Modèles tridimensionnels pour la représentation de l'état des connaissances et propositions de visualisation pour l'analyse des corpus textuels

Marie Pérès[1], Jean-Marc Leblanc[2]

[1] UPEC - CEDITEC – m-peres@wanadoo.fr
[2] UPEC – CEDITEC – jeanmarc.leblanc@u-pec.fr

Abstract

This paper, based on different academics fields, proposes a thought about data representation that leads us to develop a Statistical Analysis of Textual Data visualization tool. Through an analysis of methods of archeological restitution from "envois de Rome" to computer 3D models the role of the creator of those restitutions is enlightened and bonded to the responsibility of creating a scientific mediation. This thinking echoed positively with the thoughts of a scientist working in Statistical Analysis of Textual Data filed and the software that is develop in continuity of this methodological thinking, TextObserver, intends to created new models for textual and multimodal data analysis. It proposes new visualization functionalities that are made clearly evident by interactivity and dynamic data process of textometric results. It allows integration of diversified textual data in a multimedia framework and permit scientific experimentation in order to explore discourse variation factors. The meeting between those two questionings is the basis of the development of this exploring tool.

Résumé

Cette contribution prend appui sur des champs disciplinaires différents pour proposer une réflexion sur la représentation des connaissances qui nous à conduit à développer un outil de visualisation des données textuelles. Nous analyserons ici les méthodes et principes de restitutions archéologiques depuis les envois de Rome jusqu'aux modèles informatiques tridimensionnels interrogeant, leur forme, leur propos et le rôle de transformateur d'information que doit endosser tout créateur de restitution archéologique afin de créer un objet répondant aux contraintes de la médiatisation et la médiation de l'objet d'étude. Cette démarche de questionnement de la représentation des données développée avec le modèle informatique analysé ici a trouvé un écho favorable dans les réflexions d'un chercheur en linguistique de corpus. Le logiciel développé dans le prolongement de cette réflexion, *TextObserver*, vise précisément à introduire de nouveaux modèles de représentation des données et des résultats pour l'analyse des corpus textuels et multimodaux. Il propose des fonctionnalités originales sur le plan de la visualisation, rendues explicites par l'interactivité, et du traitement dynamique des données et des résultats textométriques. Il rend possible l'intégration de données textuelles diversifiées dans un cadre multimédia et répond en temps réel aux questionnements expérimentaux comme les facteurs de la variation discursive. La rencontre entre ces questionnements novateurs en lexicométrie et ceux développés avec la création du modèle archéologique tridimensionnel servent de fondement à la création de cet outil exploratoire.

Mots-clés : Visualisation, modélisation, textométrie, ergonomie, interfaces, représentation, corpus

1. Réflexions sur la représentation en archéologie, médiatisation et médiation de l'objet d'étude

Les réflexions présentées ici sur les représentations en archéologie en tant que médiatisation et médiation d'un objet d'étude sont faites du point de vue du plasticien spécialiste de la représentation. Le modèle informatique du *Circus Maximus* présenté ici fut l'objet de notre thèse de doctorat (PERES, 2001) et est issu d'un déplacement transdisciplinaire en vue de

créer un outil de gestion des connaissances sur un site archéologique important et pour lequel la quantité de recherches, de connaissances et d'intervenants nécessitait d'inventer d'autres moyens de se saisir du réel et de le transmettre. Il s'agit de recréer le monde pour mieux le saisir.

1.1. Données et visualisations

La question de la représentation des connaissances est inféodée à celle de la visualisation des données et interroge autant le type de données ou de connaissances à représenter que le type de représentation.

1.1.1. Le monde tel qu'on le voit, le monde tel qu'on le connait

Si l'on considère l'histoire de l'art, il est possible de situer un point clé dans l'évolution des représentations. En effet, en schématisant un peu (beaucoup) on peut dégager une période avant la Renaissance où il existe un multitude de règles de représentation dont le propos n'est pas tant de représenter le monde tel que l'on le voit mais tel qu'on le connaît ou le reconnaît. Les représentations dans le monde occidental codent ainsi l'importance relative des différents sujets représentés, le cheminement à suivre, les informations permettant une reconnaissance immédiate du sujet représenté, de son statut et de la situation. Ces représentations sont en fait éminemment discursives.

Avec la fameuse « Cité idéale » de Piero Della Francesca, la perspective à point de fuite s'exprime dans toute sa splendeur. C'est l'avènement de la représentation du monde tel qu'on le voit qui s'initie ici et qui va prédominer par la suite créant un présupposé culturel de réalisme des représentations avec point de fuite unique. Il s'agit de creuser l'espace en donnant l'impression que le réel et sa représentation sont superposables, en une sorte de calque de la réalité supprimant la distance à l'objet.

Il est cependant nécessaire de faire la distinction entre les perspectives. En effet, si la perspective à point de fuite multiple développe une représentation du monde à visée hyperréaliste du point de vue de la description de l'espace (sans jamais pouvoir l'atteindre, ne serait-ce que parce que nous avons une vision binoculaire qui n'est transposable partiellement que par des moyens techniques sophistiqués), il existe d'autres perspectives qui ont d'autres propos. Ainsi les perspectives isométrique et cavalière ont pour but de décrire au mieux la construction de l'objet en conservant les dimensions avec l'éloignement.

