Visualisation interactive de données : les apports d'une étude de cas menée auprès d'étudiants de licence

Béatrice Drot-Delange

Visualisation interactive de données : les apports d'une étude de cas menée auprès d'étudiants de licence

Béatrice Drot-Delange, Université Clermont Auvergne, Université Blaise Pascal, EA 4281, ACTé, 34 avenue Carnot, 63037 Clermont-Ferrand Cedex 1

beatrice.drot-delange@univ-bpclermont.fr

Résumé : Interpréter une représentation visuelle de données fait appel à de nombreux champs de savoirs. Dans cette étude qualitative à visée exploratoire, des étudiants de première année de licence ont analysé la représentation visuelle interactive de leur choix à l'aide des principes d'excellence graphique de Tufte. Une analyse *a priori,* élaborée à l'aide d'un corpus de textes d'experts et des concepteurs de la visualisation, est mise en regard de la production des étudiants. Le regard croisé entre analyse *a priori* et analyse produite par les étudiants montre les difficultés rencontrées par les étudiants dans cette tâche d'analyse d'une visualisation.

Mots clés : littératie, interface, artefact cognitif, interactivité

Introduction

Les données sont au cœur de la société numérique. L'individu qui collecte et diffuse des données sur son propre corps (*self-tracking*), les états et institutions qui ouvrent massivement l'accès à leurs données ou bien encore les lanceurs d'alerte et le journalisme de données sont quelques exemples des multiples facettes de cette évolution. Cette abondance de diffusion de données va de pair avec le développement de leur représentation visuelle. Il convient en effet, face à cette abondance, d'offrir les moyens au public de comprendre, d'explorer, de faciliter la recherche ou la découverte, de comparer, etc.

Les principes d'excellence graphique développés par Tufte (2001, p. 51) insistent sur la capacité de la visualisation à communiquer au lecteur un grand nombre d'idées dans un temps court tout en faisant en sorte que la représentation visuelle respecte l'intégrité des données. Cela suppose de la part du lecteur des compétences dans le sujet traité, mais pas seulement. Pour Bertin, l'objectif de toute représentation graphique est de faire comprendre, c'est pourquoi il considère qu'on ne lit pas un graphique, mais qu'on lui pose des questions (1981, p. 62). Pour Minichiello, les compétences du lecteur sont également sollicitées par le concepteur de la visualisation :

> La visualisation fait partie intégrante d'un processus d'analyse des données et offre à l'utilisateur la possibilité d'explorer des chiffres, en lui proposant des clés pour une lecture critique et, parfois, l'opportunité de découvrir des éléments inattendus. (Minichiello, 2014, p. 10)

Le lecteur d'une visualisation interactive en est aussi l'utilisateur, car sollicité pour agir sur la représentation via le choix de paramètres, de filtres, de clics de souris, de zooms, etc.

Nous proposons dans cette communication, à visée exploratoire, d'étudier via une étude de cas, dans quelle mesure différents champs de compétences (médiatique, informationnel et informatique) sont mis en œuvre par des étudiants de première année de licence lors d'une tâche qui consiste à mener une analyse d'une visualisation interactive de données.

Méthodologie

Le contexte de l'étude

Depuis 2014, l'auteur de cet article propose un cours optionnel ouvert aux premières années de licence en information et communication et depuis 2015 également ouvert aux étudiants de langues étrangères appliquées. Ce cours de vingt d'heures aborde les thèmes de la structuration des données via les langages de balise tels que le html ou le xml, de la gestion des données personnelles et de la visualisation des données. Pour cette dernière, un historique est présenté. L'objectif est de favoriser un recul critique sur les effets attendus des visualisations tel que nous le signalions en introduction, de favoriser une mise à distance des productions actuelles, mais aussi de faire prendre conscience que les formes graphiques que les étudiants connaissent et produisent le plus (camembert, histogramme, etc.) ne sont pas nés avec les tableurs. Les principaux théoriciens de la visualisation de données sont présentés. Les résultats de la psychologie cognitive sur les biais perceptifs des représentations sont abordés. Les étudiants sont amenés en dehors des séances de cours à réaliser des travaux évalués comme l'analyse d'une représentation visuelle selon une grille qui leur est proposée et à produire par eux-mêmes une représentation visuelle sur le thème de leur choix à l'aide de logiciels dédiés.

