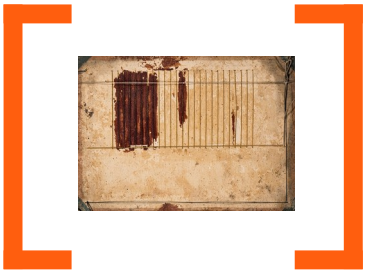
BULAC

[도서관] [शिक्षक] [ဘာသာစကား] [ሥልጣኔ]

Bibliothèque universitaire
des langues et civilisations



BILAN SCIENTIFIQUE DU PROJET MISTARA

Projets collaboratifs du GIS Collex-Persée, AAP 2018

Cécile Gobbo, Benjamin Guichard, Fanny Mion-Mouton, Maxime Tabet

Version : 30 juillet 2021

65 rue des Grands Moulins
F-75013 Paris
www.bulac.fr

T +33 (0)1 81 69 18 00
F +33 (0)1 81 69 18 99
contact@bulac.fr

TABLE DES MATIÈRES

1. RAPPEL DES OBJECTIFS DU PROJET ET DE SON PÉRIMÈTRE

1.1. RÉSUMÉ DU PROJET : ONOMASTIQUE ARABE ET MÉTADONNÉES DES LANGUES À ÉCRITURE ARABE DANS LES RÉFÉRENTIELS EN LIGNE

Les métadonnées décrivant des ressources en alphabet arabe (langues arabe, persane, ourdou et turque-ottomane principalement) présentent de nombreuses anomalies liées à la complexité du système onomastique, aux divergences phonétiques ou orthographiques entre langues ou à des erreurs de codage des caractères.

Le projet Mistara explore les possibilités d'alignement et d'enrichissement des informations relatives aux entités du monde arabo-musulman dans les outils catalographiques. Il croise les enjeux de la transition bibliographique et de la recherche sur l'onomastique arabo-musulmane en visant à :

- évaluer les adaptations nécessaire à la modélisation et à la manipulation des formats d'entités personne à la gestion d'entités multi-lingues et multi-écritures et aux spécificités de la décomposition du nom dans le monde arabo-musulman

- améliorer la gestion des référentiels produits par la recherche sur les questions onomastiques du monde arabo-musulman pour assurer leur interopérabilité et permettre l'enrichissement des outils catalographiques.

La recherche française sur le monde arabo-musulman s'est distinguée par une attention importante portée aux questions onomastiques. Cet apport scientifique est susceptible d'enrichir la modélisation des données élaborée dans le cadre de la transition bibliographique et d'évaluer les possibilités offertes par les outils d'analyse et d'alignement automatique des données bibliographiques à traiter ces informations. Le projet vise ainsi à enrichir des référentiels catalographiques (Idref au premier chef) à partir de référentiels tiers issus de la recherche : la base AlKindi (IDEO) et l'*Onomasticon Arabicum* (IRHT).

1.2. CALENDRIER

Le calendrier initial du projet était établi de juin 2019 à décembre 2020. Le lancement effectif a eu lui en juillet 2019, le recrutement prévu à début en novembre 2019. Compte tenu des difficultés de mise en œuvre dans le contexte sanitaire de l'année 2020, le GIS Collex-Persée a autorisé la prolongation du projet jusqu'au 30 juin 2021.

1.3. PARTENAIRES

Établissement porteur : GIP BULAC (Bibliothèque universitaire des langues civilisations), établissement porteur.

Établissements documentaires et équipes de recherche : Aix-Marseille université (SCD), Institut de recherche sur l'histoire des textes (section arabe, CNRS), Institut dominicain d'études orientales (Le Caire), Institut du monde arabe, Maison Méditerranée des sciences de l'Homme.

Agences bibliographiques : Bibliothèque nationale de France, Agence bibiographique de l'enseignement supérieur.

1.4. LIVRABLES ENVISAGÉS ET OBJECTIFS INITIAUX

Le projet Mistara a été conçu après le constat d'un défaut d'homogénéité dans les méthodes de traitement catalographique des données en écriture non-latines, et en l'occurrence, en écriture arabe. Mistara a été financé dans le cadre d'un appel à projet collex-persée lancé fin 2018. En avril 2019, la Bulac présente Mistara aux journées Collex-Persée en dressant un bilan de qualité des données et une liste d'outils potentiellement utiles pour l'alignement et l'enrichissement de ces données. Dans ce cadre, diverses institutions spécialisés dans le

domaine arabe ont été sollicitées en tant que partenaires de ce projet afin d'apporter leur expertise et partager leurs données (Institut du monde arabe, SCD Aix-Marseille Université, Maison Méditerranéenne des sciences de l'homme, section arabe de l'IRHT et Institut dominicain d'études orientales).

