: Océriser pour accéder aux données ? Vers une évaluation non supervisée du bruit dans les données textuelles issues d'OCR de documents du XVIIème siècle. Theses, Sorbonne Université (Sep 2022), https://theses.hal.science/tel-04700035

[23] Volk, M., Furrer, L., Sennrich, R.: Strategies for reducing and correcting ocr errors. In: Sporleder, C., van den Bosch, A., Zervanou, K. (eds.) Language Technology for Cultural Heritage. pp. 3–22. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)

[24] Wang, S., Singh, M.K.: Systems and methods for optical character recognition for low-resolution documents (Apr 2018), https://patents.google.com/patent/US20180101726A1/en, library Catalog: Google Patents

[25] Wissenburg, A.: Digitising journals and the elib programme. In: Digitising Journals Conference on future strategies for European libraries. p. 43 (2000)