Image as data: Artificial paleography and manuscript images as a source for digital humanists

Dominique Stutzmann

Machines are now reasonably well able to read ancient and medieval scripts, to reproduce the paleographical classifications, date and localize written documents, and perform writer identification [1]–[6]. They are expert systems performing some of the tasks of expertise in paleography defined as the field of history dealing with the script as an object (i.e. text as an image) and the use of the written word in past societies.

This new development changes the position of paleography within the Digital Humanities landscape, making the "text image" itself as a source for analysis, rather than human produced labels, manual measurements or scholarly metadata. After the term "digital paleography" was coined by Arianna Ciula in 2005 [7], a group of paleographers has worked in the field of "digital paleography" and cross-disciplinary studies for more than a decade [8]–[11]. In the past few years, as for the term "digital paleography", several debates have arose, partially linked to those on the notion and added value of "digital humanities"[12]. We recently proposed the term "artificial paleography" to describe the automated analysis of scripts and applying Computer Vision to "handwritten text images"[13].

This contribution proposes to give an overview of the methods applied by digital humanists and computer scientists on digital images of handwritten text and create new historical knowledge, by reading and scholarly editing texts, analyzing textual transmission, dating and localizing the (analogue) written artefacts, and better understanding the visual cultures and communication processes in past societies. This paper also addresses the new questions raised by "digitalization" and artificial intelligence in the field, be it of epistemological nature (quantitative/qualitative, sampling/exhaustiveness, objectivity and confidence measure), be it in human-machine interaction (black box and interpretation) and human resource and research processes.

[1] D. Stutzmann, "Clustering of medieval scripts through computer image analysis: towards an evaluation protocol," *Digital Medievalist*, vol. 10, 2015.

[2] V. Christlein *et al.*, “Automatic Writer Identification in Historical Documents: A Case Study,” *Zeitschrift für digitale Geisteswissenschaften*, vol. 2, 2016.
[3] F. Cloppet, V. Eglin, V. C. Kieu, D. Stutzmann, and N. Vincent, “ICFHR2016 Competition on the Classification of Medieval Handwritings in Latin Script,” *Proceedings of International Conference on Frontiers in Handwriting Recognition*, pp. 590–595, 2016.
[4] F. Cloppet, V. Eglin, M. Helias-Baron, V. C. Kieu, D. Stutzmann, and N. Vincent, “ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI),” in *14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017*, Kyoto, 2017, pp. 1371–1376.
[5] T. Bluche *et al.*, “Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project,” in *14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017*, Kyoto, 2017, pp. 312–317.
[6] J. Andreu Sánchez, V. Romero, A. H. Toselli, M. Villegas, and E. Vidal, “ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI),” in *14th IAPR International Conference on Document Analysis and Recognition. ICDAR 2017*, Kyoto, 2017, pp. 1383–1388.
[7] A. Ciula, “Digital palaeography: using the digital representation of medieval script to support palaeographic analysis,” *Digital Medievalist*, vol. 1, 2005.
[8] M. Rehbein, P. Sahle, and T. Schaßan, Eds., *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age*. Norderstedt: BoD, 2009.
[9] F. Fischer, C. Fritze, and B. Assmann, *Kodikologie und Paläographie im digitalen Zeitalter = Codicology and Palaeography in the Digital Age. 2*. Norderstedt: Books on Demand, 2010.
[10] O. Duntze, T. Schaßan, and G. Vogeler, Eds., *Kodikologie und Paläographie im digitalen Zeitalter 3 - Codicology and Palaeography in the Digital Age 3*. Norderstedt: Books on Demand, 2015.
[11] H. Busch, F. Fischer, and P. Sahle, *Kodikologie und Paläographie im digitalen Zeitalter 4 - Codicology and Palaeography in the Digital Age 4*. Norderstedt: BoD, 2017.
[12] A. Ciula, “Digital palaeography: What is digital about it?,” *Digital Scholarship Humanities*, vol. 32, no. suppl_2, pp. ii89-ii105, Dec. 2017.
[13] M. Kestemont, V. Christlein, and D. Stutzmann, “Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts,” *Speculum*, vol. 92, no. S1, pp. S86–S109, Oct. 2017.

CV

Dr. Dominique Stutzmann is senior researcher at the Institut de Recherche et d’Histoire des Textes (CNRS) and lecturer for medieval paleography and digital scholarly edition at the École Pratique des Hautes Études. He is a member of the Executive Board of Scriptorium and ICARUS. He has led or currently leads as Principal Investigator several research projects in the field of digital humanities : ANR Oriflamms and ECMEN on computer automated image analysis applied to palaeography and esp. to Vernacular palaeography ; the European research project HIMANIS

on text recognition and automated indexing of the complete collection of French medieval chancery registers from the 14th and the 15th c.; Saint-Bertin for virtually reconstructing of the former library of the Benedictine Saint-Bertin abbey in Northern France, and as co-PI the research project FAMA on Latin bestsellers.