La qualité des données textuelles

Cas d'usage à la plate-forme Cortext de l'IFRIS

Philippe Breucker, INRA SenS (UR1326 - Sciences en Société)

Journées de Recherches en Sciences Sociales (SFER), session ingénieur "Qualités des données" - Angers, décembre 2013

Introduction

Cette communication vise à envisager le traitement de la qualité des données, abordée à la manière d'une rétro-ingénierie au cours des activités de conception d'un système d'information en ligne. En effet, l'analyse de corpus de textes hétérogènes passe par la production de données calibrées analysables, et présente un réel intérêt pour le chercheur en science sociales. La qualité est dans ce cas produite, et non intrinsèque aux données qui, provenant de sources humaines ou semi-automatiques, comportent inévitablement des incohérences.

Dans le contexte de la plate-forme Cortext [1] de l'IFRIS, des méthodes et des scripts d'analyse sont produits sur différents types de corpus textuels. Les données textuelles et les résultats des analyses sont hébergés au sein de la plate-forme. Ces analyses sont mises à disposition des chercheurs en Sciences Humaines et Sociales via une architecture conçue à la plate-forme, spécifiquement pour ces traitements.

Une première version de ce service est mise en ligne depuis 2011 et est utilisée par plus de 500 utilisateurs aujourd'hui. Depuis bientôt 2 ans, l'architecture est en cours de refonte de afin d'être plus adaptable à son interconnexion avec d'autres services et au développement de futurs modules. Ce changement a amené plusieurs axes de réflexion sur les critères de qualité que devraient supporter la version 2 de la plate-forme. Ce texte est un instantané de ces réflexions sur les démarches à entreprendre pour assurer la qualité des données et la pérennisation des services apportés par Cortext à la recherche.

1. Cortext est la plate-forme digitale de l'IFRIS (Institut Francilien Recherche, Innovation, Société www.ifris.org). La plate-forme est un instrument de recherche dédié à l'étude des corpus textuels, et produit des méthodes et des outils d'analyses accessibles aux chercheurs en sciences humaines et sociales (voir www.cortext.net).

Qualité ?

La notion de qualité des données textuelle n'est donc pas triviale dans le contexte d'usage en question ici. Les notions telles que l'accès, la propriété, le stockage et le référencement doivent être prises en compte du point de vue du cycle de vie des données en amont du système, dans le système et dans les analyses, c'est à dire de *l'auteur* des données à *l'auteur* des analyses. En effet, la principale caractéristique de la donnée textuelle que nous traitons en sciences humaines et sociales est qu'elle est produite par des humains, parfois assistés de machines. Qu'il s'agisse d'un article scientifique (obtenu par exemple sur une base centralisant les publications telle que Web Of Science), d'un brevet déposé à l'office européenne des brevets ou encore d'un commentaire tiré d'un blog, ces textes ont été produits par un auteur, qui est donc soumis à des erreurs, à des imperfections dans le langage qu'il utilise, etc. Cela conduit à concevoir un système d'information - et donc une démarche qualité - adapté. La démarche doit également être bâtie sur le respect des droits d'auteurs (voir plus loin), ainsi que des droits de diffusion des données. Ces informations sont indissociables des données elles mêmes et doivent être conservées à tout moment, pendant leur cycle de vie entier. La notion d'archivage revêt également une importance particulière pour les sciences humaines et sociales, tout particulièrement lorsqu'il s'agit du web, car c'est un espace changeant qui, si l'on veut l'analyser, demande un archivage calibré.

Ces informations sont donc elles-mêmes des données, stockées sous forme de méta-données. Elles garantissent la pérennisation de la donnée elle même. Mais il est évident que notre travail sur la qualité ne peut s'arrêter à de simples informations adjointes aux données. Le système d'information conçu pour les exploiter doit pouvoir les stocker, en prévoir l'accès par un programme ou un utilisateur et éventuellement l'archivage ou la destruction.