1.1.2. La représentation scientifique au 18e siècle : les envois de Rome

Au 18e siècle deux préoccupations dans la représentation à visée scientifique explorent des axes extrêmes. En effet vont coexister les envois de Rome et les premiers graphiques.

Les envois de Rome sont les productions de fin de séjour des architectes prix de Rome envoyés dans la ville éternelle faire leurs armes et acquérir sur place une culture plastique en vue de développer leur art. Il s'agit de restitutions archéologiques de différents sites romains dans un premier temps (la zone couverte s'élargira progressivement aux grands sites archéologiques italiens puis grecs jusqu'en 1968). Ces travaux dont le sujet était choisi avec l'approbation du directeur de l'Académie de France à Rome étaient envoyés à Paris à la fin du séjour du primé. On les nomme « envois de Rome » ou plus rarement « restaurations ». Cette distinction est intéressante car il s'agit bien de propositions de restitutions archéologiques faites à partir des relevées de monuments et des observations des architectes.

Les envois de Rome sont le plus couramment des restitutions graphiques. Il s'agit de dessin au trait, sciographie ou aquarelle, projection orthographique (cartes donc) ou vue en perspective mais l'on compte également des maquettes dont la plus célèbre est sans doute la maquette du *Circus Maximus* présentée par Paul Bigot en 1900. Cette maquette d'un monument unique se verra adjoindre petit à petit l'ensemble des monuments de la Rome antique pour aboutir à un magnifique plan relief conservé à l'Université de Caen Basse-Normandie.

Parmi les représentations graphiques on peut citer par exemple les thermes de Caracalla proposées par Abel Blouet en 1826. Cette planche propose côte à côte une vue en élévation des ruines relevées par l'architecte et une proposition de restitution. Les deux sont traitées dans un style précis et sans point de fuite emprunté directement aux représentations architecturales qui permet en temps normal de représenter une future réalisation. Le traitement graphique propose au lecteur un objet à l'aura scientifique indéniable, si précis qu'il ne peut être accepté que comme exact.

Si ces représentations sont tout d'abord réservées à un petit nombre leur diffusion prend son essor au milieu du 19^{e} siècle au travers d'ouvrages traitant d'archéologie monumentale ou portant sur les grandes campagnes de fouille. Quelles que soient leur formes ces ouvrages permettent aux architectes de figurer au côté des archéologues puisqu'ils y prennent en charge les propositions de restitution. Parfois même les envois de Rome y sont repris sans aucune reprise ou modification validant ainsi sans réserve la proposition de restitution. Le type de représentation choisie fait bien sûr partie du discours… Les aquarelles adoptent par défaut les règles de représentations dont le lecteur attend culturellement un maximum de « vérité » et de « réalisme ». Mais cette attente n'est pas en adéquation avec le propos qui sous-tend la création de ces représentations. En effet, il s'agit non pas de représenter le réel tel que le préfiguraient les questions posées par les artistes à la renaissance mais bien de représenter les connaissances via ce même code de représentation. On montre les hypothèses archéologiques selon des modes de représentation les présentant culturellement comme la réalité. C'est en partie cette non-différenciation qui va provoquer en archéologie un véritable frein scientifique en figeant des modèles.

1.1.3. La représentation scientifique aux 18^{e} et 19^{e} siècles : les premiers graphiques

Le second axe que nous voudrions interroger ici, toujours au 18^{e} siècle, procède des questions de représentation statistiques. C'est la période où l'avènement des graphiques en secteur, histogrammes et autres représentations du même type souligne la préoccupation des scientifiques de permettre aux destinataires de mieux se saisir des données chiffrées ou temporelles.

En 1765, soit vingt ans avant les premières publications de Playfair, le chimiste Joseph Priestley avait le premier eu recours à une frise chronologique, dans laquelle des barres superposées de différentes longueurs permettaient de comparer les époques auxquelles vivaient différentes personnes. Priestley était d'avis que ses diagrammes, en représentant de manière schématique la disposition des parties d'un ensemble et l'évolution d'un phénomène donneraient à ses étudiants « une juste image de l'essor, du progrès, de l'étendue, de la durée et de l'état actuel de tous les grands empires ayant jamais existé de par le monde ». Le lecteur est donc prié d'associer les grandeurs géométriques visibles sur le graphique aux données que l'on cherche à représenter. Ce graphique se rapproche très fortement des graphiques des fréquences cumulées.

On attribue à William Playfair l'invention de l'histogramme, qui apparaît pour la première fois dans son *Commercial and Political Atlas*, publié en 1786. Selon Beniger et Robyn (Benier et Robyn, 1978:3), « C'est le manque de données disponibles qui mena Playfair à cette invention. Il avait, dans son Atlas, compilé une série de 34 planches sur les importations et exportations de différents pays sur plusieurs années consécutives, et les exploita en traçant ce qu'il appelait des *surface charts, graphiques de séries chronologiques* ; c'étaient des graphiques ombrés entre l'axe des abscisses et la courbe. Mais comme Playfair ne disposait pas de données pour l'Écosse, il représenta ces statistiques commerciales pour chaque année séparément, faisant figurer 34 bâtons par planche, soit un pour chacun des partenaires commerciaux. ». Il s'agit du premier graphique exprimant des donnés quantitatives mais ne les localisant ni dans l'espace (comme le ferait une carte) ni dans le temps (contrairement à la frise chronologique de Priestley).

