Cycle de webinaires (2025-2026)
Approches computationnelles du monde chinois

Présentation générale. Ce cycle d'ateliers-webinaires a pour objectif de créer un espace d'échanges et de réflexion autour de l'usage des méthodes computationnelles appliquées au monde chinois. Il part d'un constat simple : nous faisons face à "l'impératif numérique" depuis de nombreuses années, sans que le paysage académique (en termes de formation, de production scientifique) ait profondément évolué. La vague de l'intelligence artificielle (IA) qui a déferlé récemment a révélé à quel point nous étions mal préparés et largement isolés face à ces enjeux technologiques qui nous dépassent. En tant que spécialistes de la Chine, nos domaines d'étude nous confrontent à des problèmes spécifiques (liés aux langues non occidentales, à la numérisation des sources, à l'accès aux données), mais nous manquons de lieu pour nous rencontrer et pour la plupart d'entre nous, nous n'avons pas reçu de formation adaptée pour y faire face.
Ce webinaire vise à combler ce manque et à susciter une réflexion partagée autour des changements technologiques qui affectent la recherche sur la Chine, en partant de nos sources et de nos questions de recherche. L'objectif est de discuter librement des problèmes que nous rencontrons individuellement et d'imaginer des solutions ou des bonnes pratiques communes, tout en découvrant (ou redécouvrant) les travaux des uns et des autres.

Public visé. Ces webinaires s'adressent à tous les étudiants et chercheurs spécialistes de la Chine/Asie, quels que soient leur discipline, espace et période d'étude, et quel que soit leur niveau de pratique des outils numériques (de la simple découverte aux utilisateurs expérimentés).

Organisation. Chaque séance durera environ 1h30 à 2h, suivant un format souple, généralement une présentation par l'intervenant suivie d'une discussion avec l'ensemble des participants. En général, les séances auront lieu le jeudi midi en début de mois.

Pour recevoir le lien de connexion la veille, merci de vous inscrire ici.

Si vous souhaitez animer une séance, merci de contacter : cecile.armand@cnrs.fr

Programme pour l'année 2025-2026

Séance 1 (Jeudi 9 octobre, 12h-14h) : De Dunhuang à Baxian : dernières avancées dans la transcription automatique des sources historiques chinoises.

Colin Brisson, Ecole Pratique des Hautes Etudes (CRCAO)
Frédéric Constant, Université Côte d'Azur (CRCAO)

La transcription automatique de textes (HTR/OCR) est l'une des applications pionnières de l'intelligence artificielle. Avec le projet Wenyuange Siku Quanshu Electronic Version initié 1996, la République populaire de Chine a été le premier pays à s'appuyer sur cette technologie pour numériser son patrimoine littéraire. Si l'on obtient aujourd'hui de bons résultats sur les imprimés modernes, les outils existants achoppent à produire des transcriptions de bonne qualité des documents historiques. En raison de la diversité des mises en page et des styles d'écriture, le taux d'erreur élevé impose un travail de correction fastidieux et chronophage. Ce coût a jusqu'ici limité la numérisation à grande échelle aux grands projets commerciaux (Airusheng) ou collaboratifs (Shidianguji, Ctext), conduisant à une offre de corpus numériques restreinte et souvent redondante. Cette présentation montrera comment les dernières avancées technologiques permettent de surmonter ces obstacles. En nous appuyant sur deux projets en cours – Read_Chinese (BnF Datalab), qui porte sur le fonds Pelliot chinois de la BnF, et CRISOLIC (ANR-24-CE27-4500-03), qui étudie les archives administratives de la fin des Qing – nous démontrerons qu'il est désormais possible de produire à moindre coût des transcriptions de haute qualité pour des corpus manuscrits aux écritures et mises en page très diverses.

