Quelques réflexions préliminaires sur la Venice Time Machine

Frédéric Kaplan

DHLAB, Ecole Polytechnique Fédérale de Lausanne (http://dhlab.epfl.ch)

frederic.kaplan@epfl.ch

1. Les big data du passé

Les utilisateurs de Facebook partagent environ 10 millions de nouvelles photos par heure. Sur YouTube, une nouvelle heure de vidéo est mise en ligne par seconde. Chaque jour, près de 400 millions de nouveaux tweets documentent minute par minute la pulsation du monde. Le système informationnel que ces différents flux irriguent ne se cesse de croitre, constituant une base documentaire sans précédent sur notre époque. Ce déluge de données a ouvert un champ de recherche nouveau: les “Big data”. Stocker, indexer et repérer des structures dans ces flux est un des enjeux majeurs des premières décennies du XXIe siècle.

Mais cette augmentation exponentielle du volume des données est aussi un facteur de déstabilisation culturelle. Nous produirions aujourd’hui plus de données en 2 jours que l’ensemble des données numériques disponibles il y a dix ans. Chaque jour, l’Internet ressemble un peu plus à un panoptique nous donnant à voir un “grand maintenant”, complexe, infiniment dense et parfaitement documenté, mais qui sans la disponibilité de données numériques équivalentes sur le passé ne s’inscrit plus dans une temporalité longue. Ce déséquilibre est inquiétant, car une société ne saurait envisager avec sagesse son futur si elle n’est pas capable de comprendre sa trajectoire sur la durée.

Pourtant les données sur le passé existent et en très grande quantité. Les grands empires, mais aussi les états plus petits, se sont presque toujours construits grâce à la mise en place de politiques de traçabilité informationnelle consignant avec le plus de précision possible, non seulement la cartographie de leur territoire, mais aussi la vie de chaque habitant de leurs cités. Dans les grandes archives européennes, des milliers de kilomètres de registres regroupant les informations minutieusement enregistrées par ces "Google" du passé attendent d'être analysés. À ces archives s'ajoutent des constellations de documents privés, dont la volumétrie est difficile à évaluer, mais qui constituent un contrepoint informationnel précieux à la mémoire administrative. Depuis quelques années, plusieurs initiatives pour numériser, transcrire, indexer massivement ce matériel historique ont vu le jour. Il s'agit de développer les technologies et les savoir-faire pour produire à partir de ces archives, des "Big data du passé". L'enjeu de ces recherches est immense : il consiste à rendre enfin accessible des milliards d'informations qui échappent aujourd'hui à la culture numérique et donc peut-être à moyen terme, à la culture tout court. Pour les technologies de l'information, la reconquête numérique du passé est la prochaine frontière.

2. Rendre le passé présent

En 1896, à Genève, les deux millions de visiteurs de l'exposition nationale pouvaient pour la première fois découvrir leur ville depuis le ciel. Par petits groupes, depuis la nacelle d'un ballon à hélium, ils voyaient comment la ville s'était considérablement transformée dans les dernières décennies. Dans le sillage des autres capitales européennes en cette fin du XIXe siècle, l'ancienne cité médiévale était maintenant devenue une ville ouverte, aérée, cosmopolite, construite pour la promenade plus que pour la défense, une ville de son temps. Cette vue du ciel, absolument inédite à une époque ou les voyages aériens n'était encore qu'un rêve d'avenir, permettait d'un seul coup d'œil d'embrasser la structure et les motifs de la ville nouvelle. Elle donnait à voir un point de vue et une compréhension qu'aucun discours n'aurait pu résumer.

Essayons d'imaginer l'expérience de ces visiteurs de la fin du XIXe siècle, peu coutumier aux vues aériennes dont beaucoup montaient pour la première fois dans un ballon. L'ivresse du panorama provient d'une transformation soudaine du rapport entre l'espace et le temps, une compression spatiotemporelle. D'un point unique, il devient tout d'un coup possible d'entrevoir sous une nouvelle forme des lieux qu'il fallait des heures à atteindre. La surabondance soudaine de l'information disponible donne le vertige. Où regarder ? L'œil cherche désespérément des motifs globaux ou au contraire un détail singulier pour rendre intelligible ce spectacle trop riche, trop dense, pour le faire revenir dans l'ordre du discours, pour être capables une fois redescendu de dire ce que l'on a vu. Mais le panorama ne laisse que très partiellement conter. Une fois en bas, les visiteurs devront en général se contenter de parler de l'expérience vertigineuse en tant que telle plutôt que de son contenu.

Toujours à l'exposition universelle de Genève, mais sur la terre ferme cette fois-ci, comme un écho à ce premier vertigineux panorama, un autre type de spectacle était proposé dans un pavillon du parc de Plaisance. Il s'agissait d'une grande maquette de près de 30 m2, représentant Genève en 1850 et intitulée « Grand Relief du Vieux Genève ». Malgré les grandes dimensions de ce modèle en trois dimensions, l'ancienne cité était reconstruite avec une grande minutie. Les visiteurs qui surplombaient la maquette étaient de nouveau placés dans une situation de vue panoramique, mais cette fois-ci face la reconstitution artificielle d'un paysage disparu.

L'auteur de la maquette, Auguste Magnin (1841-1903), avait passé près de 18 ans à minutieusement construire ce modèle dans les combles de son atelier d'architecte. L'ensemble pesait près 670 kilos. Pour ce projet, il avait d'abord réuni une grande documentation, dessiné des relevés sur le terrain, décalqué des cartes et des vues de la ville et enfin dressé des plans définitifs de chaque groupe de maison. Magnin avait effectué des premiers modèles en carton pour évaluer la complexité de la tâche et dimensionner son projet, puis avait progressivement convergé vers une combinaison de procédés pour construire la maquette. Elle reposait sur une structure de bois soutenant un assemblage de 120 caissons, puis d'une modélisation du terrain grâce à des planchettes d'un centimètre d'épaisseur. Façades et murs

étaient bâtis en tôles de zinc soudées. Les tuiles des toits et les pavés des rues étaient imités par galvanoplastie. Par la gravure à l'acide, il avait pu reproduire le pavement des murs des fortifications. Les 1500 arbres de la maquette étaient en fonte, tous différents. Des variations dans la disposition des branches permettent de suggérer la forme d'un platane, d'un peuplier ou d'un sapin. Magnin avait choisi de garder le zinc et le cuivre dans leur couleur naturelle, privilégiant avant tout l'esthétique et le rendu global de la maquette plutôt que de poursuivre une impossible tentative de réalisme.