Son travail sur les graphiques en secteur est également remarquable. Il cherche à rassembler sur une même représentation graphique la superficie de chaque pays considéré, la population (en millions d'habitants), le total des taxes collectées (en millions de livres sterling) et le rapport entre les deux.

En 1869, Charles Minard propose une lithographie montrant le nombre d'hommes au sein de l'armée de Napoléon lors de la campagne de Russie de 1812, leurs mouvements, ainsi que les températures rencontrées lors du retour. Le tout sous la forme d'une carte (le trajet est figuré spatialement il le serait sur une carte classique mais à cette indication s'en ajoutent d'autres grâce à la largeur de trait variable et aux couleurs) qui combine information géographique, temporelle et quantitative. Cette grande complexité fait date dans l'histoire des représentations en intégrant discrètement, sans que le lecteur s'en rende compte ou presque, un ensemble de données à six dimensions (Tufte, 2001). L'ensemble est saisissant d'efficacité et lisible très facilement. Il est à noter cependant qu'un paragraphe explicatif fait office de légende et éclaire grandement la représentation.

1.1.4. Quelles représentations pour quelles données ?

Il va sans dire que toute représentation doit être observée d'un œil critique dans la mesure où elle est forcément une simplification des données chiffrées et que les déformations sont toujours possibles (surtout en cas de comparaisons de deux diagrammes où sera nécessaire de vérifier les systèmes d'échelles, les coordonnées et les choix de représentation en général).

Il est d'usage de préférer certains types de représentation graphique en fonction du type de données à présenter. La forme est en fait inféodée au discours qu'elle porte. Nous n'évoquons ici que les représentations les plus utilisées dans le but de les expliciter rapidement.

Ainsi, les données de types chronologiques seront plutôt représentées par des diagrammes de type courbes qui permettent de mettre en évidence les tendances profondes d'un phénomène, en éliminant, par exemple, les variations saisonnières des observations. Les données quantitatives discrètes donneront plutôt lieu à la réalisation de diagrammes en bâtons (seule la hauteur est importante). Ils permettent en effet de se saisir rapidement des quantités relatives et c'est l'outil idéal pour faire des comparaisons entre chaque bâton. Cependant un diagramme en bâtons n'est intéressant que si l'on prend en compte un nombre peu élevé de données et que l'on ne cherche pas à comparer des données issues de populations de tailles différentes. Si la taille des populations varie on préférera utiliser un histogramme dont la base des bâtons varie avec la taille de la population. C'est alors la surface du bâton qui est à considérer et non sa hauteur seule. Il permet plus facilement de se saisir de données

quantitatives continues. L'expression de données à caractère qualitatif se fera plutôt en utilisant un diagramme en secteurs qui nécessite que les données utilisées soient exprimées en fréquences et ramenées à un tout et permet de visualiser facilement les proportions.

Bien entendu, cette typologie évolue en fonction des données mais aussi du propos. On pourra bien sûr basculer d'un type à l'autre ou encore présenter d'autres visualisations qui n'ont pas été évoquées ici.