Nous proposons en ce sens quelques uns des points de structures essentiels à une démarche qualité :

1. **autonomie** - L'accès, le stockage et le traitement des données ne doivent pas être effectués sur la même machine. Cela garantit une indépendance matérielle et donc une meilleure disponibilité. Cela permet également une évolution de service "atomique" (chaque partie pouvant être remplacée sans changer les autres). D'autre part, chaque service (accès, calcul, stockage, archivage,...) devrait être géré de façon indépendante et asynchrone. Les services doivent pouvoir être arrêtés ou être mis à jour sans que le système en entier doivent s'arrêter. Cela assure la cohérence des méta-données dans le temps, ce qui est fondamental quant à la qualité des données.
2. **redondance** - Tous les éléments (qu'ils soient matériels ou logiciels) devraient être redondants, c'est à dire duplicables et dupliqués. Cela assure la disponibilité des services, mais également la facilité d'installation ou de remplacement d'éléments défectueux, qui nécessiterait sinon une rupture du service. C'est un point important de la démarche qualité, car

la continuité de service assure une qualité d'exploitation de ceux-ci.

3. **sauvegarde** - À ne pas confondre avec la redondance, la sauvegarde (et, dans une autre mesure, l'archivage) est avant tout un moyen de restauration à tout moment des éléments dit "chauds "(i.e. qui changent en permanence). La restauration assure la pérennité des données et des méta-données et donc la qualité de celles-ci. L'archivage assure quant à lui l'historique des données, et d'ailleurs peut-être utilisé lui-même comme source de données textuelles. La traçabilité est un point fondamental de la qualité d'une donnée : elle permet entre autre la vérification de cohérence des données et méta-données.
4. **accessibilité** - Les données doivent être accessibles, c'est-à-dire posséder une norme d'accès documentée, qui permet leur accès de l'extérieur et assure la présence des méta-données lors du transfert.
5. **normalisation** - Les données et méta-données doivent posséder une norme de stockage et de diffusion calibrée sur tous les systèmes, au risque de n'être pas exploitables. Cela assure un transfert et une exploitation corrects, ainsi que l'intégrité des données. Une mise aux normes peut s'avérer nécessaire, ainsi que l'emploi de référentiels de données tels que les ontologies par exemple.

Lorsqu'on traite des corpus textuels, on doit cependant prendre en compte le fait que les méta-données ne sont pas toujours disponibles ou récupérables à la source. Les sources ne sont pas toutes soumises à une démarche qualité (loin s'en faut) et les données peuvent également venir de sources hétérogènes, possédant différentes normes. Les méta-données peuvent également avoir été égarées, ou tout simplement séparées des données. Il s'agit donc parfois d'enrichir les données à la main ou semi-automatiquement afin de les mettre aux normes de la plate-forme.

Propriété

Naturellement nous nous posons la question de la propriété de la donnée. Un corpus [2] de qualité est un ensemble de données dont nous connaissons au moins l'auteur - ou les auteurs - (qui peut être identifié par un simple numéro unique, un pseudonyme ou par son patronyme par exemple). L'auteur du texte, lorsqu'il est connu, possède certains droits, dépendants de la manière dont les textes ont été produits, et l'utilisabilité de son œuvre dépend également de l'émetteur des données (éditeur, base en ligne, site web, ...) qui peut avoir une licence d'utilisation propre. La licence d'utilisation ainsi que les "ayant-droits" doivent donc théoriquement être conservés avec le corpus. Dans la pratique il est rare d'avoir accès à la licence elle même au niveau du texte : on doit

2. Nous parlons ici de corpus comme d'un ensemble cohérents de textes, concernant généralement une thématique ou une base bien précise. La notion de "dataset" est parfois également employée à la plate-forme pour désigner un jeu de données textuelles. La taille des corpus est très variable (de quelques dizaines de documents à plusieurs millions)

donc s'en tenir aux conditions d'utilisation de la source des données. C'est un aspect très difficile à automatiser : en effet, la plate-forme Cortext analyse des sources très hétérogènes, fournies par ses utilisateurs (au sens où c'est l'utilisateur qui envoie le corpus à la plate-forme). Les licences d'utilisation ne sont pratiquement jamais acheminées avec le corpus. Quand bien même, comment automatiser le stockage des droits d'utilisation décrits par cette licence ? Nous sommes donc limités aux informations fournies par l'utilisateur, à qui nous demandons de décrire son corpus - en terme de format notamment , qui lui-même n'a pas forcément connaissance des droits d'usage des données qu'il a récolté.

Pour éviter la diffusion de données sur lesquelles nous n'aurions pas les droits, nous ne laissons à disposition du public que les résultats des analyses, qui sont agrégés en règle générale. Les données diffusées sont donc celles produites par la plate-forme (par exemple des statistiques temporelles sur le corpus ou des fichiers de données représentant un réseau d'auteurs). C'est un "pis-aller", qui remplit certaines conditions de la démarche mais est loin d'être satisfaisant pour l'instant.