Le processus qui conduit de la recherche et la synthèse des informations disponibles jusqu'à la reproduction sous la forme d'un objet tridimensionnel efficace comporte donc de multiples choix. À partir des archives, des relevés de terrains et d'ancienne carte, Magnin avait du reconstruire la structure sous-jacente de la Genève de 1850, comprendre la grammaire architecturale et urbaine et choisir ensuite les procédés techniques pour recréer ce qui finalement une simulation de la ville disparue. Pour accroitre la lisibilité du modèle recrée, Magnin n'hésita pas à travailler sur trois échelles différentes : 1/250e pour le plan, le 1/200e pour l'élévation et le 1/100 pour le terrain. Si Magnin avait gardé une échelle constante, la ville se serait trouvée écrasée. En effectuant une distorsion généralisée des hauteurs, la maquette de Magnin gagnait en efficacité visuelle, rendait mieux compte des détails de la ville, même si par principe elle perdait en fidélité et cessait d'être un modèle absolument réaliste de la ville ancienne._

Magnin était précisément motivé pour s'atteler à cette œuvre de long allène par les grands bouleversements urbanistiques qui étaient en train d'affecter Genève dans cette seconde moitié du XIXe. Genève se voyait dotée d'une structuration nouvelle. Une nouvelle grammaire urbaine allait recouvrir la Cité ancienne. Magnin choisit la date symbolique de 1850, car elle correspond aux premiers coups de pioche portés aux fortifications de la ville huguenote, le début de la fin d'un monde. Pour Magnin, il ne s'agissait pas tant d'en enregistrer les traces que de la redonner à voir, de la manière la plus impressionnante et la plus pédagogique et utilisant la technologie de son temps. Le panorama de l'ancienne Genève devait être aussi spectaculaire que celui de la Cité nouvelle vue du ciel. Travail d'un seul homme qui luttait pour faire

vivre le passé dans un monde en pleine mutation, le relief Magnin porte en son sein les ambiguïtés du travail de reconstruction historique à fine granularité, une question qui est, comme nous allons le voir, éminemment contemporaine. Le relief Magnin n'était pas une reproduction, mais une réinvention informée du passé, une tentative pour rendre le passé présent.

3. L'interface planétaire

En moins de dix ans, notre rapport à l'espace s'est considérablement modifié. En quelques clics nous pouvons aujourd'hui depuis une vue distante de la Terre, assez semblable aux premières images prises par les astronautes de la mission Apollo, zoomer et voir à quoi ressemble une région comme si nous étions un oiseau en train de la survoler. Souvent nous pouvons également observer à quoi ressemble un bâtiment depuis une rue avoisinante et littéralement nous déplacer comme si nous étions sur place. Il ne fallut que quelques années, à l'espace urbain planétaire pour devenir un espace algorithmique, cartographié, photographié, articulé pour être explorable par l'intermédiaire des interfaces de nos ordinateurs. Le globe terrestre n'a pas été seulement mis en image, il a été machinisé.

Les conséquences profondes de ces fantastiques dispositifs de vision sont encore à étudier. Mais une remarque s'impose d'emblée : le temps semble singulièrement invisible de ce dispositif. Est-ce vraiment le présent que me donne à voir le globe algorithmique ? De quand datent ces images aériennes que me montre l'interface ? Ont-elles été prises au même moment que celles qui ont servi à construire la navigation immersive qui me permet maintenant de voir le même bâtiment de profil ? Je regarde cette image prise d'une ville à vol d'oiseau, puis je plonge dans une de ces rues sans savoir si cette nouvelle prise de vue a été capturée la même année que la vue aérienne. En fait, en y réfléchissant, je sais bien que ces deux manières de documenter et représenter l'espace ne sont pas synchronisées. Je n'ai aucune idée de quand datent les images prises par le satellite ou l'avion qui permettent de voir la ville par dessus ni quand la voiture qui collecte les images des façades des bâtiments est passée dans ces rues. Nous passons d'un dispositif de

vision à l'autre, en toute continuité, comme si nous vivions dans un utopique perpétuel présent.

Ce travestissement du temps semble être une caractéristique propre des systèmes informationnels qui se sont développés dans la première décennie du XXIe siècle. Le déluge informationnel dont nous faisons depuis quelques années l'expérience a conduit chaque acteur a parer au plus pressé : organiser l'information du monde, la catégoriser, l'articuler pour former de grands systèmes cartographiques et de grandes bases de données capables de mettre de l'ordre dans le chaos. À la possibilité de tracer en temps réel le déplacement des hommes et des marchandises, les échanges de messages, les transactions financières, le développement des systèmes de captation audiovisuelle généralisée de la vidéosurveillance à l'imagerie satellitaire, s'est ajouté l'opportunité pour chaque possesseur de dispositifs informatiques portables de documenter sa propre vie comme jamais cela ne fut possible auparavant. Chacune de ces informations captées est potentiellement indexée dans le temps et l'espace, s'intégrant dans un système de coordonnées en quatre dimensions, mais toujours selon une logique de stabilité de nos modes de représentation. Pourtant, rien n'est moins stable dans le temps que les modes d'organisation de l'information. Pour s'en convaincre, il suffit de visiter une archive très ancienne.

4. Le Google du moyen-âge

Ma première visite aux archives d'état de Venise fut un véritable choc. Je n'aurai jamais deviné que derrière cette modeste porte du *Campo dei Frari* pouvait se cacher un dédale de salles et de couloirs abritant près 80 kilomètres de documents anciens. Sur les étagères, pas des livres précieux, mais des millions de documents administratifs, compte rendu politique, déclarations d'impôts, contrats de travail, testaments, rapport de procès, plans cadastraux. La République de Venise a très tôt adopté une politique visant à garder trace des mouvements et transactions de toute nature, accumulant un trésor informationnel sans équivalent. Venise était le Google du moyen-âge.

