CONSEIL SUPÉRIEUR DES ARCHIVES

SÉANCE DU 11 DÉCEMBRE 2019

COMPTE RENDU

Étaient présents

- M. Jean-Louis DEBRÉ, président du Conseil supérieur des archives.
- Mme Annette WIEVIORKA, vice-présidente du Conseil supérieur des archives.

Membres de droit

- M. Philippe BARBAT, directeur général des Patrimoines.
- M. Louis DELESPIERRE, représentant Mme Brigitte PLATEAU, directrice générale de l'enseigneur supérieur et de l'insertion professionnelle.
- Mme Isabelle RICHEFORT, représentant M. Hervé MAGRO, directeur des archives du ministère de l'Europe et des Affaires étrangères.
- Mme Blandine WAGNER, représentant M. Sylvain MATTIUCCI, directeur du patrimoine, de la mémoire et des archives du ministère des Armées.
- Mme Laurence ENGEL, présidente de la Bibliothèque nationale de France.
- Mme Agnès MAGNIEN, représentant M. Laurent VALLET, président de l'Institut national de l'audiovisuel.
- Mme Anne DEBET, représentant Mme Marie-Laure DENIS, présidente de la Commission nationale de l'informatique et des libertés.
- M. Xavier ALBOUY, représentant M. Nadi BOU HANNA, directeur interministériel du numérique.

Personnalités qualifiées

- M. Alain CHATRIOT, professeur des universités.
- M. Thierry CHESTIER, président de la Fédération française de généalogie.
- Mme Nathalie LÉGER, directrice générale de l'Institut Mémoires de l'édition contemporaine.
- M. Jacques PEROT, président de l'Association française pour la protection des archives privées.
- Mme Sylvie THÉNAULT, directrice de recherche au Centre national de la recherche scientifique.

Représentants des organisations syndicales

- Mme Claire BÉCHU, pour la CGC.
- Mme Béatrice HÉROLD, pour la CFTC.
- Mme Violaine CHALLÉAT-FONCK, pour la CFDT-Culture.

Participaient avec voix consultative

- M. Philippe CIEREN, chef de l'inspection des patrimoines.
- M. Jean-Michel LOYER-HASCOËT, chef du service du patrimoine.
- M. Bruno RICARD, directeur des Archives nationales.

- M. Pierre LAUGEAY, directeur du Service historique de la Défense.

Représentants du service interministériel des Archives de France et intervenants

- Mme Françoise BANAT-BERGER, cheffe du service interministériel des Archives de France.
- M. Guillaume D'ABBADIE, délégué à la coordination et au pilotage des services publics d'archives (service interministériel des Archives de France).
- M. Jean-Charles BÉDAGUE, chef du bureau des études et des partenariats scientifiques (service interministériel des Archives de France).
- M. Frédéric KAPLAN, professeur à l'École polytechnique fédérale de Lausanne, directeur de l'Humanities Lab.
- Mme Marie LAPERDRIX, cheffe du Service des archives économiques et financières.
- Mme Marie-Françoise LIMON-BONNET, responsable du département du Minutier central des notaires de Paris des Archives nationales.
- M. Jean-François MOUFFLET, conservateur en chef au département du Moyen Âge et de l'Ancien Régime des Archives nationales.
- M. Angelo RIVA, professeur à l'École d'économie de Paris.

♦ OUVERTURE

Par M. Jean-Louis DEBRÉ, président du Conseil supérieur des archives

« Je ne voudrais pas ouvrir cette séance sans avoir une pensée pour Georgette Elgey, présidente de ce Conseil de 2007 à 2016, dont la personnalité, l'érudition et la pertinence des réflexions nous ont tous marqués. Elle est partie, elle nous laisse seuls, mais il faut continuer : c'est ainsi que va la vie.

« Il y a presque six mois que nous avons tenu notre dernière séance du Conseil supérieur des archives, le 28 juin dernier.

« Depuis, beaucoup de choses ont changé. Ces derniers mois ont été marqués par l'élaboration d'un nouveau Cadre stratégique de modernisation des archives, établi pour la période 2020-2024, dans le cadre du Comité interministériel aux Archives de France. Pierre angulaire de la politique des archives en France, il a été élaboré non seulement en étroite collaboration avec les administrations des archives des ministères des Armées et des Affaires étrangères, mais a également fait l'objet d'une large concertation avec les réseaux des archives ainsi que leurs partenaires institutionnels et associatifs. Il permettra ainsi aux plans stratégiques ou aux PSCE des services publics d'archives de s'inscrire dans les axes qu'il fixe.