Stockage

La qualité de la donnée passe par un accès adéquat qui est directement dépendant du mode de stockage de celle-ci. Il doit permettre la pérennité de la donnée et de ses méta-données tout en en facilitant l'accès et le traitement. Les corpus de textes peuvent vite devenir lourds en terme de taille, et l'ajout de méta-données n'arrange pas ce problème, bien au contraire (certains formats de méta-données doublent quasiment la taille du texte brut). Une plate-forme comme Cortext doit savoir gérer des dizaines de Téraoctets de données, qui doivent rester accessibles et correctement indexées. A moyen terme, la plate-forme devra se poser des questions du type de celles posées par les "big data" : même à des volumes bien moindres, les données sur lesquelles la plate-forme Cortext travaille sont en perpétuelle augmentation. Un des enjeux de Cortext est de pouvoir analyser des grand corpus de données textuelles (plusieurs millions d'articles). Les développements de la plate-forme vont dans ce sens depuis plus d'un an avec, entre autres, une refonte de notre système central de stockage. Nous essayons de tendre vers un système capable d'être complètement agnostique par rapport à la donnée, c'est à dire le moins dépendant possible à son format. Cet aspect nous paraît important dans la mesure où nous ne connaissons pas à l'avance les formats des corpus envoyés ni les traitements qu'ils supposent.

Il a été nécessaire de trouver un système générique de production de méta-données par les programmes eux-même lorsqu'ils les génèrent.

Nous avons également dû mettre au point un format de base de donnée unique pour tout type de corpus, qui permet de transformer les textes analyser en données tabulaires, mais en gardant au maximum la connexion à la source : il s'agit de stocker les données brutes directement après leur analyse syntaxique (ou "parsing"), avec une transformation minimale afin de conserver

un maximum d'information concernant le texte source, comme par exemple la ligne et la colonne auxquelles le mot a été trouvé, ou le nom du fichier duquel vient le texte dans le corpus.

D'autre part, notre système de stockage décorrèle l'emplacement de la donnée sur les espaces de stockage de l'ensemble d'informations liées à la donnée. Pour cela, notre première approche nous a amené à avoir deux systèmes distincts : l'un, optimisé pour l'accès (cf. paragraphe suivant), l'autre optimisé pour le stockage. Le stockage est organisé de façon que les données brutes soient toujours conservées intactes, assorties de méta-données fournies par l'utilisateur. Une base de donnée centrale permet de conserver la localisation exacte d'un corpus dans le système, qui potentiellement peut contenir de nombreux serveurs de stockage. D'autre part, lors du traitement, une base de données locale associée au corpus permet de stocker les informations issues du pré-traitement (analyse syntaxique, extractions de champs pour les données tabulaires, etc.).

Chaque ensemble [corpus/analyse/résultats] est conservé au même endroit, afin de permettre une reconstitution de l'intégrité des données en cas de défaillance ou de perte des informations centrales. Des méta-données sur l'analyse et sur le corpus y sont également stockées (auteur de l'analyse, paramètres, date, etc). Ainsi le système de stockage peut être à la fois distribué et autonome. Mais il est également accessible de l'extérieur comme nous allons le voir.

Accès

La notion d'accès aux données fait partie des fondamentaux de la qualité (voir plus haut). La qualité de la donnée textuelle n'existe pas si une norme d'accès n'a pas été produite et fournie en même temps que l'accès à la donnée. Un accès de qualité aux données passe par deux aspects : la facilité de localisation et de récupération de la donnée et la qualité des informations fournies avec celle-ci. La plate-forme Cortext a également planché sur ces aspects lors de la refonte de système central de gestion des données. Nous avons développé une API [3] de type RESTful [4] , qui permet l'accès par des systèmes externes aux données. Ces systèmes peuvent être par exemple des applications utilisateurs qui doivent avoir accès aux données pour les afficher ou les visualiser, comme des scripts de traitement locaux . Cette API permet une indépendance de l'accès aux données par rapport aux infrastructures (serveurs, stockage,...). De plus, et c'est un élément important de la qualité de données, elle permet la production d'informations standardisées à l'utilisateur (humain ou machine) à partir d'informations hétérogènes. En effet, si les don-

3. Application Programming Interface - ensemble de fonctions (ou méthodes) permettant d'accéder à un système depuis l'extérieur

4. REST/RESTful : REpresentational State Transfer, style d'architecture standardisé permettant à deux systèmes développés dans des standards différents de communiquer. Utilisé aujourd'hui par beaucoup de services web.

nées brutes sont “sales” et hétérogènes, parfois sans cohérence, les données calculées seront, elles, remises aux normes et fournies avec un format unique et exploitable. C'est un point essentiel dans le service rendu par la plate-forme.