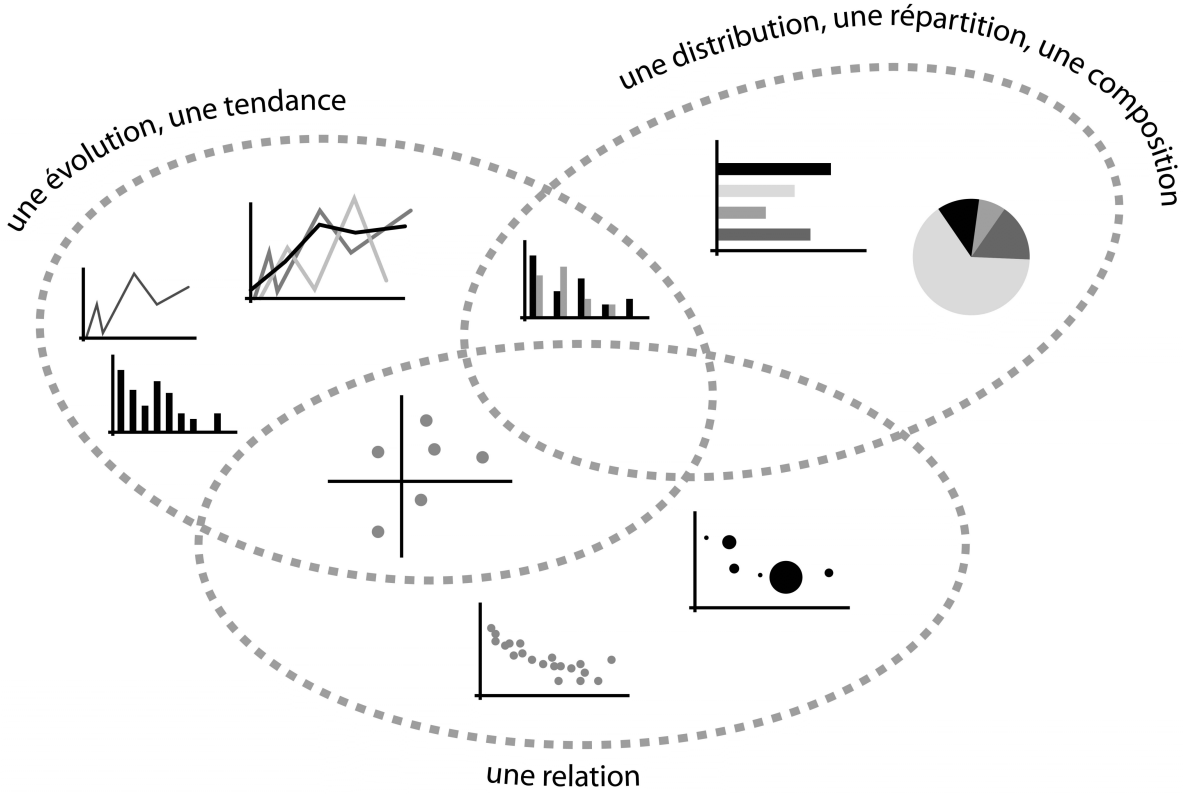


Figure 1. Tentative de typologisation des différents graphiques les plus utilisés en fonction du type d'information transmise par la présentation des données

Si l'on considère les représentations les plus utilisées en lexicométrie on peut ainsi proposer la typologie ci-dessus. Bien entendu en fonction du propos et du contexte les différents graphiques peuvent se « déplacer » dans le schéma et leur position n'est pas figée.

1.2. Représentations de données et terminologie

La multiplication des représentations à donné lieu à la création d'une terminologie dont certains termes sont extrêmement récents et qui mettent en avant les préoccupations de leurs créateurs : définir le rôle des représentations visuelles dans la communication de l'information. C'est donc à dessein que nous évoquerons ici des notions fondatrices de la visualisation de l'information.

1.2.1. Infographie

Les années 70 ont vu émerger un terme nouveau : *Infographie*. La signification du terme à longtemps fluctué, servant tout d'abord plutôt à désigner les graphismes fait par ordinateurs dans le but d'informer, que l'on appelle alors « infographies ». Ils sont alors destinés à mettre en image des informations généralement statiques. Dans les années 90, le terme désigne plus largement la création d'images numériques assistée par ordinateur (graphismes 2D,

graphiques, photos retouchées, images 3D…) et l'infographiste devient un graphiste informatisé. Le terme à encore évolué au cours du temps pour désigner de nos jours certaines productions aux qualités esthétiques travaillées représentant un concept par une image plus ou moins complexe et pouvant intégrer du texte et des chiffres.

1.2.2. Data visualisation

Forgé par Friedman en 2008 ce terme désigne les représentations de données quantifiables permettant de communiquer l'information clairement et efficacement grâce à des moyens graphiques (fonctionnels) sans que cela exclue un travail esthétique.

1.2.3. Data science

Plus qu'une visualisation, la *data science* est définie par ses auteurs comme une discipline ayant pour but l'aide à la prise de décision (et donc plutôt destinée aux consultants). En 2009, DJ Patil et Jeff Hammerbacker, proposent 3 étapes successives afin de permettre de transmettre clairement les différents éléments. La première étape consiste en la collecte de données (autrement dit la création d'un corpus), la seconde est réservée au nettoyage des données (nettoyage du corpus), et la troisième est celle de la visualisation et de l'exploitation des données. Le problème consiste à clairement indiquer les choix faits lors du rassemblement et du nettoyage des données et ceux présidant à la représentation pour que le lecteur puisse interpréter correctement.

Qu'il s'agisse d'Infographie, de *data visualisation* ou de *data science*, c'est bien la nécessaire transformation de l'information qui est interrogeable ici.

1.3. Le créateur d'objet multimédia : un transformateur d'information

Tout ce qui a été exposé jusqu'ici amène à s'interroger sur le rôle du créateur de l'objet servant à transmettre la connaissance. Avec l'évolution des moyens techniques il n'est plus simple graphiste informatisé mais bien créateur d'objet multimédia (animation, son, informations supplémentaires, accès aux sources, pluralité des médias se confrontant pour mieux exprimer les résultats…).

Nombre d'ouvrages de référence traitent des questions de représentation multimodale au delà des questions fondatrices que soulevaient la sémiologie graphique de Bertin (Bertin, 1967), la mutation de ces questions vient de la nécessaire évolution du signifiant (en tant que représentation du monde de plus en plus réticulaire) et du signifié (en particulier grâce aux moyens techniques qu'offre l'information) (Grataloup, 1996:16).