« Par ailleurs, le programme interministériel d'archivage numérique VITAM s'achève avec succès, avec la mise en place de nombreux partenariats tant publics que privés qui prennent la forme de "club des utilisateurs", et la reprise de la maintenance du logiciel par le ministère de la Culture, toujours en étroite association avec les deux ministères des Armées et des Affaires étrangères.

« Le travail sur les archives des disparus de la guerre d'Algérie, dont il avait été question lors de notre précédente séance, a également progressé de manière significative, avec la publication d'une première dérogation générale portant sur les archives relatives à la disparition de Maurice Audin, mais également la publication très prochaine d'un magnifique guide de recherche, qui a l'ambition, par une présentation simple et ajustée des fonds conservés et de leurs modalités d'exploitation, de faire comprendre à tous, des familles aux chercheurs, comment accéder simplement aux sources relatives à cette question complexe, tout en indiquant, sans ambiguïté, les difficultés et les limites de la recherche. À l'occasion de la sortie de ce guide, une journée d'étude sera organisée aux Archives nationales sur les disparus de la guerre d'Algérie, et même plus généralement sur les nouveaux fonds collectés ou classés sur la thématique de la guerre d'Algérie. Sur le même sujet, vous avez également pu découvrir, dans le dernier numéro de *L'Histoire*, un très bel article de Sylvie Thénault, qui accompagne le livre qu'elle a coordonné avec Magali Besse sur *L'affaire Maurice Audin*.

« Les premiers dépouillements de sources relatives au Rwanda ont également

commencé, dans le cadre de la commission mise en place autour de Vincent Duclert, qui a également fait l'objet d'une présentation devant vous le 28 juin dernier.

« Enfin, conformément aux engagements pris lors de notre dernière séance, une réunion a été organisée avec les personnalités qualifiées du CSA pour à la fois évoquer les différentes questions qui pourraient être abordées lors de nos séances à venir et réfléchir collectivement aux questions de collecte et d'évaluation des archives. Cette réunion, qui s'est tenue le 18 septembre 2019, a permis d'évoquer plusieurs thématiques qui alimenteront l'ordre du jour de nos prochaines séances, et tout particulièrement de celle-ci : la thématique du passage « du papier aux données », que je vais vous présenter dans un instant, mais aussi les nouveaux modes de médiation entre les archives et leurs publics ; la tension entre ouverture et protection des données et la manière dont les Archives y répondent ; les questions de transparence et de pédagogie à destination des usagers et des publics. Un cycle de réunions a d'ailleurs été lancé, qui se déroulera pendant toute l'année 2020, avec une première rencontre aux Archives de Paris qui aura pour objectif de présenter concrètement les méthodes de travail des archivistes.

« Quant à la séance d'aujourd'hui, elle aborde une question qui se pose de manière cruciale depuis peu : comment rendre visibles et exploitables sur Internet, par des machines, les sources d'archives papier ? Du fait des mutations numériques de notre société, du fait de l'émergence de méthodes de travail et d'usages nouveaux chez les publics des archives, la question est désormais centrale et interroge notre avenir. Comment, en effet, répondre aux besoin d'usagers qui ne pourront ou ne voudront plus – voire déjà ne veulent plus ! – se déplacer dans les salles de lecture ?

« Sur un autre plan, l'exploitation des sources d'archives nécessite, pour une bonne partie d'entre elles, de passer du papier aux données numériques, via la numérisation, via l'océrisation, via la transcription, via l'indexation des contenus. C'est une nécessité si l'on veut pouvoir tirer parti, à grande échelle, des ressources qu'offre le numérique, et si l'on veut pouvoir comparer, connecter, confronter des sources conservées dans des institutions différentes, à Paris et dans les territoires, en France ou à l'étranger.

« Pour ce faire, il est absolument essentiel de constituer de vastes réservoirs de données, notamment de noms de personnes et de noms de lieux, dans le contexte du bi*g data* en émergence. Une intense réflexion est en cours, notamment pour favoriser l'alimentation et l'évolution du portail national FranceArchives, en lien avec les services publics d'archives, mais aussi d'autres partenaires dans le monde de l'enseignement supérieur et de la recherche, ou encore dans la sphère de l'économie et des finances, dans le domaine de la statistique notamment, pour alimenter les travaux des économistes et des sociologues ou les réflexions des décideurs politiques.

« Nous vous proposons donc ce matin de découvrir plusieurs initiatives visant à passer du papier aux données. Je pense qu'elles vous convaincront de l'immense potentiel qu'offre en la matière les archives, qu'elles datent du Moyen Âge, de l'époque moderne

ou de périodes beaucoup plus contemporaines. »

- **LES ARCHIVES EN PLEIN TEXTE : AVANCÉES ET PERSPECTIVES DE LA RECONNAISSANCE PAR ORDINATEUR DES ÉCRITURES**

  Par M^me^ Marie-Françoise LIMON-BONNET, responsable du département du Minutier central des notaires de Paris des Archives nationales, et M. Jean-François MOUFFLET, conservateur en chef au département du Moyen Âge et de l'Ancien Régime des Archives nationales.

