Kelly CHRISTENSEN, Baudoin DAVOURY, Anahi HAEDO, Paul KERVEGAN, Esteban SANCHEZ-OECONOMO

Compte rendu de projet Git
TRANSCRIPTIONS DU CONGRÈS INTERNATIONAL DES SCIENCES ETHNOGRAPHIQUES (PARIS, 1878)

Devoir réalisé pour le master 2 TNAH de l'École Nationale des Chartes - Janvier 2021

1 Introduction

Le Congrès international des sciences ethnographiques de 1878 a eu lieu à l'occasion de l'Exposition universelle de 1878, à Paris. Édité en 1881 par l'Imprimerie nationale, le compte rendu de ce congrès a été mis à disposition par le Conservatoire numérique des Arts et Métiers.

Ce compte rendu permet de revenir aux débuts souvent problématiques de plusieurs disciplines scientifiques : ethnographie, anthropologie, sociologie. Il témoigne également des développements de l'archéologie, des études du folklore et de la culture populaire française. Le congrès est organisé en 1878, l'année où est créé le premier musée ethnographique parisien. C'est une date clé dans le développement de l'anthropologie et des études sur la culture populaire ; c'est pour cette raison que nous avons choisi de nous intéresser à cette édition du Congrès international des sciences ethnographiques.

2 Le Congrès international des sciences ethnographiques dans son contexte

2.1 La consécration des sciences anthropologiques aux expositions universelles du XIXe siècle

Les expositions universelles, carrefours commerciaux, intellectuels et techniques fédérant l'ensemble de l'activité humaine, furent des épisodes fondamentaux pour la consolidation des principaux paradigmes culturels et intellectuels du XIXe siècle. De par leur organisation encyclopédique des industries et des savoirs, elles constituaient des espaces privilégies pour formaliser et mettre en circulation des regards compréhensifs sur le monde. En établissant des vues globales et organisées de l'histoire du travail humain, elles affermirent les grandes idéologies de l'Europe libérale **(1)** ; leur mise en ordre du monde hiérarchisait les civilisations selon des critères matér
furent des espaces fondamentaux pour la consécration des sciences humaines. Ce furent les lieux de leur institutionnalisation, de leur affermissement épistémique et de leur patrimonialisation muséologique. L'anthropologie, l'ethnographie, l'archéologie et la préhistoire furent ainsi consolidées dans un même élan scientifique. Au-delà et au travers de leurs dimensions commerciales, ludiques **(3)** ainsi que de leur apologie des nouvelles vagues colonialistes, les expositions universelles permirent la vulgarisation d'une conception scientifique de la « race ». Catégorie centrale des sciences anthropologiques en gestation, cette notion manifestait un racisme positiviste qui caractérisa les pratiques savantes de la deuxième moitié du XIXe siècle. Le concept de race en circulation à l'époque des grandes expositions découlait du remaniement d'idées esthétiques, théologiques ou appartenant à des régimes de connaissance en déclin, transposées vers des cadres de pensée scientifique qui bénéficiaient désormais d'une autorité savante incontestée **(4)**. L'historiographie contemporaine propose de l'aborder avant tout (et au-delà de sa composante scientiste) en tant que processus de construction de la différence **(5)**.

Les sciences anthropologiques du XIXe siècle composent en outre un terrain dans lequel des théories protéiformes entrent en concurrence, avec pour fonction essentielle d'attester, de mesurer et de cataloguer des disparités biologiques et morales entre des groupes humains. Il est important de noter qu'il existe une corrélation entre les affiliations politiques et les postures scientifiques ; le tableau est pourtant extrêmement complexe. Des recherches récentes proposent ainsi de dépasser la mise en relation conventionnelle du racisme scientifique avec le conservatisme pour signaler l'alliance féconde entre le républicanisme et l'anthropologie raciale (6). Il est en effet indispensable de noter l'affinité idéologique entre les politiques coloniales de la IIIe

République, dont les valeurs furent matérialisées au sein des plus importantes expositions parisiennes (1878, 1889 et 1900), et les disciplines ethnographiques qui firent rentrer l'humanité dans un régime de connaissance qui légitimait l'avènement des dynamiques coloniales de la deuxième mondialisation.

2.2 Enjeux scientifiques et historiques du document retranscrit

L'exposition universelle de 1878 a été souvent caractérisée comme l'espace de consécration des sciences anthropologiques (notamment de l'anthropologie et de l'ethnographie). C'est également un marqueur pour la prolifération des congrès internationaux (7). Dans ce cadre, la société d'ethnographie de Paris organisa le Congrès international des sciences ethnographiques de 1878, tenu dans le tout nouveau palais du Trocadéro, qui abriterait désormais le nouveau musée d'ethnographie (1878-1908) (8). L'événement, mené par l'une des sociétés les plus influentes en la matière sur le plan international (l'affiliation de Charles Darwin est connue), fut ainsi organisé par certains de plus grands spécialistes de l'époque, tels Armand de Quatrefages, Paul Broca et Malte-Brun.