D'un point de vue plus large, le projet Mistara est né dans un contexte d'homogénéisation générale et internationale des modèles de structure des données. C'est dans ce contexte que le programme de transition bibliographique est lancé en France en 2015, afin d'adapter les catalogues français à ces modèles. Ce programme a engendré, en 2017, le projet de creation d'un fichier national d'entités visant à mutualiser la production et la gestion des données sur les différentes entités. Parmi ces données se trouvent les données en bi-écriture arabe dont une grande partie doit être enrichie et corrigée pour ne pas être exclue de cette grande transition.

Il y avait initialement 6 objectifs principaux au projet :

1. Rédaction d'un guide des bonnes pratiques pour la création de notices d'autorité relatives aux personnes physiques des aires culturelles utilisant l'écriture arabe, s'inscrivant dans le cadre de RDA-FR.
2. Preuve de concept de l'utilisation des outils d'alignement semi-automatisés pour le traitement de données d'autorité décrivant des personnes physiques, d'une part, et des données multilingues et multi-écritures d'autre part.
3. Proposition de préconisations pour l'alignement et l'enrichissement de notices d'autorités ayant des spécificités linguistiques à partir des outils des agences bibliographique nationales et en utilisant des référentiels tiers.
4. Preuve de concept pour l'alignement du catalogue AlKindi de l'IDEO sur le référentiel ISNI grâce à l'outil Bibliostratus.
5. Recommandation et spécifications pour l'exposition en données ouvertes liées de l'Onomasticon Arabicum de l'IRHT et son alignement sur les référentiels nationaux .
6. Mise à disposition d'un corpus de métadonnées enrichies pour la publication d'un corpus de recherche ou d'une bibliothèque numérique (corrections, dédoublonnage et enrichissement d'un corpus de notices d'autorités relatives aux personnes physiques des aires culturelles utilisant l'écriture arabe dans IDREF).

2. RÉSULTATS DU PROJET

2.1. LA PRÉPARATION DU CORPUS DE TRAVAIL ET SON TRAITEMENT

La définition du corpus s'est appuyée sur des extractions par requêtes Sparql de la base Data Idref ciblant les auteurs associés aux imprimés arabes du XIX^{e} siècle et associant différents critères de dates et de langues.

Les premiers constats sur les données et l'efficacité des outils ont dégagé deux principales limites à une homogénéisation automatique. Elles s'expliquant soit par des limites des outils ou des défaut de qualité des données utilisées. L'état des données a fait environ 3 500 notices d'autorité à enrichir et nettoyer. Un des objectifs du projet est d'évaluer les possibilités de traitement automatisées des corrections nécessaires. Un corpus de 500 notices d'autorité arabo-musulmanes, parmi les plus défectueuses, a été isolé en privilégiant les auteurs représentés dans les corpus d'imprimés du Caire, bien représentés chez les différents partenaires du projet. La majorité de ces autorités sont des philologues, des exégètes, des historiens et des théologiens. Il y a une large majorité d'auteurs arabes suivis par les auteurs persans et turc ottomans mais aussi ouzbeks. La majorité de ces entités se retrouvent dans les bases catalographiques des différents partenaires.

L'onomastique arabo-musulmane présente des complexités, d'autant plus manifestes pour les personnes ayant vécu avant le XX^{e} siècle. La structure du nom arabo-musulman médiéval correspond mal à la structure des notices qui prévoit une entrée en deux éléments représentants le nom de famille et le prénom. Le nom propre arabo-musulman contient, en effet, cinq éléments qui peuvent se répéter pour former parfois de longues chaînes onomastiques. Les divergences dans les pratiques de translittération de l'abjad arabe, entre normes de catalogage des pays anglophones, arabophone ou en Europe continentale, ajoutent à la difficulté.

Le défi du projet a été de trouver une réponse consensuelle à cette problématique. Il s'agissait de proposer une méthode qui n'élimine pas les différentes pratiques et la mention du maximum d'éléments mais les réordonne avec un ordre de préférence, l'objectif étant de produire des données complètes et clairement identifiables par les machines. L'enjeu était donc d'arbitrer des choix onomastiques susceptibles de satisfaire au mieux la lisibilité par la machine et la clarté scientifique des données.