Jean-François Moufflet commence par rappeler que la reconnaissance par ordinateur des écritures manuscrites s'inscrit dans la continuité d'un long processus de « numérisation » des archives, entamé depuis le début des années 1990. La reproduction par imagerie numérique des documents originaux et la conversion électronique des instruments de recherche traditionnels en vue de leur publication sur Internet ont ainsi permis de constituer des salles de lecture virtuelles. Aujourd'hui, c'est l'exploitation du contenu même du document qui est en jeu, avec l'indexation toujours plus fine et structurée d'informations tirées des sources, voire leur transcription intégrale, qui, dans le contexte de l'épanouissement des humanités numériques, offre d'autres usages pour la recherche.

La transcription par ordinateur des documents imprimés et typographiés était depuis longtemps possible avec les technologies d'OCR (*optical character recognition*, ou reconnaissance optique des caractères), ce qui permet aux usagers des bibliothèques numériques d'effectuer des recherches en plein texte, dans le cœur même des ouvrages. Mais tel n'était pas le cas des documents manuscrits, qui constituent une part très importante des fonds d'archives.

Le projet HIMANIS (HIStorical MANuscript Indexing for user-controlled Search), conçu et porté par l'Institut de recherche et d'histoire des textes (en l'occurrence Dominique Stutzmann, de la section de paléographie latine) de 2015 à 2017, et dont les Archives nationales ont été l'un des partenaires, s'est précisément donné pour but d'expérimenter la reconnaissance par ordinateur des écritures manuscrites sur un corpus archivistique bien identifié : les registres de la chancellerie royale des XIV^e^ et XV^e^ siècles, où les scribes recopiaient certains actes émis par le pouvoir. Il s'agit d'un ensemble bien connu des historiens, notamment pour l'étude des mesures à portée individuelle, comme les lettres de grâce et de rémission de peines, mais insuffisamment inventorié, ce qui signifie que l'on en connaît encore mal toute la richesse. En effet, dans la tranche chronologique retenue par le projet, seuls soixante-quatre registres sur cent soixante-seize, soit un tiers, ont fait l'objet d'analyses détaillées et d'une indexation complète des personnes, lieux et thèmes. Et ce ne sont pas les plus volumineux…

La recherche en plein texte dans les registres est une alternative pouvant pallier l'insuffisance des inventaires. Grâce à l'expertise technologique d'autres partenaires du

projet, notamment les sociétés françaises Teklia et A2ia ainsi que l'université de Valence en Espagne, l'intelligence artificielle est parvenue progressivement à se familiariser avec les formes de l'écriture médiévale et son système d'abréviations. Il est toutefois nécessaire, pour cela, de disposer d'un échantillon de transcriptions faites par l'homme, tout comme son intervention demeure fondamentale pour valider les hypothèses de reconnaissance de l'ordinateur et aider ce dernier à améliorer son modèle de reconnaissance. Une transcription automatisée de plus de soixante-dix mille pages a donc été faite, et une interface prototypale permet dès à présent d'interroger le texte des registres. Tout un chacun peut saisir n'importe quel terme (nom de lieu, de personne, terme courant) pour mesurer son occurrence dans les actes royaux enregistrés. La qualité de la transcription automatisée demeure variable, les registres de la fin du XV^e^ siècle posant de plus grandes difficultés. Une nouvelle reconnaissance des écritures est en cours, et l'on constate qu'elle ne cesse de s'améliorer.

La constitution d'un outil de recherche aussi puissant n'efface pas pour autant le rôle nécessaire de l'archiviste. Sa présence est d'autant plus fondamentale qu'il ne filtre plus les informations pour le chercheur : l'archiviste constituait auparavant des listes de noms de lieux ou de thèmes après un long travail d'identification et tâchait de traduire en concepts contemporains des termes anciens. Connaisseur des institutions ayant produit les documents, familier des schémas mentaux des époques antérieures, praticien du latin comme de l'ancien français, l'archiviste doit continuer à guider les chercheurs les moins aguerris.

La transcription des documents manuscrits fournit un matériau numérique qui se prête bien à des réutilisations diverses, notamment à une exploitation statistique. Ces corpus de données peuvent ensuite être traités par d'autres logiciels, qui pourront aider l'historien à en faire l'analyse. Le chercheur de demain s'apparentera-t-il à un analyste de données ? Toujours est-il que ces technologies éclairent d'une lumière nouvelle l'utilisation des archives. Les services d'archives, s'ils veulent continuer à être visibles dans l'écosystème numérique de demain, doivent expérimenter la reconnaissance d'écritures sur d'autres fonds.