Le compte rendu sténographique des séances, tenues entre le 15 et 17 juillet 1878, fut édité et publié en 1881 par l'Imprimerie nationale. Ce document est un témoin privilégié pour appréhender le champ épistémique des sciences ethnographiques à l'aube de leur institutionnalisation. Sa vocation généraliste, l'amplitude des thèmes abordés, et le concours des grands spécialistes de l'époque en fait un écrit précieux pour comprendre un terrain en formation, traversé par de violentes controverses et des idées novatrices. Un exemple en est la liste de « Questions » (p.9-13) adressée aux réseaux internationaux, formels ou informels, des sciences humaines en gestation, qui manifeste de manière claire les grandes problématiques de ces disciplines au XIXe siècle : le polygénisme et le monogénisme, le peuplement de l'Amérique précolombienne, les critères de différenciation raciale. Un autre exemple est celui des contributions assidues de Clémence Royer, traductrice de Darwin reconnue dans le monde entier : sa préface à l'Origine des espèces circula profusément dans les pays d'Amérique latine et imposa une lecture « francisée » de la théorie.

Ce projet propose la transcription et OCR d'un échantillon d'articles contenus dans le Compte-rendu du congrès international des sciences ethnographiques de 1878, en tant que première étape dans une réflexion menée autour des enjeux historiographiques, scientifiques et numériques actuels de la recherche sur les expositions universelles et sur l'histoire des sciences humaines au XIXe siècle. Ce document est représentatif d'un patrimoine imprimé large : des centaines de publications officielles, de rapports de congrès et d'autres événements liés ont été produits dans des conditions techniques et intellectuelles extrêmement similaires. Le développement d'un modèle de reconnaissance de caractères imprimés adapté à ce document serait susceptible de contribuer à créer des instruments permettant d'exploiter un corpus numérique ample, produit dans notre cas par le Conservatoire numérique des arts et métiers (CNUM) (http ://cnum.cnam.fr/thematiques/fr/1.expositions_universelles/cata_
Reste à signaler l'existence d'autres corpus mis à disposition sur Gallica (https ://gallica.bnf.fr/accueil/es/content/accueil-es ?mode=desktop) et Europeana (https ://www.europeana.eu/fr), ainsi que sur de nombreux sites à des échelles plus modestes. Ce projet pourrait ainsi contribuer à une mouvance large, portée par des acteurs tels que le Bureau international des expositions (BIE) et son nouveau World Expo Museum (Shanghai, 2015) (https ://www.editionsducerf.fr/librairie/livre/19381/paris-capitale-du-xixe-siecle-ned), dont le centre de documentation vise à fédérer la recherche internationale et à proposer des instruments numériques renouvelés. À terme, le but est de repérer, de signaler et de valoriser numériquement un patrimoine disséminé dans le monde entier.

3 Répartition des transcriptions

4 Choix techniques

4.1 Le modèle de transcription

Le document traité étant un imprimé de la fin du XIXe siècle, nous avons choisi d'utiliser le modèle `19th century prints HTR catalogs`, développé dans le cadre du projet Artlas, dirigé par Béatrice Joyeux-Prunel en partenariat entre l'Université de Genève, le centre IMAGO (financé par Erasmus+), l'EUR Translitterae, le Labex TransferS et financé entre 2011 et 2016 par l'ANR.

4.2 Le format d'export XML

4.3 Répartition des branches Git

Le projet faisant l'objet d'une collaboration entre cinq étudiant.e.s, l'utilisation de Git et d'un dépôt distant sur Github ont été nécessaires pour mener à bien le projet. Pour travailler à plusieurs sans entrer en conflit, tout en ayant accès aux travaux des autres, le dépôt Git a été divisé en 6 branches :

- `main` : la branche principale, utilisée pour conserver les données du projet une fois que celui-ci a été mené à bout. Dans le déroulement du travail, cette branche n'a pas été utilisée, sauf à la toute fin du projet, ou le contenu de chaque branche a été basculée sur `main`.
- `dev` : la branche de développement, utilisée pour tous les apports au projet qui ne concernaient pas le travail d'une personne en particulier, ou qui concernaient tou.te.s les membres du projet. C'est sur cette branche que se trouve le fichier `extraits.txt` (qui contient les références aux articles retranscrits) ainsi que le rapport et la documentation du projet.
- Ensuite, **les cinq branches personnelles** des membres du projet, qui contiennent les exports XML des documents retranscrits. L'utilisation de branches personnelles permet à chaque contributeur.ice de travailler de façon autonome, sans que les retranscriptions de chacun.e ne se mélangent. La définition de normes (noms de fichiers...) garantissent une homogénéité entre ces branches et permettent la cohérence du projet dans son ensemble.
  - `branche_anahi` : la branche utilisée par Anahi
  - `branche_baudoin` : la branche utilisée par Baudoin
  - `branche_esteban` : la branche utilisée par Esteban
  - `branche_paul` : la branche utilisée par Paul
  - `kelly-eScriptorium` : la branche utilisée par Kelly

