Jean-Philippe Moreux

Job title: Expert

Organisation: National Library of France

Abstract:

Title: Hybrid Image Retrieval in Digital Libraries - A Large Scale Multi Collection Experimentation of Deep Learning techniques

While digital heritage libraries historically took advantage of OCR technology to index their printed collections and consequently improve the scope and performance of the information retrieval services offered to users, the access to iconographic resources has not progressed in the same way, and the latter remain in the shadows: manual incomplete and heterogeneous indexation, data silos by iconographic genre. Today, however, it would be possible to make better use of these resources, especially by exploiting the enormous volumes of OCR produced during the last two decades, and thus valorize these engravings, drawings, photographs, maps, etc. for their own value but also as an attractive entry point into the collections, supporting discovery and serenpidity from document to document and collection to collection. This article presents an ETL (extract-transform-load) approach to this need, that aims to: Identify and extract iconography wherever it may be found, in image collections but also in printed materials (dailies, magazines, monographies); Transform, harmonize and enrich the image descriptive metadata (in particular with deep learning classification tools); Load it all into a web app dedicated to image retrieval. The approach is pragmatically dual, since it involves leveraging existing digital resources and (virtually) on-the-shelf technologies.

Bio:
Jean-Philippe Moreux, graduated from INSA Toulouse (Computer Science, 1990) and CERAM-CNRS Nice-Sophia-Antipolis (Software Engineering mastère, 1991), is the OCR and digital publishing formats expert at the Bibliothèque nationale de France since 2012. He works on all the BnF heritage digitization and enrichment programs and participates in research projects on these topics, and the application of research results to digital libraries. His main research topics are OCR, heritage newspapers digitization and mediation, services for digital humanities, digital accessibility. He’s also a member of the ALTO Editorial Board. Prior to that, he was an IT R&D Engineer and project manager, and then worked as a science editor and a consultant in the publishing industry.