Marie-Françoise Limon-Bonnet présente, à son tour, le projet « LectAuRep » (« Lecture automatique de répertoires »), qui consiste à soumettre à la reconnaissance automatique d'écriture manuscrite les répertoires de notaires conservés aux Archives nationales.

Elle rappelle que le répertoire de notaire est un document bien connu des chercheurs, grands utilisateurs d'archives notariales, tant aux Archives nationales que dans les Archives départementales. C'est un registre qui énumère, pour chaque jour, les actes passés dans une étude. Pour Paris, ces registres sont aujourd'hui tous numérisés, accessibles en mode image dans la salle des inventaires virtuelle des Archives nationales, mais il faut parfois patiemment feuilleter sur écran et lire des dizaines de pages de ces registres avant de tomber sur la mention de l'acte que l'on recherche, surtout si on ne connaît pas sa date exacte ou le nom du notaire qui a instrumenté.

Aussi, dans le cadre d'une convention cadre liant le ministère de la Culture et l'Institut

national de recherche en sciences et technologies du numérique (INRIA), convention destinée à faciliter le développement de recherches utilisant les technologies de l'intelligence artificielle, le Minutier central des notaires de Paris (Archives nationales) a soumis, fin 2017, à l'équipe ALMAnaCH (Automatic Language Modelling and Analysis & Computational Humanities) de l'INRIA, un projet visant à permettre, grâce au développement d'algorithmes innovants de reconnaissance d'écriture, un accès facile et pertinent aux contenus profonds des pages numérisées, par des recherches s'apparentant à la recherche en plein texte. Depuis la loi du 25 ventôse an XI (16 mars 1803) réorganisant le notariat, le document est en effet suffisamment normé pour se présenter sous forme de tableau dont les en-têtes et les colonnes pré-imprimées sont complétées de manière manuscrite par le notaire et ses collaborateurs. C'est donc ce corpus parisien remontant au XIX^e^ siècle et à la première moitié du XX^e^ siècle, constitué de plus de mille huit cents registres et comptant près de neuf cents notaires différents, qui a été retenu dans cette première expérimentation.

La phase 1 du projet (2018) a fait ressortir l'importance primordiale de la segmentation pour une reconnaissance d'écriture performante, ainsi que la nécessité d'entraîner les logiciels sur des fichiers numériques, des découpages de lignes et des mains de scribes aussi variés que possible, afin de diminuer le taux d'erreur par caractère, qui était de 40 % pour une écriture inconnue de la machine, même s'il était de moins de 10 % pour une écriture sur laquelle la machine avait été entraînée. Ayant d'abord testé la faisabilité de son projet à l'aide de la plateforme Transkribus, le Minutier central a rejoint une interface en cours de développement dans le cadre du projet « eScripta » intitulée « eScriptorium », et développée pour des projets similaires portant sur d'autres alphabets et d'autres écritures, à partir du logiciel de segmentation et d'océrisation Kraken.

Un échantillonnage d'images numériques plus diversifié a été constitué, afin de fournir des données d'entraînement plus riches, l'objectif de cette deuxième phase (2019) étant de parvenir à des modèles de segmentation et de reconnaissance d'écriture plus performants. Pour arriver à alimenter l'intelligence artificielle et la reconnaissance automatisée des écritures manuscrites, un travail de réflexion au long cours, mais aussi des manipulations très concrètes et en masse sont nécessaires ; aussi la phase 3, qui constituera le programme de recherche 2020, abordera-t-elle la problématique du passage du projet à l'échelle au moyen d'une interface collaborative ouverte aux contributeurs bénévoles.

Discussion

Thierry Chestier rappelle que le partenariat avec le monde associatif offre encore beaucoup de possibilités dans le domaine collaboratif. Il se réjouit à ce propos des liens forts qui unissent associations de généalogie et services d'archives. À propos d'indexation, il rappelle que les associations de généalogie disposent d'un nombre considérable de relevés, mais insiste sur la nécessité d'assurer la pérennité des permaliens qui y sont de plus en plus intégrés. Il signale également l'existence du logiciel Champollion, lauréat au salon Roots Tech de Salt Lake City en 2017, qui aide à la transcription de documents

manuscrits.