Parcourant pour la première fois ces interminables couloirs, je m'imaginais comme un archéologue qui aurait dans plusieurs siècles découvert les data center de notre époque, et qui réaliserait que ces racks d'ordinateurs contiennent une densité informationnelle suffisante pour refaire naitre la vie quotidienne de notre civilisation. Les documents des archives vénitiennes couvrent mille ans d'histoire. Ils ont gardé la trace des trajectoires biographiques de millions d'individus, de chaque transformation urbaine, de chaque épisode politique. Il ne fait pas de doute qu'une science historique d'un nouveau genre pourrait naître d'une exploitation systématique de ces "Big data du passé".

Pour cela il faudrait que ces informations aujourd'hui consignées des millions dans de documents manuscrits, écrits en dialecte vénitien, en latin ou en toscan puisse être accessible par informatique, cherchables, indexables, visualisables comme les autres grandes bases de données auxquelles l'Internet nous donne accès. Numériser les documents ne serait pas suffisant, il faudrait également que les machines réussissent d'une certaine manière à les transcrire, puis à en extraire des informations structurées, indexant par exemple tous les documents parlant d'une certaine personne ou d'un certain lieu et identifiant les liens qui unissent ces "entités" les unes avec avec autres. Il faudrait également pouvoir replacer ces informations dans le temps et l'espace, recréer des cartes représentant la situation non seulement de la ville, mais aussi de l'ensemble du bassin Méditerranéen au fil des années, utilisant l'énorme quantité d'information présente dans les documents des archives pour reconstituer le contexte nécessaire à l'interprétation de ces documents.

Si nous étions effectivement face à des "Big data du passé", pourrait-on imaginer de prolonger les services les plus populaires de l'Internet pour leur redonner ce qui leur manque cruellement : la durée, le temps long ? Pourrait-on imaginer un "Google map du passé" équipé d'un "slider" nous permettrait de voir le même lieu 50, 100 ou 1000 ans plus tôt ? Pourrait-on imaginer de reconstituer un "Facebook du passé", reproduisant les liens qui unissent des millions de personnes au moyen-âge, chroniquant leur vie avec une densité équivalente à celle qui caractérise nos récits de vie aujourd'hui. Peut-être que construire une machine à remonter le temps

aujourd'hui voudrait dire exactement cela : rendre le passé aussi présent que le présent, l'intégrer dans le système d'information global.

5. Le champignon informationnel

Une manière de raisonner sur ce rêve peut-être impossible est d'envisager la quantité d'information numérique que nous avons sur chaque époque comme un champignon. Si nous plaçons verticalement le temps et horizontalement la quantité d'information disponible, le déluge informationnel des dix dernières années se visualise comme un grand plateau horizontal posé sur une base qui ne cesse de se rétrécir au fur et à mesure que nous descendons dans le passé. Pour construire un Google map ou un Facebook du passé, il nous faut d'une certaine manière transformer ce champignon en un rectangle, élargir sa base pour la rendre comparable à la taille du tableau, obtenir d'une manière ou d'une autre une densité informationnelle aussi importante pour le passé que pour le présent.

Une première approche consiste à conduire de vastes projets de numérisation, puis d'extraction systématique d'information dans ces collections numérisées. Le projet Google books a ainsi numérisé près de 30 millions de livres et transcrit avec des logiciels de lecture automatique une grande partie d'entre eux, les rendant indexables et cherchables. Cette base, même si elle est constituée d'éléments extrêmement disparates (quoi de plus différent d'un livre qu'un autre livre) constitue néanmoins une source très riche d'information. Les grands projets de numérisation de la presse vont dans le même sens, permettant d'obtenir pour chaque jour une information détaillée sur les événements locaux et internationaux, mais également les cours de la bourse, les horaires de trains, les petites annonces, etc. La numérisation des données administratives donne des informations structurées de manière systématique documentant les naissances, les décès, les mariages, les testaments, les changements de propriétés ou les modifications cadastrales. Les archives privées, extrêmement nombreuses viennent s'ajouter à cette déjà très riche collection d'information avec des photographies et des lettres.

La logique archivistique et documentaire mise en place au début du XIXe siècle en France ou en Italie par exemple a encore de nombreux points communs avec celle qui est à l'oeuvre aujourd'hui. Pour cette raison, il ne fait pas de doute que sur les deux cents ans dernières années en croissant par exemple les informations extraites des articles de presse et des administrations, un tableau relativement précis de la société d'un pays et de son évolution au jour le jour puisse être reconstituée. Évidemment, plus nous reculons dans le passé, plus le nombre de documents se réduit et surtout plus les logiques de représentation ne deviennent étrangères. Les structures documentaires mises en place à la Renaissance, dans la lignée de la diffusion de l'imprimerie demande un travail spécifique pour être interprétées selon les logiques des systèmes informationnels contemporains. La situation documentaire au moyen-âge et dans les périodes plus anciennes est encore plus critique.

Pour compenser le manque d'information archivistique ou pour compléter les espaces non couverts par les archives, il nous faut adopter une stratégie complémentaire en reconstituant les données manquantes par extrapolation et généralisation. Si un historien retrouve le carnet de bord d'un capitaine de navire du XVIe siècle documentant précisément un voyage entre Venise et Corfou, il ne se contentera pas seulement d'en tirer des conclusions sur ce voyage particulier, mais pourra, sous certaines conditions, utiliser ce document pour déduire des informations sur les modes de navigation et de vie maritime de cette époque. De la même manière, nous pouvons nous servir des informations architecturales d'un palazzo vénitien de style gothique encore très bien conservé pour faire des hypothèses sur la structure et le style d'un bâtiment de la même époque et ayant des fonctions similaires, mais si ce dernier est aujourd'hui complètement détruit. Il s'agit toujours au-delà de l'information contenue dans un document particulier d'extraire des structures, des grammaires et de s'en servir pour construire des hypothèses motivées nous permettant de compléter les blancs laissés entre les archives. En informatique, ces techniques correspondent à la grande famille des méthodes de simulation, étudiées depuis plus de 50 ans. La problématique de la découverte de structures dans des données bruitées, de la généralisation à partir d'exemple, de l'extrapolation suivant certaines hypothèses de travail constitue la

base de toute une science qui s'est développée en parallèle des sciences historiques. Les "Big data du passé" annoncent peut-être une interface féconde et nouvelle entre ces domaines et invitent en tout cas à réfléchir au statut épistémologique singulier de ces "passés reconstitués".

