David Hamidovic, “Encore une base de données sur des manuscrits ! Approche pragmatique des manuscrits de la mer Morte”

David Hamidovic (UNIL FTSR) nous présente son projet *Sokoka* récemment financé pour une période de 3 ans. Original dans son ambition et sa démarche, ce projet s’inscrit sur une ligne déjà instituée sur laquelle il sera nécessaire de revenir : que sont les manuscrits de la mer Morte et quel projets digitaux ont-ils déjà inspirés ?

Les manuscrits de la mer Morte

Les manuscrits de la mer Morte sont un corpus d’environ 950 manuscrits découverts dans les années 1950 dans la région de Qumrân (entre Jérusalem et Jéricho) et vraisemblablement écrits entre la fin du III^{e} siècle avant J.-C. et le milieu du I^{er} siècle après J.-C. S’il est encore trop tôt pour mesurer tout l’apport du contenu de ces manuscrits pour nos connaissances sur les évolutions spirituelles de cette période charnière (transformation du judaïsme, christianisme primitif), passablement des études archéologiques et historiques ont déjà permis de dégager des éléments importants quant à leurs contextes de production.

Une heureuse découverte, de moins en moins mystérieuse

Ce sont des pâtres bédouins qui ont les premiers découvert ces manuscrits – pour certains conservés dans des jarres – dans des grottes de la région de Qumrân. Il est fort probable qu’ils aient été utilisés comme combustibles pour les froides nuits désertiques avant qu’une grande expédition archéologique menée par le gouvernement jordanien et l'école biblique française ne soit entreprise entre 1947 et 1956. Les pièces manuscrites excavées sont de qualités variables et vont de rouleaux très bien conservés à de minuscules lambeaux rongés par la moisissure.

D’où viennent-ils ? Et que pouvaient-ils bien faire là, pour certains dans leurs jarres, cachés dans ces grottes difficiles d’accès au beau milieu d’une région désertique ? Il faut commencer par rappeler que la région de Qumrân a subi passablement de changements politico-climatiques depuis l’époque des manuscrits. Jérusalem et Jéricho étant des centres culturels et économiques importants, il est fort probable que la région de Qumrân était un lieu de circulation, notamment

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL

marchande. Selon Pline l'Ancien, Qumrân pouvait même être considéré comme une région à plantation plutôt riche, productrice notamment de parfums et d'encens. Loin de son image actuelle plutôt austère, Qumrân était vraisemblablement irrigué d'activités humaines, potentiellement même de type scripturale.
À partir de là, certains de ces manuscrits énigmatiques auraient-ils pu être produits à même le site de Qumrân ? Sur le site de Qumrân, non loin des grottes abritant les manuscrits, un groupe d'archéologues découvrit dans les années 1950 une jarre du même type que les jarres trouvées dans les grottes. Si aucun manuscrit n'a été retrouvé à même la ruine, les recherches ont révélé une architecture très spéciale, presque monastique (ou plutôt *proto-monastique* car la forme « monastère » n'apparaîtra que vers le IV^{e} siècle en Egypte) : de très longues pièces indiquant un type d'habitation collectif, des pièces sans porte (on devait sûrement y accéder par le toit), des entrepôts, etc. Il est ainsi tout à fait possible qu'au moins une partie des manuscrits de la mer Morte aient été produits dans ce proto-monastère potentiellement en rupture avec les prêtres dirigeants du temple de Jérusalem. Mieux, il est également possible que leur mise en cachette dans des grottes ait été le fruit d'une décision actée en ces lieux. Mais suite à quel type d'événement ? Si des lignes s'éclaircissent, beaucoup d'éléments restent à documenter.