Angelo Riva s'interroge sur la manière de sélectionner les usagers qui contribuent aux entreprises collaboratives et de contrôler leurs interventions. Jean-François Moufflet lui répond que le projet HIMANIS, à la différence du projet LectAuRep, ne faisait pas appel à la participation du public, mais qu'il existe plusieurs façons de constituer et d'animer des communautés de contributeurs, selon l'ampleur des corpus et selon les projets : soit la plate-forme d'annotation est ouverte très librement, sans inscription préalable du contributeur, et l'on compte sur le public pour corriger des annotations défectueuses, soit, à l'inverse, on constitue des équipes que l'on encadre et dont on relit les transcriptions. Marie-Françoise Limon-Bonnet précise que plus la machine est entraînée, plus l'algorithme devient performant. S'agissant du projet LectAuRep, elle signale que, pour entraîner au mieux la machine, il faut passer par une série de filtres, auxquels correspondent différents profils de contributeurs, qui se voient confier des missions adaptées (transcription, vérification…) en fonction de leurs intérêts et de leurs qualités propres.

♦ DES ARCHIVES AUX DONNÉES : L'EXPLOITATION « BIG DATA » DES ARCHIVES DE LA COMPAGNIE DES AGENTS DE CHANGE DE PARIS

Par M^me^ Marie LAPERDRIX, cheffe du Service des archives économiques et financières, et M. Angelo RIVA, professeur à l'École d'économie de Paris.

Le Service des archives économiques et financières (SAEF) des ministères économiques et financiers conserve dans son centre de Savigny-le-Temple cent soixante mètres linéaires d'archives de la Compagnie des agents de change de Paris, ainsi que des collections de publications de la bourse officielle et de la bourse non officielle, dite « Coulisse ». Ce fonds est la source première de l'équipement d'excellence « Données financières historiques » (équipex « DFIH »), projet dirigé par Pierre-Cyrille Hautcœur, ancien président de l'École des hautes études en sciences sociales, et supervisé par l'économiste et historien Angelo Riva.

Depuis 2012, Angelo Riva supervise la construction d'une base de données contenant les informations relatives aux actifs cotés à la Bourse de Paris et à leurs émetteurs de 1796 à 1976. Les données de l'équipex « DFIH » sont mises gratuitement à la disposition des chercheurs via la très grande infrastructure de recherche (TGIR) Huma-Num.

Cette base de micro-données est donc susceptible d'éclairer d'un jour nouveau l'évolution du système financier français et de constituer une alternative à la base de données américaine « Center for Research in Securities Prices » (CRSP) de l'université de Chicago. L'Europe, et la France en particulier, ont un énorme potentiel de recherche en économie et finance, qui s'appuie sur une tradition d'excellence en mathématiques et statistiques.

Néanmoins, la crise financière récente a encore souligné la faiblesse des fondements empiriques des modèles d'analyse économique et financière qui sous-tendent les anticipations des acteurs, l'innovation financière et les régulations.

L'équipex DFIH est une base de données « ouverte ». Pendant la phase de conceptualisation du modèle logique des données, les évolutions futures de la base ont été pensées pour accueillir différents types de micro-données. La flexibilité des bases de données relationnelles donne la raisonnable certitude que des évolutions non anticipées pourraient trouver leur place au sein de l'équipement.

La construction de la base s'appuie principalement sur le fonds de la Compagnie des agents de change de Paris, et en premier lieu sur deux publications sérielles : l'*Annuaire des valeurs admises à la cote officielle* et l'*Annuaire Desfossés*. Les images numériques des sources sérielles numérisées pour la production de données ont vocation à être mises en ligne via la TGIR Huma-Num.

La stratégie mise en place pour alimenter la base dépend de plusieurs variables, principalement liées, d'une part, aux sources elles-mêmes et, d'autre part, à « l'état de l'art » de la technologie. En ce qui concerne les sources, l'accessibilité, la qualité des imprimés, l'organisation graphique des informations, la fréquence des changements dans les formats (et donc le type d'information contenue) sont des éléments qui ont été pris en compte. Dans le cadre de la construction de cet équipement, les contraintes et les opportunités étudiées par les archivistes, historiens et informaticiens ont conduit à la mise en place de deux stratégies de capture de données : d'une part, la saisie manuelle dans un environnement numérique *ad hoc* pour les cotes boursières jusqu'en 1950 par un prestataire privé à partir des images des sources ; de l'autre, le traitement semi-automatique des images numériques des annuaires boursiers et des cotes d'après 1950 par un logiciel spécifique fondé sur la reconnaissance optique des caractères et l'intelligence artificielle, et développé pour ce projet. En prenant en compte les coûts directs et indirects, le logiciel développé offre des gains par rapport à la saisie manuelle dans un environnement numérique. Pourtant, les deux stratégies de capture des données ont demandé un travail manuel préalable important, qui a été effectué par l'équipex dans les locaux du SAEF. Il a été nécessaire de créer, au sein de la base, une structure informationnelle qui permette aux opérateurs de saisie et au logiciel de ventiler opportunément les données dans la base.