Les règles de base de l'onomastique et de l'écriture arabe sont reprises par différentes civilisations musulmanes non locutrices de l'arabe. Ainsi, il a été nécessaire de prendre en compte les multiples spécificités culturelles et graphiques appartenant à ces langues qui influent sur l'onomastique. L'emprunt de l'abjad arabe dans des langues non-sémitiques engendre des adaptations graphiques correspondant aux traits phonétiques que ces langues n'ont pas en commun avec l'arabe. Ces spécificités ne sont pas toujours visibles à l'œil nu mais bien traitées par les machines. Les code d'écriture Unicode et leurs différents mode de translittération doivent impérativement figurer dans les données pour prévenir la production de doublons d'autorité.

2.2. RETOUR D'EXPÉRIENCE SUR LES OUTILS D'ALIGNEMENT DE LA TRANSITION BIBLIOGRAPHIQUE APPLIQUÉS À DES DONNÉES D'ENTITÉS PHYSIQUES

Les données des différents partenaires ont permis d'enrichir les données du Sudoc et de tester leur traitement avec différents outils d'alignement ou de correction des liens entre notices : Bibliostratus, Openrefine, Qualinka/Paprika.

Sur Bibliostratus, le taux d'alignement sur des identifiants ISNI ou ark produits par la BNF s'est révélé extrêmement faible, entre 7 et 10 % . Les capacités d'OpenRefine pour identifier des doublons se sont également révélées peu limitées : déclaration de nombreux faux doublons, doublons non identifiés. Qualinka/Paprika, utiliser pour créer des liens entre les notices d'autorité traitées et leurs différentes notices bibliographiques, s'est révélé peu performant pour les traitement des chaînes de caractères arabes non translittérés. Les résultats doivent ainsi être systématiquement complétés par un traitement manuel. Des données plus homogènes permettent a certains outils de donner de meilleurs résultats.

Les défauts d'analyse des chaînes de caractères arabes par ces différents outils explique que le nettoyage et l'enrichissement manuel des données ne suffit pas toujours. La translittération ISO mobilise l'emploi de nombreux signes diacritiques qui sont également susceptibles de limiter la performance de ces outils. Le projet a ainsi été l'occasion de travailler avec l'ABES à améliorer les performances des outils d'indexation de la base IDREF, en enrichissant les tables de correspondances entre caractères simples et caractères diacrités utilisés pour la translittération ISO de l'abjad.

Un rapport détaillé sur le bilan d'usage de ces outils pour des données en écriture arabe et des données décrivant des entités physiques a été élaboré avec l'ABES ; il est destiné à être partagé au sein du groupe Transition bibliographique. Le travail mené ouvre la voie d'une utilisation du logiciel Bibliostratus pour détecter des lien manquants entre notices bibliographiques et entités personnes.

2.3. LE GUIDE COMPLÉMENTAIRE POUR LE CATALOGAGE DES AUTORITÉS ARABO-MUSULMANES

Un guide complémentaire pour le catalogage des autorités arabo-musulmanes a été rédigé pour préciser et homogénéiser la prise en compte de la multitude de pratiques d'écriture et la richesse onomastique du monde islamique. Ce livrable est composé d'une première partie sur l'onomastique arabo-musulmane du point de vue du catalogage, une deuxième partie sur l'application de ces principes avec des exemples de cas particuliers et enfin une troisième partie sur les questions de translittération.

Ce guide vise à rassembler, mettre à jour, compléter et mettre en relation les informations parsemées dans les différentes documentations sur le catalogage des autorités arabo-musulmanes et les sources scientifiques sur l'onomastique. Il fixe notamment des contraintes pour la saisie des points d'accès autorisés afin d'assurer une homogénéité des données là ou nous avions plusieurs méthodes utilisées selon les équipes de catalogueurs. Il contient également divers conseils sur le contenu de chaque champ des notices. Les propositions qui figurent dans ce guide s'appuient sur les méthodes employées sur les bases de données et dans les dictionnaires biographiques de référence. L'objectif recherché n' pas été d'imposer un modèle unique mais de proposer une structure qui intègre différents modèles dans une structure homogène. La démarche a été similaire pour les questions de translittération qui constituent la troisième partie de ce guide complémentaire. Cette partie rappelle les règles existantes et précise la place de chaque système de translittération dans les notices. Pour indication, il existe officiellement 4 systèmes de translittération pour l'arabe. Il s'agit d'indiquer dans quelles conditions et dans quels champs nous devons retrouver ces systèmes dans les notices.

Validé par l'ABES, ce guide complémentaire sera mis à disposition du réseau SUDOC à l'automne 2021 et fera l'objet d'une communication spécifique auprès des membres du réseau.

En complément, un enrichissement et une correction des exemples de noms arabes figurant dans le guide RDA-FR a été proposé aux agences bibliographiques.