Participants et matériels

Les participants

Cette étude a concerné 95 étudiants de licence (71 en sciences de l'information et de la communication et 24 en langues étrangères appliquées). Les données ont été recueillies en avril 2015 (pour 44 étudiants) et en avril 2016 (pour 51 étudiants). Les étudiants ont disposé de trois semaines (à domicile) pour mener l'analyse de la représentation visuelle de leur choix.

Les consignes et matériels fournis

Une liste de visualisations est fournie mais n'est pas impérative (voir tableau 1). Les étudiants pouvaient faire leur choix en dehors de cette liste. Elle reprend des cas particulièrement populaires sur les sites consacrés au sujet ou dans les manuels (Meirelles, 2014). La liste est organisée selon une classification qui reprend les principales questions auxquelles la visualisation tente de répondre.

Tableau 1 Liste d'exemples de visualisations fournie aux étudiants

Structure temporelle	Google Trends, Baby Name Voyager, The Ebb and Flow of Movies, Historical Browser Statistics
Structure spatiale	ZIPdecode, Google Earth, MapBlast, CartoDB
Structure textuelle	Wordle, Visual Thesaurus, Web Seer
Structure hiérarchique	Money Marketmap, Federal Budget Data, D3 TReemap
Structure relationnelle : réseau	The Brain, Touchgraph, Diseasome map, Wikipedia map
Divers	D3J, Informationisbeautiful

Une grille d'analyse est proposée aux étudiants, structurée en trois parties. La première partie concerne la représentation choisie. Les étudiants rappellent le titre de la visualisation, les sources des données, l'URL de la visualisation, si les données à la base de la visualisation sont accessibles ou non. La deuxième partie est un commentaire de la visualisation. Les étudiants doivent expliquer en quelques lignes l'objet de la visualisation, les conclusions que l'on peut en tirer. Ils doivent également préciser pourquoi elle a retenu leur attention. La troisième partie consiste en une évaluation de la visualisation à partir des critères prônés par Tufte dans la conception d'une représentation visuelle (2001). Ces critères sont au nombre de huit.

1) Présentation bien conçue de données intéressantes : sujet, statistiques et conception.
2) Des idées complexes sont communiquées avec clarté, précision et efficacité.
3) La visualisation donne le plus grand nombre d'informations pour un temps d'attention le plus limité possible dans le plus petit espace possible.
4) La visualisation ne distord pas le sens des données (intégrité statistique du sujet et artistique).
5) La visualisation encourage les comparaisons.
6) Les données sont multivariées.
7) Il n'y a pas de fioritures graphiques inutiles.
8) La visualisation raconte une histoire, suscite la curiosité.

Méthode d'analyse

Nous avons recueilli 95 analyses de visualisation. Trois ont été retirées du corpus, car incomplètes ou ne respectant pas la trame proposée. Nous avons donc retenu 92 analyses.

Dans cette étude exploratoire, nous retiendrons la visualisation qui a le plus obtenu les faveurs des étudiants : BabyName Voyager. Elle a été retenue par 18 étudiants, soit environ 20% de nos effectifs.

Elle fait l'objet dans un premier temps d'une analyse *a priori* sur les dimensions médiatiques, informationnelles et informatiques, élaborée à partir d'un corpus de documents émanant d'experts et/ou des concepteurs de la visualisation. Ce corpus n'est pas exhaustif, mais permet à ce stade de notre étude d'éclairer les différentes dimensions souhaitées.

