Compte rendu du colloque : « Document Engineering (DocEng) 2015 », 8–11 Septembre 2015, EPFL.

Recensé par: Vlad Atanasiu, Université de Fribourg

Imaginez que la revue entre vos mains ait été mise en pages par des algorithmes, que les sommaires des articles et les traductions vers les quatre langues nationales soient aussi dues à des machines, enfin, que l'auteur même de ces lignes soit une intelligence artificielle ayant assisté sous la forme d'un robot au symposium DocEng. Ce sont, en essence, les objectifs poursuivis dans les travaux présentés. Pour l'historien, leur intérêt tient à l'impact sur le travail de recherche historique et aux aspects sociologiques du travail des ingénieurs, sujet de recherche en soi même.

DocEng est un symposium international sous l'égide de la plus grande société d'informaticiens du monde, l'Association for Computing Machinery (ACM) ; tenue cette année à École fédérale polytechnique de Lausanne, dans des conditions optimales d'organisation scientifique et hospitalité, elle réunit depuis 2001 des universitaires et des firmes, telle Hewlett-Packard, Adobe, et Microsoft, pour des échanges sur l'automatisation et l'amélioration du traitement du « document électronique ». Ce terme peut se comprendre originellement comme la version numérique de livres et autres imprimés sur papier – e.g. fichiers Word, LaTeX, PDFs, scans –, mais s'étend de plus en plus vers d'autre contenus multimédia et appareils d'affichage et interaction.

Parmi les idées discutées, les plus déstabilisantes, car touchant de plus près à l'activité mentale propre de l'humain, étaient « la création automatique de texte à partir des idées », le résumé d'articles et la transcription de dialogues. Un des piliers traditionnels du symposium est, cependant, la production de documents, c'est à dire le packaging de l'information en vue de sa consommation par l'utilisateur de technologies numériques.

Dans les années 1970, l'informaticien américain Donald Knuth crée le logiciel de mise en page TeX, OpenSource avant la lettre, qui transforme les informaticiens en typographes apprentis et constitue aujourd'hui le standard de publication dans nombreuses sciences exactes (mathématiques, physique, etc.). Un fichier TeX est – comme HTML, son cousin du web – un fichier texte agrémenté de balises, indiquant un rôle sémantique – e.g. paragraphes, titres – et une apparence visuelle – e.g. alignement, police de caractères – ; une fois produit, le texte balisé est mis en page par le logiciel, pour être rendu à l'écran ou imprimé. TeX implémente ainsi un concept par lequel le contenu se veut séparé de sa forme, contrairement à Word, où tout changement textuel est immédiatement visible au niveau typographique.

Ce n'est là, probablement, qu'un avatar digital de l'arbitraire du signe linguistique postulé par le linguiste suisse Ferdinand de Saussure et d'une certaine philosophie grecque, prônant l'immatérialité des idées, qu'a infusée la civilisation occidentale. Cet esprit était omniprésent au symposium. D'une part, les recherches présentées sur l'extraction d'information traitaient le texte et l'image à l'exclusion de la mise-en-page, de la matière du document, et de son contexte socioculturel, des aspects sans lesquels une paléographie, une histoire de l'art, ou le graphisme sont inconcevables. D'autre part, le rejet de la consubstantialité de l'information et de la forme mettait des bornes aux capacités analytiques des algorithmes présentés par les conférenciers.

Sous forme de texte et d'images, la langue est conçue comme un monde auto-contenu, d'où il s'agit d'extraire du sens. Par emploi de modèles quantitatifs, des schéma logiques, ou des réseaux de neurones, on arrive à des performances remarquables, comme la transcription de l'écriture manuscrite moderne et ancienne, l'identification de parties du discours, la désambiguïsation des entités nommées, la mise en relation de concepts, et, nec plus ultra, la traduction automatique.

Cependant, tant que la langue ne prend pas matière en tant qu'action dans le monde physique, les « bytes » faits de uns et de zéros erront sans sens dans les entrailles électroniques de la machine. C'est hasarder une prophétie, mais les nouveaux participants aux futurs symposia DocEng seront les robots et les drones.

En les attendant, les participants à DocEng se sont interrogés sur pourquoi ils devraient prendre soin d'espacer de façon homogène les mots imprimés et éviter les rivières, ces filets verticaux fait d'espaces intermots qui surgissent, à l'improviste, d'une page mal composée. Ces soucis de typographie fine sont difficilement défendables − y compris financièrement − dans un monde de fast foods. D'autant plus que pour nombre d'ingénieurs (je sais ce dont je parle, mon père en est un), c'est de « l'esthétique » sans substance.

Hélas, il suffirait d'un psychologue dans l'audience pour clarifier qu'espacements et rivières typographiques correspondent aux concepts psychologiques de « groupement » (les éléments d'une unité sémantique doivent aussi former une unité perceptuelle) et de « cluttering » (plus le stimulus est complèxe, moins il est compréhensible), d'où leur imapact sur l'efficacité de la communication écrite. Au point où, pourrait témoigner le dyslexique absent, lui aussi, de la salle, la vision se trouble et la lecture fait place à des migraines. Le sens de cette périphrase est de faire remarquer le peu de participation à l'élaboration des technologies présentées de spécialistes des facteurs humains, du design, de la visualisation − qui font par ailleurs, épaulés par leur collègues du marketing, le succès d'un société nommée Apple. Ce n'est donc pas étonnant qu'une action aussi simple, en apparence, que feuilleter un livre reste toujours plus informative et sans prise de tête en version papier que numérique.

Le problème de l'interdisciplinarité se pose de manière encore plus aigue pour les sciences humaines. Un champs de crevasses fait d'incompréhension mutuelle sépare les ingénieurs des humanistes. L'expertise pluriséculaire des sciences humaines est souvent peu valorisée par l'informatique, qui préfère inventer ses propres méthodes à forte constituante mathématique. Alors que les ingénieurs des documents électroniques sont amenés à redécouvrir les méthodes et le savoir des humanités, ils tendent en parallèle à vouloir les substituer par des algorithmes et les pousser vers la caducité.

Pour se restreindre au premier cas de figure, ainsi qu'au travaux du DocEng, la technologie aide la lecture de palimpsestes du Mont Sinaï par une analyse multispectrale ; elle établi toute seule des liens entre les informations de 80 km manuscrits des archives de l'État de Venise, qui prenderaient une vie de moine seulement à les lire ; extrait une hiérarchie de concepts à partir de documents ; ou encore facilite la recherche – et la recherche sur l'histoire de la recherche – en représentant la toile de citations dans les publications scientifiques.

Et si on avait qu'un seul quart d'heure a dédier au symposium, on aurait pu assister à la présentation d'un logiciel qui aide les aveugles a photographier correctement des documents qu'ils ne peuvent pas voir – ce serait le souvenir que j'emporterais, d'un vrai « document engineering » pour des vrais humains.

Vlad Atanasiu
atanasiu@alum.mit.edu
http://alum.mit.edu/www/atanasiu/