La simulation ne saurait être comprise comme une manière de simplement compenser le manque de données historiques, utilisées seulement quand les données archivistiques manquent. En fait il n'y a jamais assez de données. Aucun plan cadastral, aucune photo, aucun relevé laser ne pourraient me permettre de reconstituer précisément la structure d'une seule *calle* de Venise. Même dans des situations d'hyperdocumentation, il faut à un moment ou un autre prolonger les données selon certaines hypothèses. Simulations et données vont toujours de pair. Seules la résolution et l'incertitude changent.

6. Un bien commun scientifique

Encore aujourd'hui la plupart des historiens ont l'habitude de travailler en toutes petites équipes, se focalisant sur des problématiques très spécifiques. Ils n'échangent que très rarement leurs notes ou leurs données, percevant à tort ou à raison que leurs travaux de recherche préparatoire sont à la base de l'originalité de leurs travaux futurs. Prendre conscience de la dimension et la densité informationnelle des archives comme celle de Venise doit nous faire réaliser de l'impossibilité pour quelques historiens, travaillant de manière non coordonnée de couvrir avec une quelconque systématicité un objet aussi vaste. Si nous voulons tenter de transformer une archive de 80 kilomètres couvrant mille ans d'histoire en un système d'information structuré il nous faut développer un programme scientifique collaboratif, coordonné et massif. Nous sommes devant une entité informationnelle trop grande. Seule une collaboration scientifique internationale peut tenter d'en venir à bout.

Ce n'est pas la première fois que des disciplines se trouvent devant à une difficulté de ce type. La transformation des archives de Venise en système d'information ressemble aux défis que les chercheurs en rencontré quand ils ont tenté de

modéliser des systèmes très vastes comme le système génomique, le système cérébral ou le système planétaire. Pour constituer les bases de données nécessaires à l'étude du Génome, le Cerveau ou la Terre, les chercheurs de disciplines pourtant très compétitives ont réussi à mettre en place des programmes internationaux coordonnés. Ces programmes ont permis la création d'un bien commun scientifique, utilisable librement par tous les chercheurs du monde et jouant un rôle de pivot pour l'élaboration d'une connaissance construite collectivement.

Nous avons travaillé avec la même logique pour le projet vénitien. Le projet s'est d'abord articulé autour de deux partenaires principaux l'École Polytechnique Fédérale de Lausanne et l'Université Ca'Foscari à Venise. L'ambition pour ces deux partenaires initiaux était de mettre en place les structures et la méthodologie de travail pour constituer la plus grande base de données jamais créée sur Venise et son empire en se basant en premier lieu sur un programme de numérisation massive des archives des la cité des doges. Pourrait ensuite se joindre un nombre croissant d'équipes de recherche travaillant sur Venise venant du monde entier. Chacun des travaux de recherche de ces équipes pourrait bénéficier de l'information produite par les autres et à leur tour enrichir la base.

Pour mettre en place un régime collaboratif permettant la création d'un bien commun scientifique il faut relever un certain nombre de défis. Il a fallu d'abord convaincre les archives que leur mission de conservation et de valorisation d'un patrimoine millénaire pouvait prendre la forme d'une libération sous forme numérique des images des documents dont elles assurent la sécurité et la transmission depuis tant d'années. Une image n'est pas le document lui-même, mais leur mise à disposition sur l'Internet peut dispenser les chercheurs de se rendre physiquement à l'archive, ce qui peut tendre à marginaliser le rôle de celle-ci comme "temple du savoir". Les archivistes ont longtemps été les médiateurs indispensables pour accéder aux "Big data du passé". La numérisation massive aurait pu leur donner l'impression de se voir placer dans un rôle tout d'un coup plus périphérique.

C'est pour cette raison que nous avons choisi non pas de mettre en place un service de numérisation "clé en main" à l'archive, dans lequel une équipe de "spécialistes"

serait venue traiter les documents comme a pu le faire Google dans son projet Google Books, mais au contraire d'aider les archivistes eux-mêmes à devenir les experts de cette numérisation. C'est eux qui depuis les premiers mois pilotent le projet, choisissent les équipes et les machines, organisent les schémas de travail. Eux seuls connaissent intimement l'archive, eux seuls peuvent comprendre comment adapter au contexte très spécifique de leur lieu, certaines des innovations majeures de la numérisation.

Nous avons convaincu les archivistes de Venise qu'en abordant la numérisation de cette manière, ils seraient extrêmement positifs qu'ils libèrent les images des documents numérisés en accès ouvert, permettant à chacun non seulement des les consulter, mais aussi de les télécharger pour leurs travaux.

7. Transcrire des millions de documents

La numérisation n'est qu'une première étape. Il faut ensuite indexer le contenu des images numérisées, apprendre aux machines à les lire. Plusieurs années sont nécessaires à un philologue pour pouvoir déchiffrer la variété des écritures que l'on peut rencontrer dans les documents des archives vénitiennes. Comment une machine pourrait-elle y arriver ?

Par rapport à un tel défi, l'énorme quantité de documents que nous devons transcrire n'est pas un problème, c'est au contraire une partie de la solution. La grande quantité des documents nous permet en effet de repérer des motifs graphiques récurrents (mots, abréviation, lettrine, etc.) Un algorithme segmente les documents en repérant ces motifs, puis calcule une distance graphique entre chaque motif. Le même mot écrit par le même scripteur a une apparence similaire que l'algorithme arrive évaluer. La distance augmente quand on change de scripteur ou de mots. Par ce procédé, la masse documentaire devient un réseau de motifs graphiques reliés les uns avec les autres, avec pour chaque lien une estimation de la probabilité qu'il s'agisse du même motif. Ainsi si nous donnons la possibilité à un lecteur de transcrire un mot, proposant ainsi une correspondance directe entre une forme graphique et une séquence de lettre, une série d'inférence permet ensuite de

déduire des transcriptions possibles pour toute une série d'autres motifs présents sur les documents. Chaque transcription est associée à un degré de confiance ou d'incertitude.

