# Une pratique d’écriture au 21e siècle

Interview avec Anne-Laure, statisticienne à l’Inami

**‘Chaque graph est comme un roman pour moi,’ elle a dit un jour.**

**‘Faire les graphiques, c’est des façons d’apercevoir le monde. A travers la ‘norme’, on cherche les patterns, et on vérifie ce qui dévie, on cherche les raisons. Parfois il y a des raisons qui sont tout à fait légitimes, parfois elles sont plus discutables... Pour que cela soit bien fait, il faudrait que tout le terrain soit couvert et que la juridiction soit explicite, mais la réalité est toujours un peu plus compliquée que la nomenclature !’**

Interview avec Anne-Laure, statisticienne.
Réalisé par An Mertens à Constant, Bruxelles, juin 2013.

## Quelques genres de romans dans l’analyse des données

### Le roman policier

Anne-Laure : Une collègue par exemple examine des prestataires qui facturent le Dossier Médical Global. Normalement, tu ne peux facturer que si tu as vu le patient, et maximum une fois par an. Pour un prestataire, on voit dans sa facturation qu’il facture pour des patients qu’il n’a jamais vu, et il a une systématique dans la facturation, mais on ne sait pas encore laquelle. Souvent notre travail est cela: on cherche comment le crime est commis. On joue aux devinettes dans les données, on cherche la vie qui est derrière. C’est comme quand tu as des traces sur une scène de crime : les données sont la trace et tu cherches à reconstituer l’histoire. Tu joues au chat et la souris avec les données, c’est ludique, tu es dans le jeu. C’est drôle à faire, j’aime beaucoup les romans policiers, et ça a un côté ‘je vais le coincer’. On joue au plus malin avec le prestataire.

En fait on cherche toujours à décoder les données, à trouver dans les données la logique d’une fraude, et si on a quelque chose qui cloche, on peut aller vérifier sur le terrain, auprès des patients, par exemple.

### Le roman de genèse

Anne-Laure : Dans des analyses de données plus complexes, le but, c’est  de rouvrir les données pour

leur donner du sens. J’aime mieux parce que l’histoire est un peu plus compliquée : tu peux le comparer au moment où quelqu’un meure et où tu gardes la mémoire de sa présence. Les données, c’est un peu comme la mémoire de la vie de quelqu’un. A partir des données tu peux réévoquer, réinvoquer l’histoire de la personne. Ça laisse plus de place à l’interprétation que la méthode du "roman policier" : ce que tu recrées est plus riche, plus complexe.

On a fait une analyse, par exemple, dans la gastroentérologie, dans ce cas-ci et on regarde de manière très ouverte s’il y a des choses bizarres. Pour y arriver tu es obligé de reconstituer le tout.  Les médecins qui sont spécialisés en oncologie, par exemple, n’ont pas du tout la même pratique que ceux qui ne sont pas oncologue. Pour un , c’est tout à fait normal de faire une gastroscopie par mois à certains patients, tandis que pour les autres, on s’attend plutôt à des gastroscopies répétées au plus une fois par an. C’est compliqué de comprendre les règles d’un domaine quand tu es externe à ce domaine. Il y a plein de choses dans la facturation pour lesquelles tu ne sais pas si c’est normal ou pas. Tu as toujours besoin de l’histoire autour, du contexte, d’un médecin qui a une vue du terrain, qui fait la traduction.

Un autre exemple, c’était la question de comment on peut diminuer les dépenses sur les médicaments avec l’autorisation spéciale du médecin conseil. Il y a eu une action sur certains groupes de médicaments, et notre rôle était de voir quel avait été l’effet de cette action sur la diminution des dépenses. Derrière chaque

groupe de médicaments qu'on a analysé, tu as des histoires différentes, qui sont aussi en relation avec les entreprises pharmaceutiques particulières.  Chaque médicament a ses particularités.

On a donc la trace, on voit comment l'action  faite a changé des choses, mais pour pouvoir interpréter les différences et vraiment comprendre ce qui se passe, on est obligé de retourner au contexte. Tu peux le comparer à la comptabilité. Tu ne peux pas comprendre les raisons des dépenses en regardant la facturation, mais à partir d'elle, tu peux tout ré-ouvrir, re-narrer. Comme si tu ouvres un livre et que chaque page est une anomalie que tu vois et que tu essaies de comprendre. A chaque page que tu tournes, tu vas plus profondément dans le questionnement. Ou encore, comme quand tu rencontres une nouvelle personne, au début tu as des impressions, puis tu vas poser des questions pour comprendre la famille, des idées, une histoire particulière. Cela induit des problèmes d'interprétations, it's *messy*.

Dans le cas de ces médicaments, on est resté avec beaucoup de questions : est-ce que les coûts baissent parce que les médicaments sont passés en générique ou parce qu'il y a moins de patients? Si les traitements sont pour des pathologies chroniques, comment faire diminuer un budget ? Comment mesurer une évolution mensuelle s'il y a des patients qui reçoivent des boîtes de médicaments de trois mois… ?

On n'a pas économisé ce qui était prévu, mais par rapport à ce qui aurait été dépensé si on n'avait rien

fait, on a beaucoup économisé...

Finalement pour cette question, on a formulé 9 méthodes de calcul, 9 points de vue possibles ! On était clair sur les paramètres possibles et on a dit : décidez quelle est votre stratégie. Cela renvoie la personne qui pose les questions à sa responsabilité : se rendre compte de la complexité de la question, du fait qu'il n'y a pas moyen de diminuer vite, par exemple, et qu'une diminution, c'est une question complexe, en soi ! Et puis, il faut voir cela dans le temps, à long terme.