5 Les données produites

6 Difficultés et problématiques particulières

6.1 L'encodage des caractères spéciaux et non-latins

L'utilisation de mots en écrits en alphabets non-latin dans les articles retranscrits a été un défi pour l'encodage. Le modèle eScriptorium `19th century prints` ne reconnaît pas une partie importante des citations de langues étrangères dans les comptes rendus. Les mots en polonais, latin, allemand, et anglais ont

été correctement reconnus par le modèle. En fait, le modèle les a reconnus avec un degré d'exactitude égal à celui des mots en français, la langue principale du document. Les langues non latines, par contre, n'étaient pas reconnues : ce sont le grec, le sanscrit, le russe, le persan, ainsi que les langues peu connues tel que le vieux-slave et l'avestique que les ethnologues ont appelés par leurs nom anciens, le paléoslave et le zend respectivement. Afin de transcrire ces langues, il fallait trouver l'unicode de chaque caractère spécial.

On a utilisé deux méthodes pour trouver l'unicode des caractères spéciaux pas capturés par le modèle. Parfois les auteurs ont cité un passage d'un texte connu, tel que *l'Iliade* d'Homère. Dans ce cas, on a vérifié notre transcription de la citation avec des encodages du texte déjà disponibles en ligne. Mais la plupart des caractères spéciaux étaient imprimés sans une citation. En outre les auteurs n'ont pas toujours précisé à laquelle langue les mots qu'ils ont invoqués appartiennent. Dans ce cas, on avait besoin de, en premier, reconnaître la langue et, ensuite, de rechercher l'unicode de chaque caractère.

Pour les caractères spéciaux en grec, qui constituent la majorité, notre équipe a profité du fait que l'un des membres, Paul Kervegan, soit familier avec le grec. En travaillant en équipe pour décoder un caractère spécial, on a relevé le fait que l'un des comptes rendus a écrit le lettre *tau ( )* avec un caractère une graphie qui n'est plus en usage aujourd'hui. Malheureusement, faute d'avoir un unicode pour encoder ce caractère, on n'a pas pu garder le choix unique des auteurs. On l'a encodé avec la lettre moderne, , en privilégiant la lisibilité du mot.

Pour les caractères spéciaux en sanscrit, notre équipe a compté sur l'expertise de Christian Iyer. Il a relevé le fait que les comptes rendus de 1878 ont utilisé une fonte rare pour le sanscrit ; donc notre encodage ne ressemble pas toujours à l'imprimé. Heureusement, les auteurs ont souvent précisé le mot en sanscrit avec une translittération, ce qui a aidé à décoder la transcription d'une fonte peu commune. Ces problèmes relatifs aux choix de graphies dans l'imprimé original soulèvent des questions intéressantes vis-à-vis de l'entraînement d'un modèle d'OCR : est-il possible, ou pertinent, d'entraîner un modèle à reconnaître des graphies qui ne sont plus en usage aujourd'hui ? Est-ce qu'il vaut mieux privilégier une spécificité historique du document (l'utilisation de graphies propres au XIXe siècle) ou développer un modèle qui cherche à rendre les textes lisibles pour un public contemporain ?

Après le grec et le sanscrit dont les mots étaient le plus nombreux, nous avons encodé des mots en russe, en persan, en avestique, et le vieux-slave. Le vieux-slave a deux écritures, le glagolite et la cyrillique **(1)** ; c'est ce dernier que les auteurs ont utilisé. Sachant que le persan et l'avestique s'écrivent de droite à gauche, on a recherché caractère par caractère dans un dictionnaire. Les dictionnaires de persan, autrement connu comme le farsi, se trouvent facilement et l'encodage est normalisé. Faute d'un expert.e en farsi, on a vérifié notre encodage du mot en comptant sur la translitération « hest », traduction de « il est », fournie par les ethnologues. Pour encoder l'avestique on a consulté la recherche de Jost Gippert, spécifiquement son article *The Encoding of Avestan — Problems and Solutions* **(2)**, qui nous a fourni l'unicode des trois caractères spéciaux du mot cité dans le compte rendu.

6.2 L'encodage des variations typographiques

Par ailleurs, eScriptorium, bien que parfaitement capable de lire et restituer des mots et termes écrits en italique dans les documents choisis, ne les retranscrit pas pour autant en italique. Cet aspect, qui peut sembler mineur, peut être regrettable. En effet, il pourrait être intéressant que les fichiers XML exportés depuis les transcriptions soient capables de baliser automatiquement les termes en italique. En effet, l'intêrêt d'avoir un mot en italique est souvent de mettre en avant son aspect particulier (mot étranger ou notion avec une définition technique). Par exemple, dans les textes de Paul Kervegan et Baudoin Davoury sur les

différences entre la race, le peuple, la nation et l'Etat, ces quatre notions étaient écrites en italique afin de rappeler au lecteur que ce ne sont pas des généralités. Chacune de ces notions a reçu de l'auteur un sens précis, qui fait sa spécificité dans le texte et qui est ainsi rappelé en permanence pour que le lecteur l'ait bien en tête afin de comprendre et appréhender le contenu intellectuel du texte lu.

Table des matières