Le corpus retenu pour la visualisation BabyName Voyager est le suivant :

Communication scientifique	Wattenberg, M. (2005). Baby names, visualization, and social data analysis. Dans *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (p. 1-7). IEEE.
Manuel	Börner, K. et Polley, D. E. (2014). *Visual insights: a practical guide to making sense of data*. Cambridge (Mass.), Etats-Unis d'Amérique. p. 50-51
Article scientifique	Elmqvist, N., Moere, A. V., Jetter, H.-C., Cernea, D., Reiterer, H. et Jankun-Kelly, T. J. (2011). Fluid interaction for information visualization. *Information Visualization*. p. 333-334
Site web	Social Security Administration *https://www.ssa.gov/oact/babynames/background.html*
Site web	http://www.babynamewizard.com/voyager

L'analyse thématique du corpus des productions écrites des étudiants est ensuite mise en regard de l'analyse *a priori*. Cette méthode permet de mettre en exergue les difficultés rencontrées par les étudiants lors de cette tâche. En creux, ce sont les savoirs et les connaissances mis en œuvre, ou leur absence, qui sont définis.

Résultats

Analyse a priori

Description de la visualisation

La visualisation est un graphique à aire empilée ou en couche. Le temps est représenté horizontalement de gauche à droite par décennies depuis 1880. Chaque prénom est représenté par une aire, dont la hauteur est proportionnelle à sa fréquence d'occurrence (nombre d'occurrence par million de nouveaux nés) sur l'intervalle de temps considéré. Le code couleur permet de distinguer les prénoms attribués à des filles (rose) ou à des garçons (bleu). La saturation de la couleur renvoie à la popularité du prénom.

Le lecteur/utilisateur peut choisir de sélectionner les prénoms masculins, féminins ou les deux. Il peut saisir le début d'un prénom ou un prénom et voir afficher sa popularité par décennie. S'il clique sur le prénom, s'affiche des informations sur le prénom, fournies par les utilisateurs du site eux-mêmes.

Source des données

La source des données n'est pas mentionnée dans la visualisation. Pourtant, elle utilise les données[1] de la *Social Security Administration* (SSA) des États-Unis. Pour une période de temps choisie par l'utilisateur (année ou décennie), depuis 1900, la SSA publie les prénoms les plus populaires (top 5, top 100, top 1000) donnés aux bébés aux États-Unis pour les filles et les garçons. Les données sont issues des dossiers des cartes de sécurité sociale.

Image-interface

Le concepteur de la visualisation a nettoyé et normalisé ces données pour arriver à environ 6000 séries temporelles de prénoms distincts. Dans chaque vue, le nombre de séries temporelles est limité à 200 pour des questions de rapidité de chargement des informations et de leur affichage. N'apparaissent que les prénoms représentés par une couche d'au moins 2 pixels d'épaisseur. Les séries non représentées sont laissées en gris, quasiment imperceptibles dans la vue d'ensemble. Dès que l'utilisateur saisit une lettre dans la zone de recherche d'un prénom, la visualisation affiche les prénoms commençant par cette lettre. Elle évolue ensuite au rythme de la saisie de l'utilisateur.

Analyse des réponses des étudiants

Description de la visualisation

La description faite par les étudiants de la représentation montre parfois l'embarras dans lequel ils se trouvent pour nommer les objets qu'ils doivent décrire. Ainsi, l'étudiant (1)[2] décrit des « barres en tas » et des « barres entassées », un autre (36) pense avoir à faire à des « coordonnées parallèles ». Peu (2) se lancent dans une description précise de la visualisation. Celle-ci est souvent entachée d'erreurs sur les unités de mesure :

- *« le nombre de millions de fois que le prénom a été choisi à la naissance selon les années* » (25). En fait, c'est le nombre d'occurrence du prénom par millions de naissance.
- « *La popularité des prénoms donnés aux bébés dans le monde (et toute époque confondue). (...) le nombre de bébés qui ont le même prénom par millier de naissance en ordonnée* ». (74)

Des conclusions sont tirées par un étudiant (84) sur le nombre de naissances, ce qui est incorrect :

- « *on peut aussi se dire que cette visualisation montre l'évolution du nombre de naissances, même si tous les prénoms n'y sont pas présents puisque on ne trouve sur ce graphique que les prénoms qui ont déjà été rangés dans le top 1000* ».

