Benjamin Kiessling

Contact information	27 rue Auvry 93300 Aubervilliers France	*mobile:* +33 7 67 53 59 58 *email:* benjamin.kiessling@inria.fr
Personal information	Date of Birth Nationality	31/12/1991 French, German

Education and Degrees

April 2021	PhD, Computer Science, École Pratique des Hautes Études, Université PSL, *Advances in Optical Character Recognition for Historical Arabic Documents* Dissertation Committee: Marc Bui (Directeur de thèse) Gregory Crane (Rapporteur) Nachum Dershowitz (Rapporteur) Alicia Fornés Peter Stokes (Président) Daniel Stökl Ben Ezra
October 2012 - October 2016	M.Sc. Computer Science, University of Leipzig, *Median Strings for OCR Postcorrection*
October 2009 - February 2013	B.Sc. Computer Science, University of Leipzig

Research Grants & Project Leadership

BasHTR: Bootstrapping Arabic-Script Handwritten Text Recognition **2025 -**

Principal Investigator

- Secured €10,000 seed funding from the BnF Datalab to establish universal adapted transcription guidelines for Arabic-script ATR and create a foundational dataset for generalized recognition methods.

MiDRASH: Migrations of Textual and Scribal Traditions via Large-Scale Computational Analysis of Medieval Manuscripts in Hebrew Script (Start) **2023**

Proposal Co-Author

- Responsible for defining a multi-dimensional approach to improve the recognition of complex Hebrew manuscripts within the scientific proposal.

Scalable Methods for Text and Structure Recognition (OCR-D) **2017**

Co-Applicant & Co-Lead (Mitverantwortlicher)

- Secured €200,000 funding from the DFG for scalable full-text digitization of historical print (declined execution to join PSL).

Research Software and Corpora

Kraken ATR Engine

Role: Primary Author

- Conceived Kraken, a foundational ATR engine for historical scripts.
- Pioneered the use of trainable layout analysis and text recognition for low-resource material, establishing a new state-of-the-art baseline in digital humanities research.

Midrash Geniza Transcription Dataset

Role: Modelization

- Corpus of automatic and manual transcriptions of Cairo Geniza fragments.

CATMuS Medieval

Role: Architecture

- Ground-truth dataset for training HTR models for medieval manuscripts from the 8th to the 15th century.

BiblIA

Role: Modelization

- Ground-truth dataset for training of layout analysis and transcription models of Medieval Hebrew manuscripts.

BADAM: Baseline Detection in Arabic-script Manuscripts

Role: Principal Author

- Benchmark dataset for layout analysis and baseline detection in Arabic-script historical manuscripts.

OpenITI Classical Arabic Print Dataset

Role: Conception and Modelization

- Ground-truth dataset for classical Arabic printed texts to train OCR models for the Open Islamicate Texts Initiative.

Academic Affiliations and Work Experience

ALMAnaCH, Inria, Paris

Postdoctoral Researcher **July 2025 -**

- **BasHTR**: Leading guideline and dataset design for generalized historical Arabic-script text recognition.
- **BACK IN TIME**: Investigation into semi-open set recognition for the digitalization of ciphered manuscripts.
- **ATRIUM**: Design of object-detection based layout analysis for archeological field reports.

École Pratique des Hautes Études, Université PSL, Paris
UMR 8546 Archéologie et Philologie d'Orient et d'Occident

Research Engineer **July 2020 - June 2025**

- **MiDRASH**: Generalization and automatic selection of recognition models, complex layout analysis, and reading order.
- **Biblissima+**: Modelization of trainable reading order, Transformer-based document analysis, and unsupervised pretraining of OCR models for data-scarce settings.
- **ALTO**: Evolution of the standard for better handwriting, non-Latin text, and complex layout encoding.
- **DIM STCN**: Designed the high-performance computing strategy for the DIM STCN regional equipment grant, enabling large-scale training and inference.
- **RESILIENCE**: Conception of trainable document layout analysis methods for complex manuscripts.

Université PSL, Paris

Contractual Researcher **May 2018 - June 2020**

- **Tikkoun Sofrim** and **Sofer Mahir**: OCR for medieval Hebrew manuscripts with crowdsourcing integration for postcorrection.
- **eScripta** in PSL Scripta: Adaptation of an OCR engine for handwritten text recognition as part of the eScriptorium platform; preparation of open datasets for text recognition and layout analysis.

