

Impact : projet européen sur la reconnaissance optique de caractères : 2008 - 2011

Geneviève Cron¹
BnF, département de la Conservation, service Numérisation



Le projet européen [Impact](#) a permis à quinze grandes bibliothèques européennes de rencontrer les chercheurs et les industriels du domaine de l'OCR — Reconnaissance Optique de Caractères — ainsi que des instituts spécialisés dans le domaine des langues anciennes. Au total, 26 partenaires ont travaillé à faire évoluer la technologie de reconnaissance de caractères appliquée aux documents patrimoniaux.

Pour la France, étaient impliquées la BnF et l'[Atilf](#). La BnF a eu pour rôle initial de définir les attentes des utilisateurs bibliothécaires, de présenter des études de cas, de définir les métriques. Dans un second temps, elle a été sollicitée par les partenaires du projet pour la mise en place d'un serveur hébergeant des Web Services du projet et les a testés. Des corpus numérisés ainsi que de la « vérité terrain », c'est-à-dire un texte parfaitement transcrit, ont été fournis afin de permettre aux partenaires techniques d'optimiser les logiciels développés pour le projet.

Les outils réalisés vont du traitement d'image (binarisation, redressement, correction de courbure, OCR, retrait des bordures) aux outils contextuels (outils de correction collaborative, outils linguistiques, outils d'analyse d'un document numérique).

L'implication de la BnF sur ces thématiques se poursuit :

- en interne, au service numérisation du département de la Conservation :
 - depuis octobre 2010, le travail de recherche du doctorant Ahmed Ben Salah sur la caractérisation des documents numérisés et la prédiction de performance des techniques d'OCR a donné lieu à plusieurs publications sur l'état de l'art ;
 - un projet d'internalisation d'un OCR est actuellement à l'étude ;
- en externe, on dénombre :
 - au niveau national : les projets nationaux : [Quaero](#), [Digidoc](#), [Ozalid](#) ;
 - au niveau international : le projet [European Newspapers](#) ; développements au sein du [Centre de compétence Impact](#)

Ce projet qui a pris fin en 2012, a conduit à la création d'un Centre de compétence européen sur la numérisation et l'OCR, dont la BnF a été membre fondateur. Piloté par la [Fundación Biblioteca Virtual Miguel de Cervantes](#) (Madrid) de l'université d'Alicante (Espagne), le [Centre de compétence Impact](#) met à disposition des bibliothèques, les [outils](#) développés dans le cadre du projet : [une banque de connaissances](#) et [des services d'évaluation, de conseil et de formation](#).

¹ Geneviève Cron, ingénieure informatique, expert et chef des projets OCR au sein du département de la Conservation