Les «Big data» du passé

Frédéric Kaplan, EPFL Digital Humanities Laboratory

Les sciences humaines sont sur le point de vivre un bouleversement comparable à celui qui a frappé la biologie dans les trente dernières années. Cette révolution consiste essentiellement en un changement d'échelle dans l'ambition et la taille des projets de recherche. Nous devons former une nouvelle génération de jeunes chercheurs préparés pour cette transformation.

A l'EPFL, je donne un cours de Master dédié à cette thématique. Tout mon cours est construit autour d'un projet international, intitulé la «Venice Time Machine» et mené en collaboration avec l'université de Venise Ca'Foscari. Ce projet de recherche ambitionne de modéliser 1000 ans d'histoire vénitienne et méditerranéenne à partir de la numérisation des immenses archives de la ville, près de 80 km de documents anciens détaillant avec précision le fonctionnement quotidien de la cité des doges. Chacune des étapes de ce projet, depuis la numérisation massive de documents anciens jusqu'à l'exploration et la mise en scène des données qui peuvent en être extraites, me sert de fil rouge pour introduire les principaux concepts et méthodes nécessaires à la réalisation de ce type de projets. Au fil des semaines, les étudiants découvrent de quelles manières les humanités digitales permettent de reconstruire des «Big data» du passé, et comment cette nouvelle abondance de données transforme les approches traditionnelles en sciences humaines et sociales.

L'art des techniques de numérisation

Nous commençons par faire l'état de l'art des techniques de numérisation, pour les manuscrits, les livres imprimés, les peintures, mais aussi des objets patrimoniaux de plus grande taille, statues, façades, bâtiments, etc. Je fais découvrir aux étudiants comment nous pouvons aujourd'hui envisager des programmes de numérisation à très grande échelle, permettant de «scanner» des centaines de livres en une seule journée, par des procédés venant d'autres disciplines, comme l'imagerie médicale.

Nous travaillons dans les semaines qui suivent sur le problème de la transcription semi-automatique des documents anciens. Nous commençons par le problème plus simple de l'alignement d'une image d'un texte avec sa transcription puis étudions la manière dont il est possible de combiner les connaissances sur la structure des documents, l'histoire et la linguistique pour réussir à transformer des images numériques de document en textes indexables.

Modélisation de l'information historique

Nous abordons ensuite les questions de modélisation de l'information historique contenue dans ces documents transcrits. Nous étudions plusieurs algorithmes d'extraction sémantique permettant de repérer personnes, lieux, dates et événements dans des corpus de textes beaucoup trop vastes pour qu'un seul chercheur puisse les lire. Se pose évidemment la question des incertitudes, des incohérences intrinsèques à ces documents historiques, mais aussi celle des erreurs introduites par les processus de numérisation et d'extraction automatique. Nous découvrons ainsi l'importance de toujours modéliser les processus intellectuels et techniques qui sous-tendent l'établissement des informations historiques. C'est à nouveau la masse des données disponibles qui peut nous permettre de dégager des données fiables à partir d'une multitude d'informations incertaines.

Il s'agit ensuite de placer ces informations dans le temps et l'espace, d'adapter les systèmes d'information géographique pour qu'ils intègrent la dimension temporelle et de tester la cohérence des informations extraites dans ces nouveaux systèmes de coordonnées géohistoriques. Nous étudions comment construire des «Google maps» du passé, équipées d'une interface

permettant de faire défiler les années et de voir à quoi ressemblait une ville comme Venise, il y a 100, 500 ou 1000 ans.

Extraire les grammaires sous-jacentes

Même avec la densité d'information dont nous disposons dans le cas des archives de Venise, il nous manque toujours des éléments pour pouvoir reconstruire avec précision des cartes et des réseaux du passé. Nous pouvons alors tenter d'extrapoler et inférer à partir des données existantes les éléments manquants. Nous apprenons à extraire les grammaires sous-jacentes aux données historiques déjà reconstruites et compléter les éléments incertains en généralisant à partir des motifs identifiés. Il s'agit bien évidemment de s'interroger à nouveau sur nos procédés de modélisation et la représentation de l'incertitude attachée à ces données simulées. Les étudiants commencent alors un projet en équipe au cours duquel ils explorent comment dans un cas particulier lié aux archives de Venise ils peuvent utiliser les méthodes apprises en cours.

Nouvelle génération

Mes étudiants sont les premiers à faire chaque jour l'expérience d'un présent toujours plus dense (10 millions de nouvelles photos par heure sur Facebook; une heure de vidéo par seconde sur YouTube, 400 millions de tweets par jour). Ils seront peut-être demain ceux qui, à l'interface des sciences de l'information et des sciences humaines, seront capables de maîtriser les outils de «grand maintenant» pour reconstruire un passé dense et épais, à portée de clics.

L'auteur

Prof. Frédéric Kaplan

Prof. Frédéric Kaplan holds the Digital Humanities Chair at Ecole Polytechnique Federale de Lausanne (EPFL) and directs the EPFL Digital Humanities Lab. He conducts research projects combining archive digitisation, information modelling and museographic design. He is currently working on the «Venice Time Machine», an international project in collaboration with the Ca'Foscari University in Venice, aiming to model the evolution and history of Venice over a 1000 year period.