2.4. MISE À DISPOSITION D'UN CORPUS ENRICHI

L'enrichissement manuel, par le chargé de mission Mistara à la BULAC et les équipes de la bibliothèque de l'Institut du monde arabe, de 311 notices indexées a été réalisé en respectant cette structure. Cet enrichissement inclus des corrections partielles sur environ 1 000 notices bibliographiques. Ce corpus représente une partie infime du travail à réaliser sur les données arabo-musulmanes. Les notices modifiées sont signalées dans le référentiel IdRef par une étiquette de projet les rattachant au chantier Mistara.

2.5. AUTRES LIVRABLES

Le carnet de recherche https://mistara.hypotheses.org/ a été créé en parallèle des travaux techniques. Il contient des articles introductifs portant sur les différents aspects scientifiques du projet. Il présente notamment le guide raisonné des ressources utiles pour l'étude de l'onomastique arabe et l'identification des auteurs anciens livré par le projet[1].

Le livrable sur le schéma d'échange des données entre l'Abes et l'Idéo a rapidement abouti au constat de la possibilité d'exprimer en Marc la richesse de la conceptualisation onomastique des notices d'entité personnes créées par l'Idéo. Une démarche de conventionnement a été entamée pour permettre à l'Idéo de devenir directement producteur dans Idref de ces données ; la reprise des notices existantes pour versement dans Idref interviendra dans un second temps. Il n'a en revanche pas été possible d'intégrer la modélisation des données de l'*Onomasticon Arabicum* dans le cadre du projet.

1. https://mistara.hypotheses.org/193

3. PERSPECTIVES DE COURT ET MOYEN TERME

Le travail réalisé dans le cadre du projet Mistara sera prolonger à court ou moyen terme :

- au sein du réseau de l'ABES : amélioration de l'indexation des caractères diacrités de la translittération ISO et des caractères orignaux arabes des différentes langues ayant recours à l'abjad ; diffusion d'un guide de catalogage complémentaire pour les données onomastiques du monde islamique.

- Au sein du groupe Transition bibliographique : corrections et enrichissement des exemples arabes dans les guides RDA-FR, identification de scénarios d'usage de Bibliostratus pour l'analyse d'entités personnes et la correction de liens entre notices d'entités et notices bibligraphiques.

- Au sein de la MMSH, utilisation des modalités d'alignement et de dédoublonnement à l'aide de Bibliostratus pour poursuivre la reprise de la base d'autorités mono-écriture arabe, versée dans la base d'appui du SUDOC, pour l'alignement sur des notices validées IDREF

- Au sein de l'IMA, reprise de données de la base locale ALAM pour enrichir les notices validée d'IdRef car le projet a pu confirmer la validité et l'intérêt des champs et sources d'information contenus dans la base ; l'option d'un versement de ces sources dans la base d'appui du SUDOC est à étudier.

- Conventionnement en cours entre l'ABES et l'IDEO visant à :

    - l'interfaçage d'IDREF et des différentes bases catalographiques de l'environnement ILS-DIAMOND (couvrant des catalogues relatifs auchristianisme latin, christianismes d'orient, civilisation islamique classique) ;

    - un travail en production directement dans IdRef pour l'enrichissement des bases de l'ILS-Diamond ;

    - la reprise dans IdRef des données et notices produites dans l'ILS-Diamond.

- Au sein de la BULAC, pérennisation d'une fonction de traitement et d'alignement de référentiels onomastiques issus de la recherche en études aréales sur IdRef et pour l'enrichissement du fichier national d'entité.

- Sensibilisation des groupements d'intérêt scientifique aréalistes (Asie, Études africaines, Moyen-Orient et monde musulman) aux enjeux de l'alignement des référentiels issus de la rechercher sur IdRef : intégration d'un groupe de travail sur les référentiels et l'enrichissement de données en écriture non latin dans le projet de consorium Huma-Num DISTAAM (Digital studies Africa Asia Middle-East) ; le travail de traitement de l'*Onomasticum arabicum* sera envisagé dans ce cadre

Lorsque la correction des données aura atteint un taux conséquent les outils de traitement automatique fonctionneront mieux. Il est en effet paradoxal de devoir traiter manuellement pour faire fonctionner un outil automatique, mais les spécificités de traitement des écritures ne permettent pas d'envisager un autre chemin. Avec des données plus homogènes, les outils de détection automatique de doublons produiront moins de bruit et trouveront moins de faux doublons. Les outils d'alignement donneront des taux plus importants de liens trouvés. Nous

pouvons donc conclure que ce projet a pris un rôle préparatif plus important que prévu. Seule cette préparation des données leur permettra d'être pleinement intégrées dans les projets de transition numérique.