Fastidieux déchiffrement

Les manuscrits sont dans un état fragmentaire, abimés par le temps et les champignons. Si l'exploration des grottes et les processus d'excavation furent assez rapides (entre 1947 et 1956), le déroulement et surtout le déchiffrement des textes furent laborieux et nécessitèrent l'expertise d'une équipe internationale et pluridisciplinaire de chercheurs pas toujours facile à coordonner. De nombreux romans de gare (*DaVinci Code* en tête) habillés de fantasques complots religieux ont par ailleurs contribué à leur mise en visibilité.
La grande majorité des manuscrits sont écrits en hébreu classique, langue principale d'écriture de cette période et région du monde. Environ un tiers de cette première « classe » de manuscrits sont des copies plus ou moins rigoureuses de passages de la Bible hébraïque ou Ancien Testament appartenant au canon hébraïque. Un deuxième tiers contient des textes sacrés faisant alors autorité mais n'ayant pas été retenus par la suite au sein de ce canon « officiel ». Pour un dernier tiers, le contenu des manuscrits est de type « sectaire » et relate pour beaucoup les règles de vie communautaire des résidents de Qumrân.
Une deuxième « classe » de manuscrits contient une centaine de textes en araméen, langue populaire d'expression du Moyen-Orient très utilisée du temps de Jésus. Leur contenu est principalement des histoires bibliques détournées, voire parodiées. Le message théologique apparaît comme secondaire, l'emphase étant mise sur les situations plutôt cocasses des personnages.
En plus d'une trentaine de manuscrits écrits en grec ancien sur des supports papyrus (et non pas parchemins), une quatrième « classe » de manuscrits écrits en caractères cryptiques à la fois grecque, hébraïque et même paléo-hébraïque complète ce tableau. Ces textes ont-ils été intentionnellement cryptés ? Si oui, pourquoi ? Et pour qui ? Ces questions ont trouvés des éléments de réponse dans les années 2000 lorsqu'a été retrouvée à Jérusalem une très vieille tasse ayant appartenue à une lignée de prêtre du temple de Jérusalem et sur laquelle étaient inscrites des écritures cryptiques en tout point similaires à celles retrouvés dans ces quelques manuscrits de Qumrân. Grande découverte qui avalise une hypothèse faite depuis maintenant plusieurs décennies : il est tout à fait possible que cette quatrième « classe » de manuscrits aient

été écrits par des prêtres du temple de Jérusalem potentiellement en rupture théologique avec les prêtres dirigeants. Dans leur fuite/exil du temple de Jérusalem, cette famille de prêtres en rupture aurait alors pu emporter ces textes jusqu'au proto-monastère de Qumrân. Mais à quelle fin véritable ? Et sur quels points de désaccord précis ? Beaucoup de détails restent encore incertains.

Les manuscrits de la mer Morte et les DH

Les manuscrits ont été rattachés à des projets digitalisation, de mise en base de données et de traitements automatisés (ce qui sera plus tard rassemblé sous le terme « DH ») dès les années 1960. Schématiquement, 3 types de technologies ont été développés plus ou moins spécifiquement pour l'exploration de ces contenu numérisés.

Logiciels d'aide à la lemmatisation

Le premier type de technologies regroupe les supports d'aide à la lemmatisation des textes (découpage en fonction des préfixes et suffixes), à leur vocalisation et autres dictionnaires intégrés. Ces applications – qui sont en fait très proches de logiciels comme *BibleWorks* – donnent souvent accès aux images numérisées des manuscrits en plus ou moins haute définition dans une optique d'exégèse. Google a par ailleurs récemment lancé un projet de re-numérisation en ultra-haute définition en collaboration avec les autorités israéliennes. Elément intéressant, ce projet qui se veut une exploration au plus près des manuscrits de la mer Morte se heurte aux transformations subies par les matériaux originaux : depuis les années 1950, les manuscrits se sont dégradés et surtout rétractés, faisant parfois disparaître plusieurs centimètres en bordure de page, ce qu'une digitalisation en ultra-haute définition ne peut à elle seule pas faire réapparaître.

Luminescence infrarouge

Un deuxième type de technologies provient des techniques de la luminescence infrarouge, développées depuis les années 1960. Passablement de manuscrits ayant été retrouvés dans un état les rendant indéchiffrables, des techniques de scannage infrarouge et de traitement informatique ad hoc sont parvenus à rendre visibles des signaux auparavant inaccessibles. Les résultats de ces procédés ont parfois été intégrés à des bases de données d'exégèse et leurs fonctionnalités de lemmatisation. Par la suite, le procédé a été miniaturisé et son traitement accéléré au point d'être insérés dans des supports portatifs capables de dialoguer avec des micro-ordinateurs.

Traitements du signal pixellique

Un troisième type de technologies traite des signaux numériques propres aux caractères manuscrits des textes. Dans le cadre de sa thèse de doctorat, David Hamidovic a lui-même mis au point une méthode d'apographie afin de proposer des hypothèses de reconstruction de certains passages manquants. Plus précisément, à l'aide d'un logiciel de dessin professionnel

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL

permettant d'analyser – à la main – les caractéristiques formelles de l'écriture d'un manuscrit (contours des caractères, espacement type entre les caractères et les mots, etc.), la méthode fut capable de restituer des caractères détériorés – voire parfois des phrases entières – uniquement à partir de points existants.