Saxon Academy of Sciences, Leipzig

Researcher **April 2018**

- **Bibliotheca Arabica**: Conception of semantically enriched text recognition methods for Arabic manuscript catalogues.

Faculty of Information Sciences and Knowledge Studies, University of Tehran (faculty sponsor: Prof. Fatima Fahimnia)

Visiting Researcher **December 2017 - February 2018**

- Development of a text line layout analysis method for a corpus of Iranian newspapers and periodicals.
- Survey on characteristics of historical Persian language manuscript page layout.

Alexander-von-Humboldt-Chair of Digital Humanities, Department of Computer Science, University of Leipzig

Researcher **February 2017 - March 2018**

- Conception of OCR and page layout analysis methods based on artificial neural networks.
- Evaluation of the developed methods in various historical non-Latin material and for style detection.
- Investigation of distributed training data creation using a crowdsourcing platform.
- **OpenITI/ShariaSource**: Machine-learning lead in a joint project for a distributed corpus-building tool.

Research Assistant **February 2014 - January 2017**

- Development of a distributed OCR pipeline
- Survey on dictionary-based OCR postcorrection for polytonic Greek texts.
- **EUDAT**: Lead of a pilot project on the long-term archival of corpus data.
- Training of neural networks for recognition of classical Arabic texts.

Lehrstuhl für Rechnernetze und verteilte Systeme, Department of Computer Science, University of Leipzig

Research Assistant **August 2011 - February 2014**

Academic Service

2025-2026	**ICDAR 2026 Competition on Multilingual Medieval Handwriting Recognition (CMMHWR '26)**, Principal Organizer, *20th International Conference on Document Analysis and Recognition (ICDAR)*, Vienna, Austria, 2026
2025	**TranscriboQuest 2025**, Advisor
2023-	**eScriptorium Steering Committee**, Member
2022	**DAHTR 2022**, Scientific Committee Member, École nationale des chartes, Paris

Teaching and Supervision

2020-2024	*Fouille de données et apprentissage (Data Mining and Machine Learning)*, 2nd year Master Digital Humanities, lectures and tutorials (12h, 24h since 2023), University of Tours
2020-2021	*Systèmes d'exploitation et Web (Operating Systems and Web)*, 1st year Master Digital Humanities, lectures, tutorials, and practical classes (12h), University of Tours
2017-2018	*Seminars on the principles of OCR*, M.Sc. Information Science (12h), University of Tehran

2021 Pierre Gaillard, *Evaluation of Augmentation Methods for Optical Character Recognition*, 2nd year Master Computer Science internship, 2 months

PhD thesis

Benjamin Kiessling. "Advances in Optical Character Recognition for Historical Arabic Documents (Avancées en Reconnaissance Optique des Caractères pour les Documents Arabes Historiques)". PhD thesis. École Pratique des Hautes Études, Université PSL, Paris, Apr. 2021. HAL: tel-03854403

Book Chapters

P.A. Stokes and **Benjamin Kiessling**. "Sharing Data for Handwritten Text Recognition". In: *Digital Humanities in Practice*. Routledge Companions to the Digital Humanities. Routledge, 2025, pp. 95–104. HAL: hal-04444641

Benjamin Kiessling. "HTR & OCR History". In: *Apprendre à lire aux machines*. Parcours Numériques. In Press. Presses Universitaires de Caen, 2025. HAL: hal-05163931

Journal Articles

Benjamin Kiessling. "Transcription Guidelines for Generalized Automatic Text Recognition". Submitted to *Computational Humanities Research* (Cambridge University Press). Preprint available on HAL. 2025. HAL: hal-05429033

Alan J. Wecker, Vered Raziel-Kretzmer, **Benjamin Kiessling**, Daniel Stökl Ben Ezra, Moshe Lavee, Tsvi Kuflik, Dror Elovits, Moshe Schorr, Uri Schor, and Pawel Jablonski. "Tikkoun Sofrim: Making Ancient Manuscripts Digitally Accessible: The Case of Midrash Tanhuma". In: *J. Comput. Cult. Herit.* 15.2 (Apr. 2022). ISSN: 1556-4673. HAL: hal-04938049

Benjamin Kiessling, Gennady Kurin, Matthew Thomas Miller, and Kader Smail. "Advances and Limitations in Open Source Arabic-Script OCR: A Case Study". In: *Digital Studies/Le champ numérique* 11.1 (2021). HAL: hal-04445437

P.A. Stokes, **Benjamin Kiessling**, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem. "The eScriptorium VRE for Manuscript Cultures". In: *Classics@* 18 (1 2021). ZENODO: 18202517