À cette logique de correspondances graphiques se superposent les connaissances sur la langue et la structure des documents dont nous pouvons disposer pour une époque particulière sous la forme de corpus de textes déjà transcrits. Les philologues et les historiens travaillent depuis longtemps pour établir des éditions critiques des textes et séries documentaires les plus importantes des archives. Nous pouvons extraire de ces corpus de textes datant d'époques différentes, des modèles statistiques. Plus précisément, nous pouvons construire, à partir de grande quantité de textes transcrits dont nous disposons, un modèle calculant les probabilités de transitions entre les mots pour une année spécifique du passé. Ce modèle peut-être utilisé pour améliorer les inférences lors du processus de transcription semi-automatique des documents. Si graphiquement deux transcriptions peuvent correspondre à un même motif visuel, une d'entre elles est peut-être beaucoup plus probable si l'on se base sur les statistiques des textes de l'époque correspond.

Ce jeu d'inférences statistiques transforme en profondeur les méthodes de transcription. L'objectif n'est plus forcement d'obtenir une transcription parfaite, fruit d'un travail parfois de plusieurs années, mais d'améliorer constamment la qualité des transcriptions partiellement automatisées en transcrivant spécifiquement certains documents ou passage. Les algorithmes eux-mêmes peuvent donner des indications sur les passages qu'il serait le plus utile de transcrire pour améliorer globalement la qualité de la transcription générale.

8. Construire le graphe sémantique

À partir des transcriptions complètes ou partielles, il possible de repérer ce que nous appelons les "entités nommées". Il s'agit typiquement des personnes, des lieux et des institutions. Un double processus permet de le faire. Nous pouvons d'une part repérer certains noms à partir de liste de lieux ou de personnes déjà établis et d'autre part inférer selon la syntaxe la présence d'entités nommées nouvelles. Dans

un contrat de travail ou un testament par exemple, le nom de l'apprenti ou du défunt est en général introduit à un moment particulier. Le repérage le plus efficace de ces éléments variables du texte se fait durant le processus de transcription, où un lecteur peut spécifier qu'un segment graphique particulier correspond à un nom ou un lieu. Le système d'inférence peut ensuite tenter de prédire, selon la structure des documents quand ces éléments sont susceptibles d'être présents.

Les entités nommées participent à ce que nous appelons des entités temporelles, typiquement des évènements. Pour encoder des informations historiques, il est en général préférable de modéliser les changements plutôt que les états. Par exemple plutôt que noter dans une base de données la date de naissance d'une personne, il est plus astucieux de créer une entité temporelle correspondant à l'évènement de sa naissance. Nous pouvons en fait dans ce cas plus facilement associer à cette entité un lieu, la présence d'autres personnes, etc.

Les entités nommées et les entités temporelles constituent les sommets d'un graphe immense dont les arêtes précisent les relations particulières que ces entités entretiennent les unes avec les autres. La structure de ce graphe a des similarités avec celle qui code l'information dans les réseaux sociaux. D'une certaine manière, nous essayons de construire ici, un "Facebook" du passé. Dans ce graphe nous pouvons par exemple facilement indiquer qu'une personne X a participé à un évènement Y ayant lieu dans un lieu Z durant un intervalle temporel T. Ce codage sous forme de graphe nous permet plusieurs types d'inférence. Si deux personnes ont participé à un même évènement, cela implique qu'elles ont été ensemble dans le même espace pendant le même intervalle temporel. Cela suppose également que ces personnes doivent être nées avant cet évènement et pas encore mortes. Ces contraintes qui semblent évidentes lorsque l'on code une information historique prennent toutes leurs intérêts quand des millions d'informations de ce type sont extraites de grandes séries de documents. Chaque nouvelle information doit pouvoir s'inscrire dans l'espace délimité par les autres informations pour former un ensemble cohérent. Si des incohérences apparaissent, il devient nécessaire de considérer qu'il existe différentes reconstructions possibles du passé, incompatibles les unes avec les autres.

Le codage des lieux pose des problèmes spécifiques. La notion de lieu est en effet un concept relativement complexe. Coder un lieu par ces coordonnées GPS, une idée qui semble naturelle aujourd'hui, se révèle être une stratégie très mal adaptée dans un contexte historique. Un pays comme la France ou une ville comme Athènes peuvent être considérés comme des lieux correspondant à une géométrie spatiale relativement stable. Mais il existe d'autres types de lieux qui bougent dans le temps. Certains évènements peuvent avoir lieu sur bateau. D'autres dans un bar dont l'adresse pourra changer plusieurs fois, mais qui conservera une même identité dans le temps. La notion de lieu n'est donc pas a priori une notion géométrique. Les lieux se définissent toujours relativement à d'autres lieux plus larges et sont toujours potentiellement en mouvement par rapport à ces référentiels. Ce type de précaution est crucial lorsque l'on conçoit les systèmes d'information qui doivent accueillir les données historiques.

9. Encoder les informations métahistoriques

Comment passer des documents transcrits à ce type de codage sémantique ? Le travail peut être fait manuellement par des historiens qui interprètent les documents et encode les informations dans une base de données. Dans certains cas, le processus peut être en partie automatisé lorsque les documents suivent une logique très régulière, par exemple pour certains documents administratifs comme des contrats qui spécifient toujours de la même manière les relations qui lient un certain nombre d'acteurs. Un contrat d'apprentissage vénitien comportera ainsi le nom de l'apprenti, le nom du maître, l'adresse de son atelier, le salaire qui sera versé.