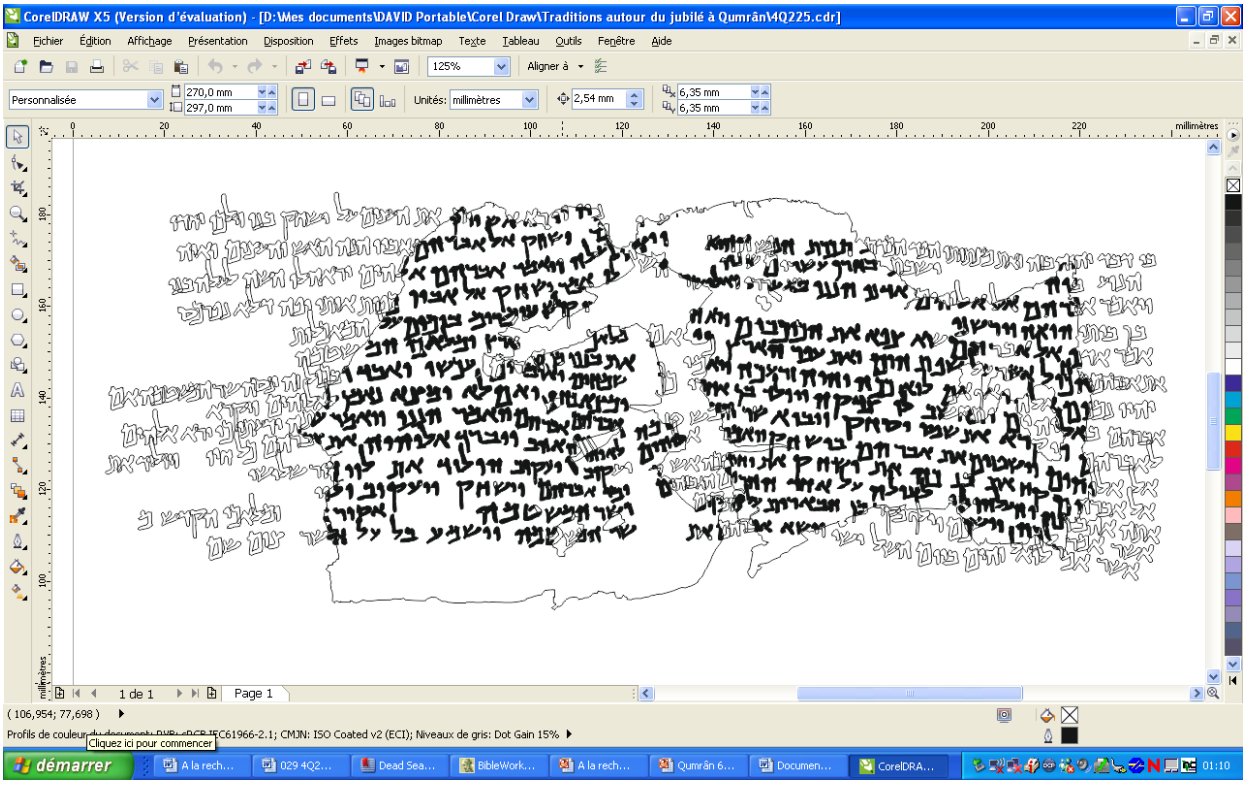

Figure 1: exemple d'hypothèses de reconstruction obtenus par la méthode d'apographie proposée par David Hamidovic. Une limite – mais également un apport – de cette méthode est qu'elle prend littéralement la "main" du scribe et ne peut donc fonctionner que pour les manières d'écrire d'un scribe pour un manuscrit.

Un projet différent mais qui se base également sur les propriétés pixelliques des caractères est actuellement en cours à l'université de Tel Aviv et tente d'analyser statistiquement les variations manuscrites afin d'identifier des profils de scribes. Si le projet est intéressant, il semble néanmoins négliger des variables importantes telles que le fait qu'un même scribe pouvait parfois orthographier le même mot différemment ou que plusieurs scribes pouvaient participer à la rédaction d'un *même* manuscrit. Toujours à Tel Aviv, un autre projet en cours est le façonnage d'un système d'information audacieux qui devrait parvenir à reconnaître les caractères d'un passage des manuscrits puis de s'en servir comme « accroche » pour restituer automatiquement les interprétations concurrentes de ce passage, stockées jusqu'alors dans plusieurs bases de données. À l'Ecole Normale Supérieure de Lyon, un autre projet de reconnaissance manuscrite se base sur les courbes des caractères afin de retranscrire puis traduire automatiquement des passages de manuscrits. Les résultats préliminaires sont a priori assez bluffants : entre 70 et 85% de traductions correctes. Mais qu'est-ce qu'une traduction « correcte » ?

Le projet Sokoka

Tout récemment lancé pour une période de trois ans, le projet Sokoka (nom araméen probable pour le site de Qumrân) s'accroche lui-aussi aux propriétés pixelliques des caractères des manuscrits mais avec l'ambition d'en apprendre davantage sur les personnages historiques que sont leurs différents scribes.