Le Service des archives a su s'adapter aux exigences d'un tel projet en travaillant en étroite collaboration avec les chercheurs de l'équipex et les informaticiens associés. Une adaptation de la politique des publics a été nécessaire pour faire face à des équipes de chercheurs nombreux tout en assurant une ouverture permanente de la salle de lecture et des conditions de travail optimales pour tous les agents du service et les autres lecteurs.

La réalisation de l'équipex DFIH constitue par ailleurs pour le SAEF une opportunité majeure de valorisation et de réflexion sur le métier d'archiviste et les archives auprès de

l'ensemble des intervenants de la chaîne archivistique. En amont, auprès des services versants, c'est un exemple concret de l'apport que peut avoir l'exploitation des archives tant sur leur travail quotidien que sur la prévention des crises économiques. Le service des archives n'est plus uniquement le service qui gère la bonne gouvernance de l'information en interne aux ministères économiques et financiers, il produit de nouvelles données valorisables, diffusables, utilisables par les services des ministères économiques et financiers, et peut ainsi contribuer à l'évolution des politiques publiques : en interne au SAEF, en offrant au personnel l'occasion d'une réflexion sur les modifications apportées par les technologies numériques dans l'appréhension des documents et l'exploitation des données qu'ils contiennent ; en aval, par l'ouverture du centre des archives à un nouveau public et à de nouveaux modes d'utilisation des archives mises à sa disposition tant en France qu'au niveau européen.

Discussion

Bruno Ricard signale combien ce projet illustre le potentiel qu'offrent les archives, et en particulier les archives nativement numériques, pour de nouveaux usages, plus variés, autres que le travail historique ; en l'occurrence, ici, la prise de décision politique. Les services d'archives conservent en effet des masses considérables de données, notamment statistiques, qui nécessitent, pour des projets de ce type, des changements dans les modalités d'accès aux données et de leur mise à disposition, et notamment un accès sécurisé à distance, offrant des possibilités d'appareillage. Il faut que les services d'archives, dans les années à venir, soient en mesure d'offrir un tel mode d'accès et d'exploitation.

Laurence Engel souligne combien la transformation des collections en données – ou celle des données en collections – ouvre des perspectives heureuses. Elle estime que les projets mis en œuvre dans les services d'archives convergent avec ceux qui le sont dans les bibliothèques, et singulièrement la Bibliothèque nationale de France : elle signale ainsi un projet en cours de transformation d'une salle de lecture en salle de fouille de données à destination des chercheurs, qui autorisera un nouvel usage des données conservées à la Bibliothèque nationale de France.

À une question de Philippe Barbat sur les détenteurs des droits sur les outils, logiciels et algorithmes créés et la manière d'en faire profiter les usagers, Angelo Riva répond que les outils développés dans le cadre de l'équipex DFIH sont « ouverts » (les codes sources sont ainsi publiés sur le site du projet) et que leur développement n'a été mené qu'avec des institutions universitaires, pour ne pas être lié, par exemple, à une entreprise. Marie-Françoise Limon-Bonnet ajoute que les outils mis au point aux Archives nationales sont aussi « ouverts ».

Françoise Banat-Berger signale que la difficulté réside surtout dans l'incorporation *a posteriori* de tels outils dans les systèmes d'information des services d'archives.

- ## CONSTRUIRE AVEC LES USAGERS LES SERVICES DONT ILS ONT BESOIN : RETOUR SUR LE HACKATHON DES ARCHIVES NATIONALES (ET QUELQUES AUTRES EXPÉRIENCES)

  Par M^me^ Béatrice HÉROLD, directrice de l'appui scientifique aux Archives nationales.

Béatrice Hérold précise que son intervention a pour objet de rendre compte de quelques expériences menées ces dernières années aux Archives nationales qui s'inscrivent à l'articulation de leurs deux grandes stratégies : d'une part la volonté de mettre les usagers au cœur des politiques qu'elles conduisent, d'autre part la nécessité d'ajuster l'institution à l'évolution numérique de la société.