Source des données

[1] Accessibles à l'adresse https://www.ssa.gov/oact/babynames/decades/index.html (consulté le 26 aout 2016). Toutes les informations de ce paragraphe sont issues de ce site.

[2] Dans la suite du texte, les nombres entre parenthèses identifient la grille d'analyse d'un étudiant donné.

La question de la source et de l'accessibilité des données révèle des confusions chez les étudiants. En effet, la source n'est pas indiquée sur la visualisation. Les étudiants confondent source des données et site hébergeant la visualisation.

De même, bien que les données brutes ne soient pas directement accessibles, deux étudiants l'affirment en expliquant que :

- « *tout le monde peut accéder au site et aux visualisations* » (40) ;
- « *elles sont accessibles lors d'un double clic sur le prénom* » (79).

Concernant la source de ces données, les étudiants confondent l'hébergeur de cette visualisation et ses partenaires avec la provenance des données. Deux ne répondent pas à la question et 2 indiquent ne pas connaitre la source des données. Cette source n'est en effet pas indiquée sur le site lui-même, il faut faire une recherche et lire un article scientifique rédigé par le concepteur de la représentation pour la connaitre (voir la section précédente sur l'analyse *a priori*).

Image-interface

L'auteur de la visualisation écrit que cette forme de graphique est familière pour le grand public (Wattenberg, 2005, section 2.2 Visualization Method), elle ne semble pourtant pas si facile à comprendre.

- « *On ne comprend pas dès le début si la ligne la plus haute représente le nom le plus donné* » (45)
- « *En regardant le graphique, il parait complexe et ne nous donne aucune information. Mais dès que nous passons le curseur de notre souris sur une courbe, les informations essentielles s'affichent et on comprend rapidement comment fonctionne la visualisation* ». (25)

Cette première impression est partagée par d'autre étudiants (38, 39, 58). Certains précisent même que si « *on ne déplace pas le curseur de la souris sur les différentes branches du graphique, on ne peut obtenir aucune information concrète* » (78).

Bref, il est nécessaire de s'approprier l'interface pour comprendre ses potentialités :

- « *Le graphique est clair mais pour cela il faut trouver et comprendre comment on se sert de la barre de recherche des prénoms (...), comprendre que nous avons un lien pour avoir des critères de recherches (...)* » (84).

Parfois, l'interface ne répond pas à toutes les attentes :

- « *On ne peut pas faire de comparaison dans un même graphique entre deux prénoms qu'on recherche* » (36).

Traitement des données

A la question de l'éventuelle distorsion des données (soit statistiquement soit esthétiquement), les étudiants sont unanimes pour affirmer que cette visualisation est respectueuse des données, puisqu'on accède aux chiffres précis en cliquant sur les aires. Ils accordent une confiance totale aux traitements faits par le concepteur, traitements dont il n'est pas certain qu'ils aient conscience, sur des données dont ils ne connaissent par ailleurs pas la source.

Discussion

L'étude de cas présentée ici met en exergue quelques composantes d'une littératie de la visualisation de données (Lee *et al.*, 2016; Womack, 2014). Tout d'abord, les étudiants rencontrent des difficultés de lecture et d'interprétation des visualisations : l'échelle ou les unités de mesure ne sont pas toujours interrogées ou comprises. De même, les conclusions ou les inférences faites à partir de ces visualisations sont parfois fausses (inférer le nombre de naissances à partir du nombre de prénoms

attribués par million de naissances).

