In search of ground-truth for artificial palaeoghraphy

Hannah Busch, KNAW Huygens ING

Large scale digitization projects of the past twenty years and the possibility of exploitation with the help of the International Image Interoperability Network (IIIF) have substantially contributed to reach a critical mass which allows the application of deep learning for the study of medieval scripts. The project "Digital Forensics for Historical Documents. Script Analysis in a World of Anonymous Writers" (Huygens ING, KNAW Humanities Cluster)[1] attempts to create a digital tool, based on convolutional neural networks, in which the unique characteristics of a certain script sample will be matched with similar script samples by making use of digitized manuscript collections available in the world wide web and their IIIF APIs.

In the medieval period scripts developed over time, from the cursive scripts of the antiquity to Caroline minuscule, Gothic textualis, humanist and early modern scripts. The discipline of studying those chronological and regional variances of script is called palaeography, and it requires a deep and detailed knowledge of historical script features. The work of palaeographers is characterized by their high level of expertise and their skill of minute manual comparison with the scope to date and localize handwritings. Therefore, there are and there have been only a few authorities in the field. The decisive reason for the application of computational approaches is not the distrust in the opinion of the few experts, but the fact that due to mass digitization there is much material available then there are experts in the field of palaeography. Experimenting with new approaches becomes inevitable.

The work presented here is focusing on the challenge on defining and finding"ground-truth" for digital, computer-assisted, or artificial palaeography (see Ciula 2005 and 2017; Stokes 2009; Kestemont et al. 2017). In AI ground-truth data is highly relevant for the successful training of neural networks. I am going to present two possible approaches to label datasets for medieval palaeography. The first approach is concerned with the condition of the collection's metadata and the question how to deal with uncertainties. In theory, the metadata of the digital objects contains a valuable amount of reference data, but information about origin, dating, provenance, script classification are often not reliable. The reliability of the ground-truth data is an important factor for the palaeographical meaningfulness of the computational results. How can we overcome this objection and select suitable training data from the IIIF collections? In my presentation, a primary survey of the condition of digital manuscript collections will be discussed: What is the state of the art of metadata in digitized collections? Which standards are available and which are the most commonly used ones? How can we take advantage of the provided data regardless its (digital) condition and uncertainty for our research purposes?

The second approach to acquire ground-truth is the involvement of experts and their knowledge. Can I benefit from expert-sourcing to create training data that represents the (dis-)similarities of script in the Middle Ages without enlarging the black box of palaeography? In which ways can experts be involved in the data selection? And, how can experts be convinced in sharing their data with the project?

---

[1] Subproject 2: Script Analysis in a World of Anonymous Writers, Huygens ING, and DI Humanities Cluster, KNAW, funded by the KNAW 2018-2022.

References

Ciula, Arianna. “Digital palaeography: using the digital representation of medieval script to support palaeographic analysis”. *Digital Medievalist* vol. 1, no. 0, Apr. 2005, doi: http://doi.org/10.16995/dm.4

Ciulia, Arianna. “Digital palaeography: What is digital about it?” *Digital Scholarship in the Humanities,* vol. 32, no. suppl_2, 2017, pp. 89–105, doi:https://doi.org/10.1093/llc/fqx042

Kestemont, Mike, et al. “Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts“. *Speculum*, vol. 92, no. S1, Oct. 2017, pp. 86–109, doi:https://doi.org/10.1086/694112.

Stokes, Peter. “Computer-Aided Palaeography, Present and Future”. *Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age,* BoD, 2009, pp.309–38, http://kups.ub.uni-koeln.de/2978/.