Cette évolution existe déjà en germe si l'on considère les modalités de représentation d'ordre numérique et analogique. Une simple horloge est un bel exemple des différences qu'induisent un choix entre numérique et analogique. Une représentation analogique de l'heure (avec un quadrant à aiguilles) permet de connaître l'heure qu'il est mais il permet aussi de visualiser la proportion des minutes écoulées dans l'heure par rapport à celles qu'il reste ou encore de faire la même chose pour l'heure par rapport à la durée d'une demi-journée de 12h (dans sa forme la plus courante). La mise en relation au contexte (précisé ou sous entendu) est ici très importante dans la représentation. Le sablier ou le cadran solaire ont exactement les mêmes propriétés. L'approche contextuelle en rattachant le particulier au général permet de comprendre le message tout en le situant par rapport à des paramètres de reconnaissance. Le mode de représentation analogique code ainsi un contexte graphique en même temps que l'information. Le mode numérique (pur) lui présente l'information hors contexte. Une horloge

numérique nous indiquant 08:30 est parfaitement lisible mais nécessite que le lecteur forme lui-même le cadre de rapport au contexte. De plus une information évoluant rapidement sera difficile à saisir par un affichage numérique malgré sa précision plus grande que l'affichage analogique. De nombreuses représentations existent, mêlant les modes analogique et numérique. Elles permettent de mieux se saisir de l'information de manière pertinente et rapide. L'approche analogique (souvent première) permet l'appréhension globale de l'information dans son contexte et est renforcée et précisée par le numérique (le chiffre – donnée – source, à la décimale près). Cet exemple déjà signifiant si l'on examine une représentation graphique « simple » évolue grandement si l'on ajoute une dimension temporelle à la représentation ainsi que, par exemple, des liens vers les sources.

1.4. Représentation et archéologie : L'exemple du modèle informatique du Circus Maximus

Plus qu'une maquette tridimensionnelle d'un site archéologique, le modèle informatique du *Circus Maximus* est un objet complexe et évolutif. Constitué d'un certain nombre d'éléments il comprend une maquette en trois dimensions dans laquelle l'utilisateur peut se déplacer virtuellement. Cette maquette est l'interface première aux informations contenues dans le modèle.

1.4.1. Du réel à la réalité virtuelle une nécessaire médiation

La représentation que constitue le modèle nécessite une médiatisation et constitue une médiation. Il est donc nécessaire de s'interroger sur la (relative) neutralité des médias mis en œuvre, sur la (possible) neutralité de la représentation et sur la (nécessaire) possibilité d'avoir accès aux sources.

Le modèle actualisé, visible à un moment donné (ce qui inclue les connaissances déjà acquises par l'utilisateur au cours de manipulations antérieures) se pose comme l'interface d'accès au réel, l'objet d'étude. Le modèle lui-même reste insaisissable dans son ensemble trop complexe et étendu. Ce sont les relations des différents éléments entre eux, activés par un utilisateur donné qui donnent sa réalité virtuelle à l'objet informatique et garantissent la pertinence de la représentation. L'utilisateur actionne des connexions en élaborant des hypothèses et en cherchant des informations. Le résultat est pertinent de son point de vue en cohérence avec son questionnement. Un autre utilisateur ne peut se saisir des résultats sans devoir faire le cheminement inverse et s'interroger sur les questions qui y ont présidé.

1.4.2. Visualisation, interfaçage et enrichissement du modèle

Sa mise en forme permet dans un premier temps d'identifier le degré de certitude de la restitution. Un objet attesté sera représenté de manière hyperréaliste (l'obélisque central par exemple qui est connu et se trouve *Piazza del Popolo* à Rome), alors qu'un objet représenté seulement par l'élaboration d'hypothèses sera représenté dans un code visuel simple (couleur) permettant à la fois de déterminer un matériau principal et de visualiser son degré de certitude.

Toute action sur un des objets constituant la maquette permet de plus d'accéder à un certain nombre d'informations (grâce à une requête vers des bases de données) telles que le lieu où l'objet a été retrouvé, son emplacement de conservation, les travaux ayant été rédigés à son propos ou des représentations alternatives.

Par défaut la maquette est présentée dans son état au 1er siècle. En effet, le site a une histoire longue et complexe et l'état initial choisi fut celui où le cirque était à son apogée

architecturale. On peut également par le biais d'un clic sur un objet faire évoluer la visualisation de la maquette de manière à visualiser les apparitions des différents objets les composant (objet antérieur à celui désigné par le clic en couleur matériaux, objet datant de la même période en rouge et objets n'étant pas encore été construits en gris). L'apparence des objets issus d'hypothèses est donc volontairement simplifiée à la rencontre de l'objet antique désigné et de son imagerie moderne en vue d'être identifiable le plus rapidement possible.

L'enrichissement du modèle se fait au fur et à mesure des nouvelles hypothèses et des découvertes. Chaque chercheur impliqué dans le projet peut ajouter des éléments aux bases de données et si après concertation entre les différents membres du projet une nouvelle proposition de restitution venait à apparaître, elle serait intégrée à la maquette. L'ancienne proposition serait alors mise en mémoire et accessible à tout moment en tant qu'alternative précédente, accompagnée de l'argumentaire qui a conduit à l'infirmer.

2. *TextObserver*[1], un outil d'observation et d'exploration des données textuelles et multimodales

Dans une précédente contribution aux JADT (Leblanc, 2010), nous formulions un certain nombre de propositions visant à rendre plus ergonomiques les dispositifs textométriques et à implémenter un certain nombre de fonctionnalités de visualisation, de calcul, de retour au texte. Si une grande partie des propositions prenait pour terrain d'expérimentation l'analyse factorielle des correspondances, le projet dans son ensemble n'a pas pour objet ces seules représentations.