Lien de connexion séance 1

Séance 2 (Jeudi 4 décembre, 12h-14h) : Bridging the Close/Distant Reading Divide : Analyses de corpus historiques sur la Chine moderne (19-20e siècles)

Cécile Armand, CNRS (IAO)
Christian Henriot, Aix-Marseille Université (Irasia)

Depuis une trentaine d'années, les sources historiques numériques accessibles aux chercheurs de la Chine moderne se sont multipliées. Cette numérisation massive ouvre de nouvelles perspectives, mais soulève aussi des défis inédits, liés non seulement à l'échelle des corpus, mais aussi à leur multilinguisme, à la diversité des genres et des supports, ainsi qu'aux biais introduits par des programmes de numérisation souvent opaques. L'essor de l'intelligence artificielle (IA) a renforcé ce mouvement, invitant à repenser les corpus non plus seulement comme des textes à lire individuellement, mais comme des réservoirs de données requérant des méthodes de traitement automatiques. Pour autant, les outils computationnels demeurent encore largement sous-exploités, voire mal compris par les historiens, de sorte que la richesse de ces corpus et leur potentiel de transformation historiographique restent en grande partie inexplorés.
Cette session présentera les travaux du projet ENP-China, qui s'efforce de relever ces défis en intégrant pleinement les méthodes computationnelles au cœur de la recherche historique. Le projet s'attache à structurer et à enrichir sémantiquement de vastes corpus textuels (presse, annuaires, archives, journaux intimes, dictionnaires biographiques, données du web), disponibles en texte intégral, et à développer des méthodes adaptées pour extraire, organiser, et analyser l'information historique qu'ils contiennent. A travers une collaboration inédite entre

historiens et chercheurs en informatique, ce travail a conduit au développement de la Modern China Textual Database (MCTB) et de l'application HistText, un outil basé sur l'IA dédié à la construction et à l'analyse de corpus sur mesure. Des études de cas viendront illustrer concrètement comment cet écosystème peut renouveler notre connaissance de la Chine moderne et contemporaine.

Lien de connexion séance 2

Séance 3 (Jeudi 5 février, 12h-14h) : Entre accès restreints et algorithmes opaques : impasses et détours d'une recherche « tout en ligne » en Chine

Virginie Arantès et Cinzia Losavio (CNRS, project Chine CoRef)

À partir de deux recherches postdoctorales menées dans le cadre du projet Chine CoRef, cette communication interroge les limites d'une enquête reposant exclusivement sur des ressources numériques en Chine et explore les moyens de dépasser certaines de ces contraintes.
Au-delà des obstacles d'accès (identité numérique chinoise, numéros locaux, restrictions d'API), l'enjeu majeur tient à la double médiation des données : par les algorithmes de filtrage et de classement, mais aussi par des formes narratives et idéologiques qui transforment les contenus (annonces immobilières incomplètes ou filtrées par des logiques commerciales et politiques, posts esthétisés et patriotiques sur les parcs).
La présentation discutera ainsi des ressources et compétences nécessaires pour mener des enquêtes plus fiables – outils techniques (téléphones chinois, VPN spécialisés), maîtrise pratique des applications, analyse critique des récits produits – et montrera pourquoi, dans ce contexte, une recherche « tout en ligne » doit nécessairement être articulée à d'autres méthodes de terrain.

Lien de connexion séance 3

Séance 4 (Jeudi 2 avril, 12h-14h) : "Ethnic ordering in the People's Republic of China: a text-as-data approach"

Jérôme Doyon, Science Po Paris

This project proposes to systematically exploit volumes produced by the Chinese party-state about its own organizational structure and, more specifically, the Materials on the Organizational History of the Chinese Communist Party (组织史资料, Organizational Histories). These underused sources not only help us compensate for the increasing difficulty of access to the field and data in the study of Chinese politics; they also provide the richest available account of the history of the party-state structure and its human resources. The project is structured around two working packages (WP), a methodological one and a thematic one. As part of WP1, the research team plans to generate a large textual corpus suitable for data extraction based on the Organizational Histories to produce open-access data that will greatly advance and transform future research on Chinese politics. With WP2, we will leverage this data to advance our understanding of ethnic politics in China, a topic made particularly relevant but also sensitive in today's Chinese politics, given the Party-state's highly repressive approach to the management of ethnic minorities, especially in Xinjiang and Tibet.

Lien de connexion séance 4

Séance 5 (Jeudi 4 juin, 12h-14h) : Analyse de l'évolution du chinois du point de vue syntaxique

Wu Qishen, Inalco

Plus d'informations très bientôt.

Lien de connexion séance 5