De l'image au texte - From image to text

Franck Lebourgeois*†[1]

[1]Laboratoire d'InfoRmatique en Images et Systèmes d'Information (LIRIS) – Institut National des Sciences Appliquées de Lyon, Université Lumière - Lyon II, Université Claude Bernard - Lyon I, CNRS : UMR5205, Ecole Centrale de Lyon – Blaise PASCAL 69621 VILLEURBANNE CEDEX, France

Résumé

La numérisation des documents historiques constitue une véritable révolution dans l'accès au patrimoine culturel. Cette révolution va changer radicalement la manière d'apprendre, et elle va avoir un impact majeur dans le travail des chercheurs en Sciences Humaines et Sociales. Cette révolution ne peut se faire que si les documents historiques de notre patrimoine culturel sont numérisés en mode image et convertis en documents éditables en mode texte. Sans transcriptions, tous les documents numérisés historiques resteront inaccessibles par les moteurs de recherche actuels. Dans cette présentation, nous décrirons toutes les étapes de ce long processus visant à transformer des images de documents numérisés en documents textuels. Nous discuterons des limites de la reconnaissance optique de caractères (OCR) et nous présenterons les études actuelles pour améliorer les résultats de l'OCR (prétraitement, post-traitement, la localisation automatique des erreurs ...) et alternatives (crowdsourcing, sous-traitance, ...). Nous présenterons également les plates-formes collaboratives pour l'enrichissement de documents historiques.

The digitization of historical documents constitutes a real revolution in the access to cultural heritage. This revolution will radically change the way we learn, and it will have a major impact in the researchers works in Humanities and Social Sciences. This revolution can be done only if the historical documents of our cultural heritage are digitized in image mode and converted into editable documents in text mode. Without transcriptions, all the digitized historical documents will remain inaccessible through current search engines. In this presentation, we will describe all the steps of the long process to transform images of digitized documents into textual document. We will discuss the limits of optical character recognition (OCR) and we will present the current research to improve OCR results (preprocessing, postprocessing, automatic localization of errors ...) and alternatives (crowdsourcing, outsourcing, ...). We will also present collaborative platforms for the enrichment of historical documents.

*Intervenant

†Auteur correspondant: Franck.Lebourgeois@insa-lyon.fr