Ces propositions étaient les suivantes:

- Représenter l'analyse factorielle des correspondances sur trois axes, lorsque cela faisait sens, car il est des cas de signatures statistiques où la représentation tridimensionnelle peut occulter des phénomènes qui apparaissent plus distinctement sur deux axes (comme le phénomène de temps lexical).

- Implémenter le mouvement dans les représentations des résultats, ce qui permet de mieux saisir les phénomènes de variations[2].

- Proposer une navigation entre les différentes partitions d'un corpus.

Les développements de l'outil se sont poursuivis, se nourrissant de la réflexion qui vient d'être présentée.

Nous pourrions évoquer les fonctionnalités qui ont été développées selon la typologie suivante:

2.1. Fonctionnalités de lecture/affichage

Il s'agit d'apporter ici une aide à l'interprétation. Si nous nous en tenons à la seule analyse factorielle des correspondances, *TextObserver* permet d'afficher les colonnes seules, les colonnes et les lignes, uniquement les lignes, mais aussi de choisir les points ligne que l'on

[1] Conçu par Jean-Marc Leblanc, *TextObserver* est développé par Sébastien Jacquot, Amani Daknou, Marie Pérès.

[2] Le mouvement est implémenté dans les analyses factorielles des correspondances. Les développements actuels portent désormais sur l'implémentation du mouvement dans les réseaux de cooccurrences.

souhaite projeter sur la représentation factorielle. Ces affichages sont disponibles en 2D, 3D ou en version extrudée (voir ci-dessous).

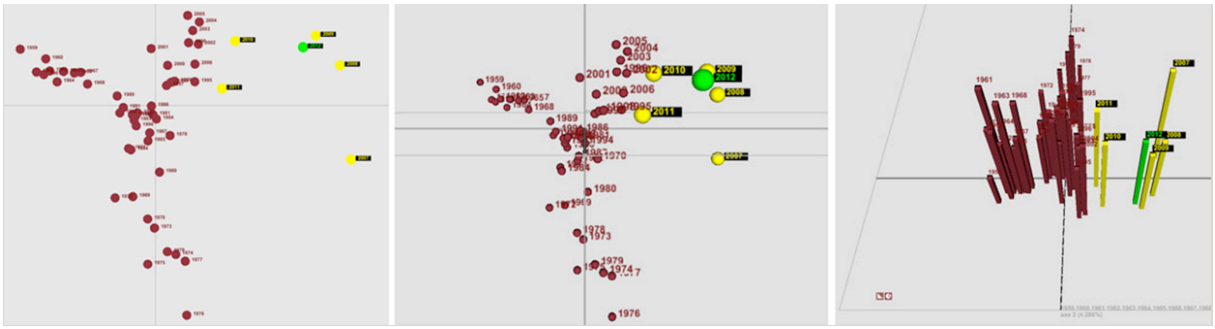


Figure 2. Affichage des points colonnes, en 2D ; 3D et version extrudée

La représentation de droite (version extrudée) permet d'éprouver le corpus sur la base de caractéristiques quantitatives. Sur cette figure, la hauteur des histogrammes représente la taille des parties du corpus. Cette représentation permet en outre de porter un regard sur les grandes oppositions du corpus, ici sur la partition par année. Ce visuel original permet donc d'appréhender simultanément deux phénomènes. Sur chaque représentation, l'utilisateur a la possibilité de zoomer et d'explorer la carte ainsi produite.

TextObserver permet en outre, en temps réel, d'attribuer une couleur à tous les éléments que l'on souhaite mettre en évidence. Ici en rouge les années 1959-2006 puis en jaune les années 2007-2011 (Sarkozy). En vert, décembre 2012 (Hollande). Les mêmes visuels sont disponibles avec les points lignes ou un extrait de ceux-ci[3].

L'affichage des points lignes peut se faire au moyen d'expressions régulières mais aussi par la création de listes thématiques.

Une innovation importante en termes d'affichage ou de lecture est la possibilité de visualiser, au moyen d'un curseur les points contributifs des différents axes, offrant à l'utilisateur de visualiser de façon instantanée les formes qui sont responsables de la configuration factorielle.

Parmi les fonctionnalités de lecture, il convient de souligner que *TextObserver* permet de calculer et d'afficher dans une même interface, des AFC provenant de corpus différents ou d'états différents d'un même corpus. En outre ces AFC peuvent être dupliquées, superposées, modifiées, manipulées. Enfin, il est possible de soumettre à *TextObserver* des tableaux de toutes sortes et non uniquement des tableaux lexicaux, ce qui permet de mieux comprendre le fonctionnement de l'analyse factorielle mais aussi de croiser les résultats issus d'un corpus textuel à des données autres.

2.2. Fonctionnalités de calcul

Les requêtes évoquées ci-dessus peuvent aussi s'appliquer au tableau lexical, c'est à dire aux données soumises à l'Analyse Factorielle, non plus au moment de l'affichage mais lors du calcul.

[3] Le corpus utilisé ici est constitué des vœux des présidents de la Cinquième République, de 1959 à 2012.

