
La Base de français médiéval et le consortium CAHIER : dix ans d'échanges et de collaborations

Alexei Lavrentiev^{*1} and Céline Guillot-Barbance²

¹Institut d'Histoire des Représentations et des Idées dans les Modernités – Centre National de la
Recherche Scientifique : UMR5317 – France

²Institut d'Histoire des Représentations et des Idées dans les Modernités – École Normale Supérieure -
Lyon – France

Résumé

Le projet de la Base de français médiéval (BFM, <http://txm.bfm-corpus.org>) fait partie des membres fondateurs du Consortium CAHIER. Les origines du projet remontent à la fin des années 1980 et son évolution a suivi, et parfois anticipé, les grandes tendances du développement des humanités numériques. L'expérience de la BFM a permis de contribuer à plusieurs chantiers du consortium CAHIER : l'accès libre aux données, les normes d'encodage (et notamment l'usage de la TEI), la typologie textuelle, la mise en place de chaînes éditoriales ouvertes. Les échanges qui se sont produits dans les groupes de travail et lors des ateliers du consortium permettent à leur tour d'améliorer les pratiques d'encodage et les outils proposés aux utilisateurs de la BFM et d'assurer une plus grande interopérabilité et pérennité des données.

La constitution de la BFM a commencé par la numérisation de l'édition de la *Queste del saint Graal* d'A. Pauphilet (1923) et la base a été enrichie au fil des ans grâce à des vacations, aux contributions de doctorants de Ch. Marchello-Nizia, aux échanges avec des collègues et, plus récemment, grâce à des financements ANR. A ce jour, la BFM comprend 170 textes composés entre le 9e et le 15e siècle, soit près de 4,7 millions de mots. Pour la *Queste del saint Graal* l'édition de Pauphilet a d'ailleurs été remplacée dans le corpus par une édition numérique originale (Marchello-Nizia et Lavrentiev 2019). Une augmentation importante du corpus est prévue en 2021. Les textes de la BFM sont étiquetés en morphosyntaxe et lemmatisés (avec ou sans vérification) et bénéficient du balisage XML-TEI enrichi. En particulier, le discours direct est balisé dans l'ensemble du corpus, ce qui permet de mener des recherches sur l'oral représenté (Guillot-Barbance et al. 2018). De nombreuses thèses et travaux de recherche ont été réalisés grâce aux données de la BFM. Notamment, la partie médiévale du corpus de la *Grande grammaire historique du français* (Marchello-Nizia et al. 2020) est entièrement issue de la Base de français médiéval. La BFM est accessible en ligne grâce au logiciel "portail TXM" (<http://textometrie.org>), les textes peuvent être consultés librement et l'accès au moteur de recherche et d'analyse est donné gratuitement sur simple inscription (Guillot-Barbance et al. 2017).

Quand la TGIR Corpus (prédécesseur d'Huma-num) lance en 2010 l'appel à la création de consortiums de corpus, la BFM faisait déjà, depuis 2004, partie du Consortium international pour les corpus de français médiéval (CCFM, <http://ccfm.ens-lyon.fr>). Cette organisation

*Intervenant

informelle, n'ayant jamais bénéficié d'un financement spécifique et dont l'activité s'est estompée après 2008, a néanmoins permis d'entamer la réflexion et de publier des documents de travail sur les normes communes d'encodage et de description des textes, ainsi que sur les conditions d'accès aux corpus et d'échange de données. L'expérience du CCFM a sans doute inspiré l'initiative de la TGIR Corpus et a servi de point de départ pour certaines activités de CAHIER (notamment pour l'organisation de groupes de travail et pour la rédaction de guides de bonnes pratiques).

L'un des premiers groupes de travail de CAHIER visait à traiter les questions juridiques liées aux droits d'auteurs (et d'éditeurs) et à la mise à disposition de corpus. La BFM avait commencé comme un ensemble de concordanciers échangés dans un cadre privé entre chercheurs sous la forme de CD-ROM ou de tirages papier, puis elle avait progressivement ouvert l'accès à l'interrogation et au téléchargement du corpus sur Internet. Elle avait une longue histoire de relations complexes avec les éditeurs commerciaux. Denise Pierrot, qui s'est occupée des questions juridiques pour la BFM, a également joué un rôle important dans le groupe de travail correspondant de CAHIER et a contribué à la rédaction du Guide des bonnes pratiques. La situation juridique a évolué suite aux décisions de justice dans le procès Droz contre Garnier numérique (2014 et 2017), ce qui a rendu possible la mise à disposition libre de textes historiques (hors appareil critique).