Cette littératie implique aussi d'être en mesure de s'interroger sur les sources des données, sur le message délivré, bref, d'être critique par rapport à la visualisation proposée. La visualisation des données s'inscrit dans un processus itératif faisant intervenir de nombreux acteurs et étapes intermédiaires entre les données brutes et leur visualisation (Börner et Polley, 2014). À chaque étape des choix sont faits, des traitements opérés sur les données.

L'interactivité incite les étudiants à penser que la visualisation est fidèle et fiable, puisque d'un clic ou d'un survol de souris, des nombres apparaissent, qui sont bien ceux représentés graphiquement, prouvant en quelque sorte le bien-fondé de la visualisation et confirmant alors le respect des données. Ces actions peuvent engendrer la croyance d'une objectivité des données ainsi visualisées, puisque décidées et mises en forme par l'utilisateur. Tout se passe comme si la visualisation s'auto-validait ou s'authentifiait pour reprendre le terme de Treleani :

> (…) le site de visualisation de données nous promet un monde authentifiant et objectif, car nous savons que ces statistiques énoncées grâce à nos actions de manipulation dérivent de données non filtrées : n'importe quel autre usager pourrait arriver aux mêmes résultats en effectuant les mêmes manipulations. L'interactivité nous fait croire en l'objectivité de données qui sont évidemment filtrées et éditorialisées à l'avance, comme si le seul fait de les exposer n'était pas déjà une interprétation. (Treleani, 2014, p. 9)

Conclusion

Nous avons mené une analyse *a priori* d'une visualisation interactive choisie par des étudiants de première année de licence en information et communication et en langues étrangères appliquées. Ces étudiants ont mené leur propre analyse en respectant une grille élaborée à partir des critères d'excellence graphique de Tufte.

Nous avons montré que l'analyse de visualisations interactives de données nécessitait des savoirs relevant de nombreux champs : statistique, informatique, médiatique et informationnel.

Les résultats obtenus montrent les difficultés rencontrées par les étudiants à prendre en compte l'ensemble de ces dimensions. Le processus éditorial ainsi que les processus opérés sur les données brutes pour concevoir et produire une visualisation ne sont pas toujours perçus ou connus.

Bibliographie

Bertin, J. (1981). Théorie matricielle de la graphique. *Communication et langages*, *48*(1), 62- 74.

Börner, K. et Polley, D. E. (2014). *Visual insights: a practical guide to making sense of data*. Cambridge (Mass.) : The MIT Press.

Grammel, L., Tory, M. et Storey, M.-A. (2010). How information visualization novices construct visualizations. *IEEE transactions on visualization and computer graphics*, *16*(6), 943- 952.

Lee, S., Kim, S.-H., Hung, Y.-H., Lam, H., Kang, Y. et Yi, J. S. (2016). How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking. *IEEE transactions on visualization and computer graphics*, *22*(1), 499- 508.

Lloveria, V. (2015). Data design-moi un mouton. *Communication & Organisation*, (46), 99- 112.

Meirelles, I. (2014). *Design de l'information: représenter visuellement les informations* (traduit par C. Vair). France : Parramón.

Minichiello, F. (2014). La visualisation de données en éducation. *Revue internationale d'éducation de Sèvres*, 66, 10- 14.

Reyes, E. (2015). La datavisualisation comme image-interface. *I2D – Information, données & documents*, *52*(2), 38- 39.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. Dans *IEEE Symposium on Visual Languages, 1996. Proceedings* (p. 336-343).

Treleani, M. (2014). Dispositifs numériques : régimes d'interaction et de croyance. *Actes Sémiotiques*, (117). Récupéré du site de la revue : http://epublications.unilim.fr/revues/as/5035

Tufte, E. (2001). *The Visual Display of Quantitative Information*. Connecticut : Graphics Press.

Wattenberg, M. (2005). Baby names, visualization, and social data analysis. Dans *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (p. 1- 7). IEEE.

Womack, R. (2014). Data Visualization and Information Literacy. *IASSIST Quarterly*, *38*(1), 12- 17.