Auteurs et non pas copistes

Aujourd'hui, peu de connaissances ont été produites à propos des scribes des manuscrits. Dommage car d'un point d'histoire des religions, ces figures sont importantes tant ils ont une fonction d'auteur : plus que recopier, ils racontent, ajoutent, réarrangent, modifient en fonction de sensibilités différentes. Si les messages des textes ne se voient pas radicalement transformés, les façons d'y parvenir – parfois – changent. Car en tant que personne historique, chaque scribe a une culture, une sensibilité et produit ainsi des interprétations qui sont autant de traces laissées par tout un univers d'affections. Ils peuvent être ainsi considérés comme des connexions directes aux glissements confessionnelles de l'époque.

À partir de là, on peut faire l'hypothèse de l'existence de *projets éditoriaux* inscrits en creux dans les manuscrits de Qumrân et qui pourraient être reconstruits en se basant sur les caractéristiques formelles et sémantiques du contenu des manuscrits. L'ambition du projet Sokoka pourrait ainsi être conçue en termes de rétro-ingénierie en histoire des religions qui tentera in fine de reconstruire les diverses processions/interrogations de foi propres à des personnages historiques importants (les scribes) de cette période charnière (entre le III^{e} siècle avant et le I^{er} siècle après J.-C., entre Jérusalem et Jéricho).

Data selfies *et apprentissage non-supervisé (supervisé)*

Le concept de « data selfie » n'a pas seulement pour vocation d'être vendeur : il représente assez bien ce que le projet tentera de mettre au point, à savoir des profils de scribes nourris par le croisement de données sémantiques (mots, tournures de phrases) et formelles (courbures, espacements) issus des manuscrits. Car il ne faut oublier que leur rédaction s'étale sur une période de 250 ans, suffisamment longue pour générer plusieurs générations de scribes aux intérêts différents projetés sur des parchemins et papyrus de façons certes « pliées » et indirectes mais que les méthodes actuelles de *machine learning* ont – du moins théoriquement – les moyens de reconstruire.

Afin de procéder systématiquement à l'analyse formelle des manuscrits, il faut à coup sûr commencer par une entreprise de saisie de données et de formation de répertoires sur laquelle pourra venir se greffer des modèles d'apprentissage plus ou moins supervisés. Mais quel est le *meilleur* protocole à mettre en place pour la construction de cette base de données initiale ? Faut-il s'appuyer sur la méthode d'apographie précédemment présentée ? Son avantage est la précision mais son gros désavantage reste sa lenteur et donc son coût. Une autre possibilité pourrait être de découper/encadrer les images des manuscrits en fonction de leurs caractères.

UNIL | Université de Lausanne
LADHUL - Laboratoire de cultures et humanités digitales de l'UNIL

Mais une autre question apparaît aussitôt : faut-il se fonder sur les nouvelles images produites par Google en ultra-haute définition du manuscrit ou sur les plus anciennes, plus complètes (cf. supra) mais de moins bonne qualité ? De plus, est-il vraiment nécessaire de découper ou extraire tous les caractères de chaque manuscrit ? Et comment s'assurer que ce travail de saisie est correctement effectué ? Pour le moment, ces questions restent ouvertes.

La ligne de mire du projet est claire : celui-ci devra proposer sur une base de données capable de supporter des fichiers images et des répertoires de chablons – soit encadrés/découpés, soit dessinés – d'opérer sur eux, les croiser, les comparer, les classifier afin d'isoler des profils de scribes. L'idée est belle et aguicheuse : les manuscrits sont un corpus précieux, renommé et engager des partenariats avec des informaticiens actifs dans le domaine de l'analyse automatisée de caractères manuscrits ne devrait pas être chose impossible. Mais le développement de modèles d'apprentissages automatisés nécessite un gros travail préparatoire à la forte inertie : comment un historien de l'Antiquité pourra-t-il s'y retrouver dans les propositions informatiques et les irréversibilités qu'elles tendent à instaurer ? Comment commencer à construire cette base de données préparatoire constituée de caractères sans faire d'erreurs trop coûteuses à corriger ? En gros, vers qui se tourner lorsqu'il s'agit de débuter un tel projet *de la meilleure des façons* ? C'est précisément là que le LaDHUL – en tant que plateforme interdisciplinaire – se doit d'être un outil d'aiguillage tant il sait maintenant qu'une mise en base de données est un acte collectif à l'inertie forte, qui permet l'accrochage d'opérations tout en barrant la route à d'autres. En somme, même si des solutions rapides pour la base de données préliminaires de ce projet Sokoka sont nécessaires, ce « rapide » ne peut être que lent du fait des projections à faire, des scénarios à envisager et des compositions à instituer.

Florian Jaton