Quelle que soit la méthode, il est extrêmement important que les processus techniques et historiques qui conduisent à construction de l'information soient documentés avec précision . En effet, ces informations métahistoriques sont nécessaires pour comprendre la nature de l'information encodée dans le graphe sémantique. Elles peuvent expliquer des éventuelles incohérences dans les données encodées. Elles permettent d'envisager avec précaution la fusion de jeux de données venant de groupes de recherche différents, utilisant des méthodes variées. Elles

ouvrent la voie à la reconstruction de passé possible en ne considérant que certains types de documents ou de méthodes.

Les informations métahistoriques peuvent se documenter sous la forme d'un graphe similaire à celui qui sert à coder l'information historique. Des acteurs (les chercheurs) participent à des événements (une transcription, la vectorisation d'une carte) qui ont lieu dans l'espace et le temps et conduisent à la production de certains objets (cartes, images, transcription, etc.). Selon les interfaces utilisées, elles peuvent être documentées automatiquement sans faire perdre du temps au chercheur.

Ainsi chaque étape du processus de construction de l'information historique peut être documentée. Il devient possible d'expliciter le processus de sélection des sources, les phases de transcriptions et les différents processus interprétatifs, qu'ils soient réalisés par des humains ou des machines. Cette approche ouvre la voie à des systèmes informatiques capables de gérer de situations dans lesquelles il n'y a pas qu'une seule vérité historique, mais diverses reconstructions possibles, correspondant à des espaces de connaissances particuliers, tous parfaitement documentés.

10. Le déploiement spatiotemporel

Le graphe sémantique peut ensuite être déployé dans le temps et l'espace, mais pour cela il faut reconstruire des cartes du passé. Il existe pour une ville comme Venise de très nombreuses cartes détaillant avec précision les évolutions urbaines. Les relevés cadastraux des 200 dernières années peuvent être facilement adaptés pour intégrer un système d'information géographique historique. Chaque relevé cadastral donne une "photographie" à une date précise de la structure urbaine de la ville. Une des difficultés est de compléter l'espace temporel qui sépare deux de ces "instantanés". Un bâtiment peut, par exemple, être présent dans le cadastre napoléonien du 1805, mais pas sur le cadastre autrichien, environ trente ans plus tard. En l'absence d'autres informations, il faut faire une inférence probabiliste si nous souhaitons représenter la situation dans une année intermédiaire. Plus nous serons proches de

la date du cadastre napoléonien, plus il y aura de chance que le bâtiment soit encore présent. Dans tous les cas, il nous faut à nouveau accepter l'idée qu'il n'y a pas une seule bonne carte, mais plusieurs plus ou moins probables.

Pour exploiter les cartes du XVIIIe, XVIIe et du XVIe siècle nous devons procéder différemment. En effet la plupart de ces cartes offrent des vues en perspective de la ville, très détaillées et riches sur le plan architectural, mais pas directement superposable aux cartes actuelles. L'information sur ces cartes est avant tout topologique. Nous pouvons déduire quel palazzo jouxte quel campanile. Nous pouvons croiser cette information avec d'autres sources, comme les déclarations d'impôts qui elles aussi donnent des informations topologiques, comme la description de la position d'une boutique par rapport à la proximité d'un palais. En combinant ces sources, un jeu de contraintes se met en place, permettant d'élimer certaines solutions, mais laissant de la place à diverses configurations. Il s'agit en quelque sorte de résoudre un Sudoku géant en partant de certaines hypothèses et en évaluant si elles sont compatibles avec le reste des données disponibles.

Faute de cartes détaillées, nous devons pour reconstruire Venise avant 1500 en nous basant seulement sur des données archéologiques et des témoignages indirects. L'exercice est plus périlleux, la place des hypothèses de travail étant extrêmement importante. Néanmoins, dans la mesure où ces choix sont parfaitement documentés, il est envisageable de simuler les configurations urbaines qui font décrivent l'évolution progressive de la ville depuis les premiers ilots terraformés jusqu'aux cartes détaillées de 1500. Pour reconstruire ainsi la « morphogenèse » de la ville, l'historien doit extrapoler à partir des données existantes, extraire les régularités — des grammaires —, qui gouvernent les formes connues pour simuler leur application dans un nouveau cas non directement documenté. Ces grammaires ont certaines similarités avec celles que nous mentionnions pour la transcription des textes, mais elles sont ici utilisées non pas pour aider à la reconnaissance, mais pour générer des configurations urbaines possibles.

11. Les approches procédurales

En informatique, ce type d'approche appartient à la grande famille des techniques d'extrapolation, et plus particulièrement à ce qu'on appelle les approches génératives ou procédurales. Une grammaire générative permet par exemple de tracer un réseau de rues ou de visualiser des façades de bâtiments qui n'existent plus. Ces «architectures» procédurales peuvent cohabiter avec des modélisations plus traditionnelles comme dans le cas du projet Rome Reborn, une reconstruction de la Rome antique en 320 AD. Le modèle de la ville utilisé dans ce projet combine aussi bien des éléments de classe I, dont la position, l'identification et l'apparence sont connues avec une assez grande précision, et des éléments de classe II pour lesquels nous ne disposons que d'informations indirectes et imprécises. Seuls 250 bâtiments parmi les 7000 bâtiments que compte la reconstitution appartiennent à la première catégorie. Tous les autres sont construits selon des hypothèses architecturales et historiques formalisées dans un algorithme.

Les approches procédurales qui encodent les hypothèses historiques dans des algorithmes ne se limitent aucunement au domaine des reconstitutions architecturales. Pour s'en convaincre, prenons une autre famille d'exemples, cette fois-ci centrés sur la modélisation des routes maritimes historiques. L'étude des arts nautiques, les traités de navigation, l'étude des navires et bien sûr les recherches sur la circulation des biens ou des informations ont donné lieu à un grand nombre de monographies. En appliquant ces recherches aux routes maritimes vénitiennes et en les associant aux très riches informations disponibles dans les Archives d'État, une reconstitution fine des échanges en Mer Méditerranée est envisageable.