La pérennisation des données, grâce notamment à l'usage de l'encodage XML-TEI pour les textes et les annotations, a été la priorité pour la BFM depuis le début des années 2000 (Guillot et Heiden 2002). Le consortium CAHIER a pu bénéficier de la documentation de la BFM relative à l'encodage TEI du corps du texte et de l'entête (teiHeader) des documents. La BFM a été l'un des premiers projets à se conformer au modèle élaboré pour l'outil Weboai de CAHIER permettant le moissonnage des métadonnées (<http://weboai.cahier.huma-num.fr>). Toujours dans le domaine des métadonnées, les descripteurs typologiques de BFM, tels que le genre ou le domaine du texte, ont servi de base au thésaurus élaboré par le groupe de travail " Typologie textuelle " de CAHIER. Ce thésaurus, très riche et soigneusement structuré, permettra à son tour de préciser les métadonnées des futurs corpus de la BFM et de faciliter l'analyse de données de corpus agrégés à partir de plusieurs sources.

La pérennisation des données de recherche conformément aux principes FAIR (Findability, Accessibility, Interoperability and Reusability) est l'activité du Consortium CAHIER fortement encouragée par la TGIR Huma-Num. Même si la vision qui semble se dégager des recommandations du conseil scientifique d'Huma-Num et qui consiste à exiger avant tout le dépôt des données dans l'outil Nakala nous paraît un peu réductrice, il est certain que l'archivage pérenne et l'accessibilité des données sont extrêmement importants. Le soutien que CAHIER assure pour le dépôt des textes et des images dans NAKALA est très précieux pour la BFM.

La formation aux outils d'édition, d'analyse et de publication de corpus ouverts, ainsi que l'échange de bonnes pratiques éditoriales, a toujours été l'une des principales activités de CAHIER. La BFM, qui se développe en étroite collaboration avec la plateforme de préparation, d'analyse et de préparation de corpus TXM (Heiden et al. 2010) a pu partager son expérience, et des membres de l'équipe BFM ont animé de nombreuses séances de formation lors des ateliers CAHIER.

Le consortium CAHIER a également favorisé les échanges entre l'équipe TXM et le Pôle document numérique de la MSH de Caen qui développe la chaîne éditoriale Métopes (grâce notamment au financement d'un stage en 2017). La BFM bénéficie actuellement de certains éléments de Métopes pour la mise en page de ses éditions au format PDF et des scripts de traitement automatique permettant d'importer dans TXM des documents XML-TEI créés avec Métopes ont été élaborés.

La BFM est un projet qui a commencé bien avant la création du Consortium CAHIER et qui va sans doute continuer à se développer après la disparition de CAHIER dans sa forme actuelle de consortium de la TGIR Huma-Num. Quel que soit l'avenir du Consortium,

nous sommes convaincus que les ressources numériques, les méthodes et les outils de travail élaborés grâce ou avec le soutien de CAHIER ainsi que les relations humaines et les partenariats de recherche qui se sont tissés au cours des dix ans de ses activités continueront à jouer un rôle important dans la communauté des humanités numériques.

Références

Guillot-Barbance, Céline, Heiden, Serge et Lavrentiev, Alexei. 2017. " Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique ", *Diachroniques*, 7, 167-184.

Guillot-Barbance, Céline, Alexei Lavrentiev, Serge Heiden et Bénédicte Pincemin. 2018. " Diachronie de l'oral représenté: délimitation et segmentation interne du dialogue (IXe-XVe siècle) ", in Wendy Ayres-Bennett, Anne Carlier, Julie Glikman, Thomas Rainford, Gilles Siouffi et Carine Skupien Dekens (éds.) *Nouvelles voies d'accès au changement linguistique. Actes du colloque de la SIDF*, Paris, Classiques Garnier, 279-296.

Marchello-Nizia, Christiane, Combettes, Bernard, Prévost, Sophie et Scheer, Tobias (éds.). 2020. *Grande Grammaire Historique du Français*, Berlin, De Gruyter.

Heiden, Serge, Jean-Philippe Magué et Bénédicte Pincemin. 2010. " TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement ", in S. Bolasco, I. Chiari et L. Giuliano (éds.) *Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, Rome, Edizioni Universitarie di Lettere Economia Diritto, 1021-1032.

Heiden, Serge et Barbance-Guillot, Céline. 2003. " Capitalisation des savoirs par le web: une application de la TEI pour l'encodage et l'exploitation des textes de la Base de Français Médiéval ", in P. Kunstmann, F. Martineau, et D. Forget (éds.) *Ancien et moyen français sur le Web: enjeux méthodologiques et analyse du discours [Actes du colloque d'Ottawa, 4-5 oct. 2002]*, Ottawa, Éditions David, 77-92.

Marchello-Nizia Christiane et Lavrentiev, Alexei (éds.). 2019. *Queste del saint Graal*, Lyon, ENS de Lyon.

Pauphilet, Albert (éd). 1923. *La Queste del Saint Graal. Roman du XIIIe siècle*. Paris, Champion.

Mots-Clés: linguistique diachronique, français médiéval, textométrie, accès ouvert, typologie textuelle, TEI