La stratégie numérique des Archives nationales a pour premier objectif d'adapter l'institution à la transformation numérique de l'État, et donc à la prise en charge de toutes sortes de formes de données dont elles doivent assurer l'intégrité, l'authenticité, la sécurité et l'intelligibilité dans le temps long de l'histoire. Elle s'incarne dans le projet ADAMANT, lui-même inscrit dans le programme interministériel VITAM. Pour construire l'offre de services autour de ces archives nativement numériques, il est apparu nécessaire de rencontrer les usagers des données, en particulier des statisticiens et des historiens qui les manipulent régulièrement dans le cadre de leurs recherches. Un groupe de travail s'est réuni en 2018-2019, copiloté par Claire Lemercier, directrice de recherche au CNRS, spécialiste de l'histoire sociale des activités économiques et familière des méthodes quantitatives. Les échanges ont permis de confirmer que les chercheurs avaient en effet besoin des données « brutes », telles que les Archives nationales les avaient collectées sans les manipuler ; mais ils ont émis le souhait qu'elles puissent leur offrir un service supplémentaire, c'est-à-dire leur restituer les données sous plusieurs formats, correspondant à des usages et des exploitations différentes. L'institution a donc bien pris en compte cette question de la multiplicité des formats de restitution dans le chantier d'ADAMANT portant sur l'accès et la diffusion des archives. Autre résultat de ce groupe de travail, Claire Lemercier et Étienne Ollion, chercheur au CNRS et professeur associé à l'École polytechnique, dont les travaux portent sur l'État et le pouvoir, ainsi que sur la numérisation des sciences et des sociétés, ont accepté de former quelques archivistes à leurs propres outils, pour partager de nouvelles compétences.

La question de l'offre de services à constituer pour les usagers est bien sûr plus large. Elle fait l'objet de l'attention des Archives nationales depuis des années. En 2013, l'ouverture de la salle des inventaires virtuelle projetait en masse les clés d'accès aux archives sur Internet. Mais cette offre de service traditionnelle, qui convient à une requête ponctuelle effectuée par un humain, est-elle suffisante ? En 2017, Françoise Banat-Berger, alors directrice des Archives nationales, eut l'intuition qu'il fallait aller plus loin, que d'autres usages étaient en train d'émerger, d'autres usagers à trouver. C'est ainsi qu'est née l'idée de faire un *hackathon*.

Pour commencer, les Archives nationales ont organisé en décembre 2017 un *barcamp*. Il s'agit d'une journée d'ateliers participatifs, très cadencés par une série d'exercices permettant de mettre à jour des idées cadrées de projets qui pourraient être développés, en l'occurrence à partir des données des Archives nationales. Une cinquantaine de participants extérieurs à l'institution ont été invités à venir réfléchir à la valorisation d'une vingtaine de jeux de données, autrement dit des inventaires et des référentiels.

Un an plus tard, en décembre 2018, avait lieu le premier *hackathon* organisé par un service d'archives en France. Partant des résultats du *barcamp*, ses objectifs se déclinaient autour de trois axes : faciliter l'accès aux ressources (amélioration de l'expérience utilisateur, *data visualisation*, géolocalisation, etc.), améliorer l'exploitabilité des données pour les rendre interopérables (dans la perspective d'alignement de données avec d'autres institutions) et, plus généralement, identifier des services numériques correspondant à de nouveaux usages pour les publics. Les Archives nationales ont restreint à cinq le nombre de jeux de données mis à disposition, en les sélectionnant par rapport à la thématique « Archives et citoyenneté ». L'ensemble des inventaires et des référentiels étaient également mis à disposition. La préparation de l'événement a été longue et très soignée ; elle a permis de se rendre compte que le format des inventaires n'était pas forcément adapté à leur réutilisation par des développeurs, ce qui a donné lieu à leur conversion dans des tableurs.

Après une soirée d'ouverture aux Archives nationales, le *hackathon* s'est déroulé pendant tout un week-end dans un lieu dédié à l'innovation numérique, le Liberté Living Lab, avec près de soixante-dix participants. Le projet lauréat, dénommé « Une minute ago », s'était attaché aux actes de notaire ; il permettait d'analyser et de visualiser l'activité du notaire à partir des données de l'inventaire, par le biais d'un affichage en carte (indiquant la localisation géographique des personnes à l'origine des actes) ou en graphe (pourcentage des actes par typologie), proposait aux internautes de compléter l'indexation fine des actes, et offrait la génération automatique de la demande de reproduction avec toutes les informations nécessaires, en particulier la cote de la liasse et la référence précise de l'acte. Enfin il suggérait un rebond vers d'autres ressources externes, comme Wikipédia, pour élargir la recherche au-delà des Archives nationales.

Certains de ces projets ont eu des suites, aux Archives nationales ou ailleurs. L'équipe lauréate du coup de cœur « innovation » avait eu l'idée de développer un *captcha* patrimonial ; ce projet est en cours de développement dans le cadre du dispositif « Service numérique innovant » soutenu par le ministère de la Culture. Les Archives nationales elles-mêmes sont en train de développer une nouvelle application pour mieux explorer et exploiter la base « LEONORE » des dossiers de titulaires de la Légion d'honneur, à partir d'un des prototypes fabriqués lors de ce *hackathon*.