Benjamin Kiessling, Matthew Thomas Miller, G Maxim, Sarah Bowen Savant, et al. "Important New Developments in Arabographic Optical Character Recognition (OCR)". in: *Al-Uṣūr al-Wusṭā* 25 (2017), pp. 1–13. arXiv: 1703.09550

Conference Papers

Benjamin Kiessling. "Version 5 of the Kraken ATR Engine for the Humanities". In: *Document Analysis and Recognition – ICDAR 2025: 19th International Conference, Wuhan, China, September 16–21, 2025, Proceedings, Part III*. Wuhan, China: Springer-Verlag, 2025, pp. 443–458. ISBN: 978-3-032-04623-9. HAL: hal-05144723

Benjamin Kiessling and Thibault Clérice. "Does Context Matter ? Enhancing Handwritten Text Recognition with Metadata in Historical Manuscripts". In: *CHR2024 – Computational Humanities Research Conference*. Aahrus, Denmark, Dec. 2024. HAL: hal-04704547

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, Agnès Boutreux, Avery Manton, Simon Gabay, Patricia O'connor, Wouter Haverals, Mike Kestemont, Caroline Vandyck,

and **Benjamin Kiessling**. "CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond". In: *ICDAR 2024: International Conference on Document Analysis and Recognition*. Vol. 14806. Lecture Notes in Computer Science. Athens, Greece: Springer Nature Switzerland, 2024, pp. 174–194. HAL: hal-04453952

Gundram Leifert, Christel Annemieke Romein, Achim Rabus, Phillip Benjamin Ströbel, **Benjamin Kiessling**, and Tobias Hödel. *Evaluating State-of-the-Art Handwritten Text Recognition (HTR) Engines; with Large Language Models (LLMs) for Historical Document Digitisation*. Fourth Conference on Computational Humanities Research, Paris, France, December 6-8, 2023. Dec. 2023. ZENODO: 8102666

Benjamin Kiessling. "CurT: End-to-End Text Line Detection in Historical Documents with Transformers". In: *Frontiers in Handwriting Recognition: 18th International Conference, ICFHR 2022*. Vol. 13639. Lecture Notes in Computer Science. Hyderabad, India: Springer International Publishing, Dec. 2022, pp. 34–48. HAL: hal-04036249

Daniel Stökl Ben Ezra, Bronson Brown-DeVost, Pawel Jablonski, Hayim Lapin, **Benjamin Kiessling**, and Elena Lolli. "BiblIA - a General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset". In: *HIP@ICDAR 2021: The 6th International Workshop on Historical Document Imaging and Processing, Lausanne, Switzerland, September 5-6, 2021*. ACM, 2021, pp. 61–66. HAL: hal-04937186

Benjamin Kiessling. "A Modular Region and Text Line Layout Analysis System". In: *17th International Conference on Frontiers in Handwriting Recognition, ICFHR 2020, Dortmund, Germany, September 8-10, 2020*. IEEE, 2020, pp. 313–318. HAL: hal-04442992

Daniel Stökl Ben Ezra, Bronson Brown-DeVost, Nachum Dershowitz, Alexey Pechorin, and **Benjamin Kiessling**. "Transcription Alignment for Highly Fragmentary Historical Manuscripts: The Dead Sea Scrolls". In: *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. 2020, pp. 361–366

Benjamin Kiessling, Daniel Stökl Ben Ezra, and Matthew Thomas Miller. "BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts". In: *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2019, Sydney, NSW, Australia, September 20-21, 2019*. ACM, 2019, pp. 13–18. HAL: hal-02167164

Benjamin Kiessling, Daniel Stökl Ben Ezra, Rodney Ast, and Holger Essler. *Aligning Extant Transcriptions of Documentary and Literary Papyri with their Glyphs*. 29th International Congress of Papyrology (Lecce). 2019

Benjamin Kiessling, Robin Tissot, Daniel Stökl Ben Ezra, and Peter Anthony Stokes. "eScripta: A New Digital Platform for the Study of Historical Texts and Writing". In: *Proceedings of the DH* (July 2019). HAL: hal-02310781

Benjamin Kiessling. "Kraken - A Universal Text Recognizer for the Humanities". In: *Digital Humanities 2019*. Utrecht, Netherlands, July 2019. HAL: hal-04936936