Certaines routes sont particulièrement bien documentées, notamment le système des galées du marché voyageant chaque année en convoi entre la fin du 13e siècle et le milieu du 15e siècle. Sur ces trajets, il est possible à partir des données disponibles de tracer les itinéraires de chaque convoi, année après année. À partir de l'encodage de ces informations, nous pouvons reconstruire un véritable simulateur de trajets qui nous permet de répondre à des questions comme : En juin 1322, quel est le prochain bateau de Constantinople à Corfou ? Combien de temps prendra le voyage ? Quelles sont les chances de rencontrer des pirates ?

Mais malgré leurs détails, les données des archives ne descendent pas en dessous d'une certaine précision dans la description des itinéraires. Par exemple, seules les escales commerciales obligatoires sont notées. Or il est extrêmement probable que les navires fassent des escales intermédiaires notamment en Mer Adriatique ou lors de recherches d'informations (présence de dangers, piraterie…). Si l'on souhaite aller à ce niveau de détails pour simuler, mois par mois, la position probable des navires, plusieurs hypothèses peuvent être proposées.

Ces hypothèses se basent sur les connaissances recueillies sur la navigation de l'époque, mais, comme dans tout processus de recherche académique, elles peuvent différer les unes avec les autres, notamment dans la part relative qu'elles accordent au cabotage ou à la navigation en haute mer. Ici, comme en architecture, il peut y avoir débat. Et dans le nouveau régime visuel des cartes géohistoriques, ce débat prend la forme d'algorithmes contradictoires qui encodent des trajets différents.

Ces exemples nous illustrent comment la cartographie en étendant le domaine de représentation de l'histoire effectue en même temps un déplacement du débat scientifique. Qu'il s'agisse de tracés de rue, de réseaux d'information, de modélisation architecturale, de flux commerciaux ou de trajets maritimes, les chercheurs ne peuvent plus se contenter de débattre linguistiquement avec des argumentaires et des raisonnements, mais en opposant dans le domaine cartographique leurs hypothèses appareillées sous la forme d'algorithmes.

12. Une éthique de la représentation

La force des systèmes d'information historique est d'uniformiser les données produites par des méthodes différentes dans un seul modèle informationnel capable de reconstruire à la volée des cartes plausibles pour n'importe quelle date des 1000 dernières années. Il faut néanmoins prendre des précautions quant à la manière de représenter ces cartes hypothétiques.

La représentation sous forme de cartes, en faisant basculer une représentation essentiellement textuelle et narrative en une représentation visuelle et exploratoire,

introduit un mouvement visant à expliciter chaque étape des processus de constitution des données historiques sous-jacentes aux représentations. Pour ne pas tromper celui qui lit les représentations visuelles et synthétiques d'un passé reconstitué, il faut développer des outils et des processus garantissant une « éthique de la représentation ». Ces développements impliquent la constitution de nouvelles normes visuelles.

Les lecteurs de cartes attendent peut-être implicitement un niveau de précision éventuellement incompatible avec le niveau de certitude que les données peuvent raisonnablement fournir. Alors qu'un nombre toujours plus grand de projets s'attelle à reconstruire des cartes du passé, la question de la représentation de l'incertitude a pris à grande importance dans la communauté des chercheurs qui travaillent sur ces questions. Nous avons discuté plus haute qu'une manière d'approcher ce problème est d'associer à chaque processus intellectuel et technique qui permet la construction des cartes ou d'autres formes de représentation visuelle à des données documentant la manière dont sont constituées les données historiques représentées. D'un point de vue éthique, plus les reconstitutions du passé sont susceptibles d'impressionner, plus rigoureux et transparent devrait être le processus intellectuel et technique sous-jacent.

Différentes méthodes ont été proposées pour visualiser l'incertitude de données. Il est possible d'utiliser par exemple différents niveaux d'opacité pour signifier le degré d'incertitude présente sur les cartes historiques. La même approche peut être utilisé dans le cadre de représentations architecturales urbaines (opacité très dense quand le bâtiment est encore présent, dense quand le bâtiment existe encore, mais a subi des rénovations importantes, moyen quand il n'existe plus qu'une ruine, très peu dense quand la seule source de présence de ce bâtiment est indirecte, indiqué par exemple par une position sur un plan). Ces informations permettent de qualifier la nature de la reconstitution 3D résultante. Elles sont une première indication qui peut être complétée par un accès aux données métahistoriques documentant la chaine de processus sous-jacents aux données représentées.

13. La rencontre entre des dizaines de milliers d'étudiants et des millions de documents

L'ensemble des facteurs que nous venons d'analyser nous laisse à penser que 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. Il nous faut former une nouvelle génération d'étudiants capables de maîtriser les algorithmes comme outils de connaissance et de débats, mais aussi de s'interroger sur ces changements. En rapprochant recherche et éducation, cette nouvelle génération d'étudiants pourrait non seulement bénéficier de ces projets de grande ampleur, mais jouer un rôle crucial dans leur succès ou leur échec. Il s'agit "simplement" d'organiser, chaque année, la rencontre entre des dizaines de milliers d'étudiants et des millions de documents.

Pour tenter d'illustrer comment cette convergence entre éducation et recherche pourrait avoir lieu, prenons un exemple. J'ai décrit plus haut les techniques générales qui permettent d'envisager la transcription semi-automatique de documents anciens. Nous sommes en train de préparer un cours en ligne sur ce sujet particulier. Le cours présente à la fois les bases en terme de philologie pour s'initier à la lecture des écritures et l'interprétation des documents anciens et introduit les techniques générales pour automatiser en partie leur transcription. Comme dans tous processus partiellement automatisés, ces techniques gagnent beaucoup en efficacité quand elles sont adaptées aux caractéristiques spécifiques de chaque famille de document. Si l'algorithme peut bénéficier d'information sur la langue, la structure des documents, les caractéristiques des écritures, ses performances seront décuplées. Or appliquer des connaissances générales apprises en cours à des cas spécifiques est la base d'une démarche pédagogique efficace. C'est ce que souhaitons proposer aux étudiants de ce cours.