Les applications développées à la plate-forme peuvent donc se baser sur ce format d'échange d'information sans être liées au système de stockage en interne ou encore à la localisation précise des données. Le cas échéant, un changement complet du système de stockage (pour passer par exemple d'un stockage centralisé à un stockage distribué) serait transparent pour les applications externes qui y avaient accès car elle auront exactement la même API présentée par le système. C'est un principe qui s'est démocratisé ces dernières années sur tous les grands services web modernes (Google, Twitter, ...) et qui commence à être adopté par les communautés scientifiques. Le courant de l'open data y a grandement contribué, et on voit fleurir de plus en plus d'API publiques (en France par exemple, le service “etalab”, chargé de la diffusion de données publiques, a prévu de développer en 2013 une API permettant l'accès aux données). L'API est donc un moyen moderne, indépendant des diverses modifications des systèmes, capable d'assurer la qualité d'accès aux données.

Enrichissement

La qualité d'un corpus ne se résume donc pas au corpus lui même mais bien à l'ensemble du système d'information qui l'entoure, permettant de fournir un service optimum à l'utilisateur des données. À ce titre, l'enrichissement des données ou des méta-données est un aspect important. Il peut s'agir d'améliorer les méta-données en y ajoutant par exemple une typologie, d'y adjoindre une mise à jour (un corpus d'articles sur un sujet précis évolue en permanence) ou tout simplement la suppression de certaines données. Le système conçu par la plate-forme Cortext permet, via son API, de compléter ou de supprimer un corpus donné, ce en préservant au maximum la qualité des informations. Cela est réalisé par un système de mise à jour des informations contextuelles du document, rendu possible par la décorrélation du document (ou de l'ensemble des documents) et de l'information contextuelle. Les informations sont conservées dans une base de donnée locale (cf. plus haut) qui est enrichie au fur et à mesure des étapes, et qui pourra être utilisée par la suite comme base pour d'autres traitements. Ces modifications sont rendues possibles par la normalisation des méta-données, un autre point fondamental de la qualité des données textuelles. En effet, différents modules de la plate-forme pourront y avoir accès, et en complétant ces méta-données, améliorer la qualité de celles-ci. Les modules d'analyse linguistique pourront par exemple adjoindre des mots-clé extraits des textes aux documents, ce qui enrichit l'information associé à celui-ci. On voit donc bien comment l'augmentation de la qualité globale des données passe par un enrichissement local des informations.

Conclusion

La qualité globale de la donnée textuelle passe par un ensemble d'aspects qui doivent être considérés séparément afin de garantir un service pérenne. La séparation de ces aspects ainsi que leur maintenance et leur documentation sont cruciales pour un maintien de ce service. Les solutions que nous avons trouvées lors de nos réflexion autour de la conception de notre système ne sont évidemment pas idéales mais, nous l'espérons, découlent de bonnes pratiques et principes d'une démarche qualité, et permettent d'être évolutifs par rapport aux nouveaux standards du web ("web de données", "web sémantique","big data"...) ainsi que des productions scientifiques, qui passent de plus en plus par ce média. La qualité des données est un débat qui évolue avec ces standards au sein de la plate-forme et qui nous guide dans l'évolution de nos développements. Les étapes à venir sont la production d'une documentation complète sur les formats acceptés et les traitements de la plate-forme, ainsi que sur l'accessibilité des services et des données de façon normalisée via une API, elle aussi documentée en détail.

À moyen terme, nous prévoyons également le développement d'un déploiement automatisé de la plate-forme et de ses modules sur des grilles de calculs (type cloud). Cela permettra l'ajustement à la volée des ressources en fonction des traitements, assurant une qualité de services en plus de la qualité des données produites. Il est souhaitable que les standards de qualité des données (textuelles ou non) passent par l'unification des standards du web, ce qui rendra le travail de normalisation plus automatisable. Cela demande une rigueur dans la production des données, ce qui dépend bien-sûr des volontés de ceux qui les produisent. La plate-forme en tout cas continuera de s'adapter à ces normes en tentant de fournir un service de qualité aux chercheurs.