La préparation de ces rencontres a permis de poser un diagnostic différent sur la qualité des données produites sur les archives originales. Les inventaires réalisés par les services d'archives offrent une description à la fois littéraire et très hiérarchisée, selon une arborescence des informations qui va du général au particulier ; ils sont faits pour être lus par des humains. Ce format n'est pas adapté à l'exploitation par une machine. C'est la

raison pour laquelle le troisième événement numérique des Archives nationales se concentrera sur ce matériau clé, la (méta)donnée. Le *datathon* des Archives nationales, organisé en décembre 2019, a pour ambition de se familiariser avec les enjeux du web de données.

Avec ces différents projets, les Archives nationales se positionnent comme un acteur du numérique, mais surtout expérimentent de nouvelles manières de se confronter avec leurs usagers, habituels et inhabituels. Elles espèrent ainsi inventer les services de demain et répondre toujours mieux aux attentes de leurs publics.

♦ PRÉSENTATION DE « TIME MACHINE PROJECT »

Par M. Frédéric KAPLAN, professeur à l'École polytechnique fédérale de Lausanne, directeur de l'Humanities Lab.

Frédéric Kaplan présente « Time Machine Project », projet soutenu par la Commission européenne à hauteur d'un million d'euros pour mettre au point une feuille de route de dix ans destinée à construire une plate-forme européenne de données archivistiques. Il rappelle le prototype qu'avait constitué le « Venice Time Machine », lancé en 2012, qui avait pour but de « construire un modèle multidimensionnel collaboratif de Venise en créant une archive numérique ouverte du patrimoine culturel de la ville couvrant plus de mille ans d'évolution ». Ce projet ne passait pas par la transcription complète des documents, mais par le repérage de mots-clefs dans des textes (par exemple le cadastre), qui permettait la création d'une base de données spatio-temporelle, associée à de l'iconographie. Elle offrait une représentation à la fois synchronique et diachronique de la ville.

Aujourd'hui, une vingtaine de villes en Europe se sont intéressées au projet, qui vise désormais à créer autant de « local time machines » : partant d'un modèle en trois dimensions auquel on attache des données, ces projets ont pour ambition de changer nos connaissances sur la ville, mais aussi de planifier son évolution – ce qui ouvre des perspectives en matière de prise de décision tant pour la puissance publique que pour les entreprises.

« Time Machine Project » s'appuie sur une organisation dont le siège est à Vienne, qui comprend trois cent quatre-vingt-quinze membres, dont cinquante-sept sont financeurs. L'objectif, à terme, est de créer un « *big data* du passé », dans l'horizon temporel propre à l'Europe.

Discussion

Sylvie Le Clech se demande si un tel projet ne pourrait pas séduire des services d'archives territoriales et s'interroge sur la taille critique que doit avoir un service ou une

collectivité pour adopter ce type d'outil.

Frédéric Kaplan lui répond que ce n'est pas plus simple pour une grande ville que pour une ville moyenne – au contraire peut-être –, mais que l'attention portée sur des grandes villes a d'abord été utile pour emporter l'adhésion symbolique et donc le financement des projets de recherche. Il espère, à terme, disposer d'un service directement utilisable par un service d'archives départementales. Du fait de la particularité du réseau français, un « pilote » adaptable à tout service d'archives départementales serait, selon lui, une piste très intéressante.

À une question de Nathalie Léger sur l'existence de protocoles, par exemple de collecte ou de structuration des informations, qui permettent de telles représentations de données, Frédéric Kaplan répond que, sur le plan technique, une « time machine box » est mise à la disposition du service pour l'accueil des sources premières, le choix des données à représenter se faisant ensuite avec le service d'archives (c'est ainsi la thématique de l'information urbaine qui a été retenue à Venise). Le processus est décentralisé, et laisse à chaque institution la liberté du choix du détail du projet.

Françoise Banat-Berger souligne combien ce type de projet est propre à rassembler tous les patrimoines (au-delà des archives et des bibliothèques, les monuments historiques, les objets archéologiques, les données de l'Inventaire…).

Frédéric Kaplan insiste sur l'intérêt des ressources de l'Institut géographique national, qui sont très utiles pour la reconstitution en trois dimensions de Paris, grâce aux photographies aériennes qui y sont conservées et qui permettent des restitutions calculées et non pas hypothétiques de la ville. Il évoque également les ressources audiovisuelles, sept entreprises de télévision publique étant partenaires du projet.

La séance est levée à 12 h 30.