Benjamin Kiessling, Robin Tissot, Peter A. Stokes, and Daniel Stökl Ben Ezra. "eScriptorium: An Open Source Platform for Historical Document Analysis". In: *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22-25, 2019*. IEEE, 2019, p. 19. HAL: hal-04937198

Benjamin Kiessling, Daniel Kinitz, Christoph Gümmer, and Parivash Mashhadi. "Script and Emphasis Detection using Recurrent Neural Networks". READ2018: International Interdisciplinary Symposium on Reading Experience and Analysis of Documents. 2018. HAL: hal-05439572

Oral Communications and Invited Talks

Benjamin Kiessling. "When Automatic Text Recognition doesn't work and how to fix it". presented at the Digital Humanities Tech Symposium at DH2025, Lisbon. July 2025. HAL: hal-05411082

Benjamin Kiessling. "Present and Future of ATR in the Humanities". presented at the AI Meets Humanities and Social Sciences, Austrian Academy of Sciences, Vienna. June 2025. HAL: hal-05411097

Benjamin Kiessling. "Large Multilingual ATR Models and Humanities Practice". presented at the 2025 Workshop SCOOP - Source Codes of the Past, Institute for Advanced Study, Princeton. June 2025. HAL: hal-05150070

Benjamin Kiessling. "Techniques d'analyse automatique des documents pour la paléographie". presented at the Ecole d'été en analyse de l'histoire de la mer et des océans (ANHIMO 2023), Institut de l'Océan de l'Alliance Sorbonne Université, Paris. June 2023. HAL: hal-05438654

Benjamin Kiessling. "Why We don't Postcorrect". presented at the International Workshop on Error Correcting HTR, Ca' Foscari University, Venice. Nov. 2022

Benjamin Kiessling. "New Developments in Kraken and eScriptorium". presented at the Documents anciens et reconnaissance automatique des écritures manuscrites workshop, Ecole Nationale des Chartes, Paris. June 2022. HAL: hal-05438436

Benjamin Kiessling. "eScriptorium : une plateforme d'applications web pour la transcription automatique de documents manuscrits ou imprimés". presented at the Atelier Digit-Hum, ENS, Paris. Oct. 2021

Benjamin Kiessling. "Kraken - A Universal Text Recognizer for the Humanities". presented at the Open Islamicate Texts Initiative Workshop, University of Maryland, College Park. Jan. 2020

Benjamin Kiessling. "The Limits to Digitization". presented at the Dark Archives conference, Faculty of English, Oxford University, Oxford. Sept. 2019. HAL: hal-05438427

Benjamin Kiessling. "The Intricacies of Arabic Manuscript Layout Analysis". presented at Digital Humanities Conference: Crowdsourcing and Citizen Science, University of Haifa, Haifa. May 2019

Benjamin Kiessling. "OCR-Processing in Leipzig's OpenPhilology Project". presented at 14th PhilTag meeting, KALLIMACHOS, University of Würzburg. 2017

Benjamin Kiessling. "The State of Arabic Script OCR". presented at meeting on the current state of Arabic OCR, Digital Library of the Eastern Mediterranean, Harvard University. 2017

Technical Workshops and Tutorials

2025	*Unpacking Large Language Models: Design, Limitations, and Solutions for Humanities Research*, Ars Inquirendi - The 2025 Dark Archives Conference, workshop (3h), University of Oxford
2023	*Atelier d'initiation à l'usage d'eScriptorium*, Journées annuelles du cluster 3 de l'EquipEx Biblissima+, tutorial (4h), Campus Condorcet
2022	*Intelligence artificielle pour les SHS : initiation à la transcription automatique des documents (Artificial Intelligence for the Humanities: Text Recognition Tutorial)*, PSL-week, lectures and tutorials (8h), EPHE, PSL University
2019	*The Kraken OCR Engine for Historical Documents*, ManuSciences '19, Franco-German summer school, workshop (10h), Fréjus
2017	*Kraken and the Future of Arabic OCR*, workshop (4h), Genealogies of Knowledge group, University of Manchester

Public engagement

	2021	*Paroles de chercheuses et chercheurs*, seminar, science communication for high school students, Lycée Gustave Monod, Enghien-les-Bains

Skills

Programming languages	C, Python, Perl, Bash, Tcl
Software	pytorch, tensorflow, lxml, numpy, scipy
Formats and Standards	METS, MODS, ALTO, hOCR, Page XML, MARCXML, TEI, Handle, Unicode, protobuf, JSON, Arrow

Languages

German	native speaker
English	bilingual
French	intermediate/upper-intermediate