Ainsi à partir d'un tableau lexical entier, c'est-à-dire qui contiendrait l'ensemble des mots d'un corpus textuel, il est possible d'effectuer de nouvelles analyses factorielles en supprimant des lignes ou des colonnes.

On supprimera par exemple l'ensemble des pronoms personnels et adjectifs possessifs d'un corpus et on effectuera une nouvelle AFC sur ces données pour mesurer l'incidence des marques personnelles sur un corpus donné, ou au contraire on supprimera toutes les lignes (tous les mots) du corpus à l'exception des ces marques de l'énonciation pour effectuer une analyse factorielle sur ces nouvelles données. Dans tous les cas de figure, une interpolation de mouvement, permet de mesurer en temps réel l'évolution de la configuration factorielle, c'est-à-dire de visualiser le déplacement des points entre l'état initial et les autres états de l'AFC[4].

La même opération peut s'appliquer aux points colonnes, c'est-à-dire aux différents textes composant le corpus.

2.3. Retour au texte

Parmi ces fonctionnalités, il convient de souligner que le retour au texte est possible à partir d'un clic droit sur un point de l'AFC, qu'il peut porter sur une recherche à partir de la forme graphique ou d'une étiquette morphosyntaxique ou sémantique, selon le catégoriseur qui aura été appliqué au corpus[5].

TextObserver a été conçu comme un outil d'expérimentation, au moyen duquel l'utilisateur peut faire varier un certain nombre de paramètres (seuils, suppressions de parties du corpus, rétention ou suppressions de formes ou de groupes de formes) et observer la modification induite par ces paramétrages en temps réel, au moyen du mouvement. Au cœur de ce dispositif, le tableau lexical est un objet dynamique, et l'analyse factorielle devient interactive et cliquable: elle permet d'accéder au retour au texte, à des calculs de probabilité, de concordance, de cooccurrences...

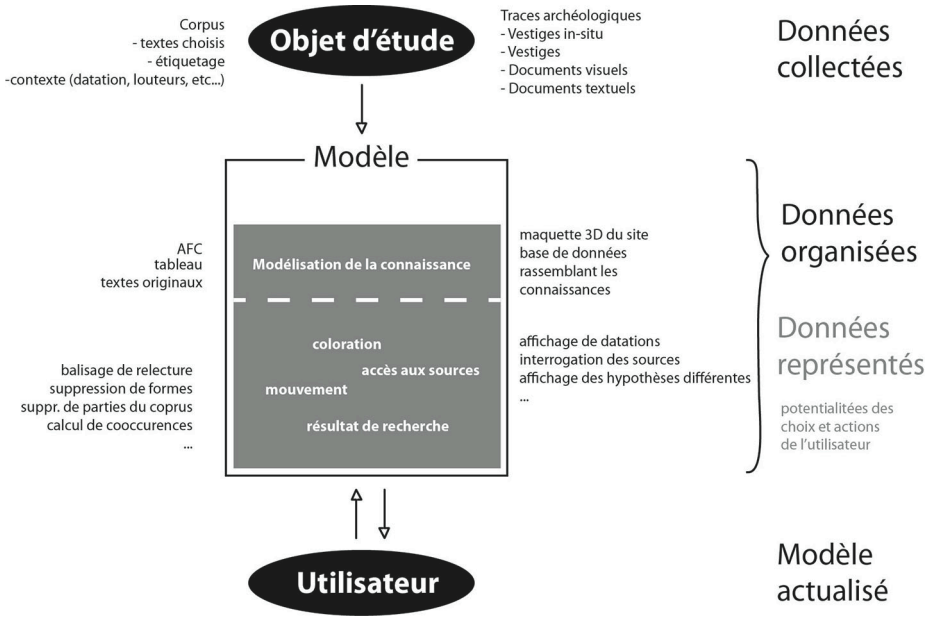
3. De l'archéologie à la textométrie… de la constance des questionnements

Nous avons souligné dans la première partie que le logiciel est une médiatisation plus ou moins explicite et qu'il convient donc de s'interroger sur la neutralité relative des médias et de la représentation. Dans tous les dispositifs on vise à la plus grande neutralité possible afin de ne pas déformer l'information.

Soumettre un texte à un logiciel de textométrie revient à le modéliser comme nous avions modélisé le *Circus Maximus*, cela ne le rend pas saisissable dans son ensemble mais permet un certain nombre de reconstructions par son utilisateur. (Une concordance, un graphe de cooccurrences, une recherche de motif sont autant de recompositions partielles). Le simple fait d'effectuer cette recherche et d'obtenir un résultat modifie l'image mentale que l'utilisateur se fait du texte, soit en la confortant, soit en la transformant. C'est aussi pour cette raison que le retour aux sources - ici au texte - est primordial. L'image mentale pourrait sinon complètement basculer et ne plus avoir de rapport avec le texte original tout comme le *Circus Maximus* pourrait devenir un cirque "rêvé".

[4] *TextObserver* est disponible au téléchargement à l'adresse http://textopol.u-pec.fr/textobserver.

[5] Un utilitaire de conversion est disponible à l'adresse http://textopol.u-pec.fr, rubrique boîte à outils, qui permet de transformer tout document tabulé en un fichier XML utilisable par l'outil *TextObserver*.