Depuis quelques années les cours en ligne massifs (MOOCs) ont gagné en popularité. Certains cours de l'École Polytechnique Fédérale de Lausanne sont suivis, chaque semestre par des dizaines de milliers d'étudiants dans le monde entier. Ces cours

sont exigeants et demandent un travail régulier de la part des étudiants. Sur 50 000 étudiants inscrits la première semaine, seulement 20 % arrivent en général au bout du cours. Mais, même avec un taux d'abandon à 80 %, cela représente encore 10 000 étudiants motivés.

Imaginons pour simplifier que les documents de notre archive sur lesquels nous souhaitons que les étudiants travaillent s'étalent sur un millier d'années. Découpons cet immense corpus en décade regroupant des documents ayant une certaine homogénéité formelle et linguistique. Nous pouvons former ainsi 100 groupes de documents homogènes. Avec 10 000 étudiants nous pouvons encore associer à chaque 100 étudiants à chaque catégorie de document. Faisons les travailler en groupe de 5, constitué de profils complémentaires (historiens, informaticiens, linguistes, etc.) et demandons à chacun des 20 groupes ainsi constitués d'adapter les principes généraux vu en cours aux spécifiés de la famille de documents dont ils ont la charge. Tous les étudiants apprendront par une telle démarche et il est très probable que, sur les 20 groupes, au moins un obtienne un résultat de très bonne qualité. En un semestre, ce sont 100 nouveaux algorithmes optimisés pour la transcription automatique de familles de documents homogènes qui sont disponibles. Les étudiants ont appris, le projet a fait un bon avant, la procédure peut être réitérée au semestre suivant.

14. Comment le travail de chacun s'enrichit du travail de tous

Il est essentiel de comprendre que ce changement d'échelle dans les sciences humaines n'exclue pas les modes de recherche plus traditionnels. Au contraire, les grandes bases de données multidimensionnelles ainsi constituées offrent un environnement qui facilite le travail sur des sujets spécifiques.

Si j'étudie un peintre du XVIe siècle, je pourrais profiter du travail d'extraction que d'autres chercheurs ont effectué sur les contrats d'apprentissage documentant ainsi l'ensemble des ateliers de la ville et tous les apprentis (déclarés) qui y travaillaient.

Je voudrais peut-être savoir les couleurs disponibles à cette époque et trouverait l'information dans le modèle des échanges maritime construit par un autre groupe de recherche à partir des archives. En me basant sur les cartes reconstituées montrant le tissu urbain et simulant certaines des façades que ce peintre a peut-être peintes, je pourrais contredire ou compléter les informations présentées dans la base géohistorique urbaine du projet. Mon travail en retour viendra densifier la base de données grâce aux informations biographiques et iconographiques nouvelles qu'il a contribué à structurer.

Dans le même ordre d'idées, la grande base de données multidimensionnelle pourrait me permettre de mieux comprendre un moment particulier dans l'évolution de la musique vénitienne au tournant de l'âge baroque. Comment par exemple expliquer le grand bouleversement qui, avec l'arrivée de la musique baroque, conduit les partitions musicales à redevenir manuscrites, phénomène peut-être unique dans l'histoire des médias. Au XVe et XVIe siècle, Venise dominait le marché des partitions imprimées. Mais les changements musicaux introduits par le baroque, à l'époque de Monteverdi, ont posé un problème à l'industrie typographique. La nouvelle musique ne pouvait plus s'imprimer sans inventer de nouveaux systèmes techniques. Cette crise se conjugue avec un déclin économique, car cette transformation apparaît au moment où la cité des doges perd du terrain en Méditerranée face aux Ottomans. Enfin la nouvelle musique cesse d'être une musique populaire pour au contraire conquérir les élites. Dans ces relations imbriquées, quelles sont les causes et quelles sont les conséquences ? Comprendre comment la musique redevient manuscrite nécessite de combiner les points sur vues artistiques, techniques, sociologiques et économiques.

Le défi de la Venice Time Machine est d'amener toutes ces recherches à s'articuler les unes par rapport aux autres, de façon à livrer une vision globale de la ville et de ses réseaux à travers 1000 ans d'histoire. Le défi n'est pas que technique. Il faut que les différents chercheurs qui contribuent cet environnement de recherche sentent que ce projet est aussi le leu, un bien commun collectivement construit.

15. Un modèle multidimensionnel de Venise comme prototype pour un programme européen plus vaste.

L'histoire de Venise, celle d'un modeste village de pêcheurs protégé par une lagune qui devient en quelques siècles le plus grands empire maritime de la Méditerranée, est fascinante à reconstituer. Comme nous l'avons vu, nous disposons de données innombrables pour reconstruire cette histoire, car le développement de Venise s'est très vite accompagné de la mise en place d'un système bureaucratique dans lequel tout était documenté. L'enjeu est de constituer une base ouverte multidimensionnelle et collaborative, en perpétuelle densification.

La première dimension de cette base est environnementale : l'évolution de la lagune, sujet de grande attention aujourd'hui, a toujours joué un rôle crucial dans l'histoire de Venise. La ville s'est construite en coévolution avec elle. La seconde évolution est urbaine et architecturale. Nous possédons des informations assez précises qui permettent de reconstruire la « morphogenèse » de la ville, construite d'ilot en ilot au fil des siècles. La troisième dimension regroupe les activités des hommes. Grâce aux très riches archives de la ville, nous pouvons reconstituer des informations démographiques assez détaillées. Nous pouvons également modéliser les circulations et les échanges, au sein de la lagune puis avec le développement de l'empire maritime sur une grande partie de la Méditerranée. Enfin, la dernière dimension concerne les productions humaines, linguistiques, culturelles et artistiques, résultant de toutes les influences issues du fantastique réseau socio-économique de Venise en Europe. Comme nous l'avons vu, toutes ces dimensions s'influencent les unes et les autres.

Venise est un exemple emblématique, mais la même approche pourrait être déployée sur d'autres archives et d'autres villes. Il ne fait pas de doute que L'Europe est la mieux placée mondialement pour être en pointe dans ces nouveaux champs de recherche: il y a ici une histoire millénaire, riche et complexe, les meilleures spécialistes de ce passé et le multilinguisme comme culture profonde. C'est une chance que nous devons saisir.