La figure ci-dessus met en parallèle les deux objets d'étude que nous confrontons dans le cadre de cette contribution. L'objet d'étude - corpus textuel ou traces archéologiques - est d'abord contextualisé, (partitions en dates, auteurs pour le corpus textuel, typologisation des traces archéologiques) puis ces données sont organisées en modèle c'est-à-dire en représentations de la connaissance. L'analyse factorielle qui typologise le corpus selon une partition donnée, correspond à la maquette du cirque. Les deux représentations donnent accès aux connaissances. L'analyse factorielle telle qu'elle est envisagée dans *TextObserver* rend possible le retour au texte, mais aussi d'autres calculs de fréquences ou de cooccurrences.

La troisième phase correspond aux possibilités de l'utilisateur de modifier l'objet représenté. Il peut ainsi colorer les points de l'AFC (points lignes ou points colonnes) afin de mettre en évidence les caractéristiques de son corpus, de même que l'utilisateur de la maquette du cirque peut modifier couleurs et textures en fonction de l'état des connaissances ou des hypothèses concernant l'objet d'étude. La configuration de la représentation factorielle peut également être modifiée en fonction de paramètres que l'utilisateur fera varier (suppression de formes, de parties du corpus...), en temps réel et au moyen, du mouvement, de même que la datation (états des connaissances à un moment donné) peut transformer les objets visibles selon les même modalités pour ce qui concerne la maquette du cirque.

4. Conclusion

En prenant appui sur l'étude d'une production venant d'un champ disciplinaire différent de ceux mettant habituellement en œuvre des logiciels de textométrie cette communication met en avant les constantes dans la réflexion menant à la représentation des connaissances dans le développement logiciel d'outils de visualisation des données.

Le logiciel *TextObserver* explore ces pistes et cherche dans cette ligne à introduire de nouveaux modèles de représentation des données et des résultats pour l'analyse des corpus textuels et multimodaux. Les fonctionnalités de visualisation qu'il propose sont explicitées par l'interactivité et le traitement dynamique des données et des résultats textométriques. Il

répond en temps réel aux questionnements expérimentaux comme les facteurs de la variation discursive. Tout comme le modèle du *Circus Maximus*, *TextObserver* est un outil exploratoire. Cependant, il prend une toute autre ampleur puisque le corpus n'est pas figé autour d'un objet d'étude comme l'est le modèle du *Circus Maximus* et cherche à proposer une médiation ouverte dont les objets peuvent être des corpus textuels mais aussi des corpus multimédias et multimodaux.

Références

Barats C., Fiala P. et Leblanc J.M. (2013). « Approches textométriques du web : corpus et outils », in *Manuel d'analyse du Web* (Dir Christine Barats), Armand Colin, Paris, 100-124.

Beniger J.R. et Robyn D.L. (1978). « Quantitative graphics in statistics: A brief history », in *The American Statistician*, no 32, pp. 1-11.

Bertin J. (1967). « Sémiologie graphique : les diagrammes, les réseaux, les cartes », Paris Mouton Gauthier-Villars, 431p.

Bonin S. (1983). « Initiation à la graphique : transcription visuelle des données statistiques et cartographiques », Épi éditeurs.

Cibois P. (2000). « L'analyse factorielle, Presses Universitaires de France » - PUF (Que sais-je ?), (5e éd.), Paris.

Guilmeau-Shala S. (2011). « En quête de la couleur : publication de dessins réalisés lors de voyages d'études en Grèce », in *Bibliothèques d'atelier. Édition et enseignement de l'architecture*, Paris 1785-1871, INHA (« Les catalogues d'exposition de l'INHA »).

Lebart L., Morineau A. et Piron M. (2000). « Statistique exploratoire multidimensionnelle », Dunod, Paris.

Leblanc J.M. (2010). « Nouvelles fonctionnalités pour la visualisation des données textuelles et des résultats : Pour une approche ergonomique des dispositifs lexicométriques » in actes des JADT 2010, 08-11 mars 2010, Rome.

Lechleiter F. (sous la direction de Foucart B.) (2008). « Les envois de Rome des pensionnaires peintres de l'Académie de France à Rome de 1863 à 1914 », thèse de doctorat, Université Paris IV.

Friendly M. (2002). « Visions and Re-Visions of Charles Joseph Minard » in *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 1 (Spring, 2002), pp. 31-51.

Pérès M. (sous la direction de Golvin JC.) (2001). « Réflexion sur le modèle informatique du *Circus Maximus* ». Thèse de doctorat, Université Michel de Montaigne - Bordeaux 3.

Tufte E. (2001). « The Visual Display of Quantitative Information », Cheshire, CT, Graphics Press, 2e éd. (1re éd. 1983).

Pérès M. (2006). « De la modélisation à l'image virtuelle : image et réel ». in *Figure de l'art*, vol.(11): 197-208.

Viprey J.M. (2006). « Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle : une projection « géodésique » » in actes des JADT 2006, 989-1000.

Wildbur P. et Burke M. (2001). « Le graphisme d'information, Cartes, diagrammes, interfaces et signalétiques », Thames & Hudson.