## Méthodes

### An : Comment j'imagine ton travail dans le concret ?

Je reçois des extractions de données de la facturation des assurances de maladies. Ce que j'ai dans les données, c'est la date de l'acte, les numéros de l'acte dans la nomenclature des maladies (c'est assez précis, il y a des 20 000 codes de nomenclature, on peut le voir sur le site de l'INAMI), le numéro du prestataire, le numéro de registre national du patient, son nom, son adresse, le montant facturé, la date à laquelle l'acte est introduit dans la mutuelle, l'hospitalisation, le lieu. Il s'agit d'informations très comptables, mais très détaillées.

Je reçois les données brutes. On a accès à tout ce que nécessite le travail de contrôle. On a accès à la facturation d'un médecin, avec des patients identifiées pour pouvoir vérifier les déclarations du prestataire. Et à partir de ces données, on reconstitue des liens entre les choses. Il y en a qui habitent à la même adresse et qui ont le même nom de famille, c'est un lien. Il y a des patients qui reviennent chez leur médecin, c'est un lien. Il y a des prestataires qui partagent une patientèle, c'est un lien..

A partir de données complètement aplaties et détaillées, c'est à moi de faire les liens en fonction des variables: même prestataire ou pas, même prestation ou pas, même lieu ou pas, séquences d'événement etc...

**An : Est-ce qu'il y a moyen de voir le processus ? Est-ce qu'à la fin du processus on peut le lire dans les graphs ?**

Anne-Laure : Je me rends compte que je documente très peu les étapes, la manière dont mon travail s'oriente ! Heureusement, comme c'est tout en programmation, il existe une trace dans le code. La structure régulière du code en SAS pourrait être :

*Datastep* : tu élimines les données dont tu n'as pas besoin, tu les vérifies, tu regroupes des catégories, etc..

*Proc sql* : tu sélectionnes des données et tu en fais une somme ou un autre calcul

*Histogram* : pour la variable 'quantité' tu visualises les résultats de p.ex. l'année 2011 et/ou tu les compares avec le même calcul pour 2008

Si tu le fais bien, tu mets des titres, qui s'afficheront dans les graphiques que j'exporte dans un document word. Sinon, c'est impossible de te retrouver dans le code après, quelle graphique vient de quelle étape ! Mais c'est une habitude qu'on prend au fur et à mesure, celle de mieux documenter, de supprimer les parties de codes que l'on n'a pas utilisées, pour s'y retrouver même après quelques mois.

On travaille par étapes. On fait une étape, on teste, on le garde ou pas.

J'essaie de séparer dans le code la partie de la préparation des données (avec un Datastep stable) et

l’analyse des données, mais cela ne fonctionne pas comme ça au moment où tu travailles : on fait les deux en parallèle : code & contexte.

L’autre jour, une collègue me dit, les patients qui ont fait une analyse d’oncologie, je pense qu’il faut les enlever de l’analyse, parce que c’est des patients avec un profil particulier. Il faut donc changer les données de départ et éliminer tout ça. En plus, parfois je fais des calculs sur tous les patients, parfois sur une partie seulement. Souvent je ne sais plus très bien ce que j’ai fait, donc je fais tout tourner, comme ça je suis sure que je n’ai pas loupé une étape. Parfois ça tourne 1,5 h ou 2 h, parfois je change un petit truc et je dois tout refaire. Ceci dit, je ne fais ce cycle complet que 5 à 6 fois par an, pas plus.

Il faut du temps et il faut bien documenter ! En général, je ne comprends plus rien à mon code quand j’y reviens dessus un mois après. Heureusement, plus tu programmes, plus c’est systématiquement fait.

**An : Quels outils utilises-tu ?**

Anne-Laure : On travaille sur des bases de données très grosses, donc il faut un gros système (SAS) qui est compliqué à utiliser. Les données digitales elles-mêmes en général remontent à deux ou trois ans, mais parfois on travaille sur des données de la fin des années 90.

Puis, on publie les résultats dans un rapport rédigé, de 20 à 30 pages. J’y mets les questions, les hypothèses et puis les détails des analyses.

**An : Et les graphs.**

Anne-Laure : Les deux. Il y a des gens qui comprennent bien les graphiques et d'autres des tableaux. Personnellement, j'aime bien les squatter plot, qui utilisent 1 point par personne sur deux axes. Ils permettent plus de complexité, ils rendent les évolutions plus fines, plus visibles. Ils montrent la variabilité entre les individus. Parfois, on envoie un courrier aux prestataires, dans le genre de 'la limite normale serait 10 %, vous êtes à 15 %, il faudrait faire attention, ça coûte à l'assurance maladie...' En général, ils en tiennent compte. Avec un seul graphique, tu peux montrer comment chaque individu du groupe à modifié (ou non !) son comportement.

## Le point de vue de la chorale

**An : A partir de tes calculs tu trouves des points de vue précis, tu pourrais en choisir un, et ne pas laisser le choix à l'autre.... Est-ce que tu n'as pas envie d'appliquer tes connaissances dans d'autres domaines que l'Inami?**

Anne-Laure : Je l'ai fait à Garance, une association à Bruxelles qui lutte contre toutes les sortes de violences. J'ai travaillé sur les résultats de leurs questionnaires d'évaluation. Il y avait des questions ouvertes, c'est l'horreur en quantitatif parce que tu es obligée de recoder toutes les réponses. On avait 200 ou 300 réponses à 2 questions différentes. Quand tu as des questions ouvertes, tu a toutes les voix individuelles qui sont représentées, chacune avec leur différences, leur particularité, tandis que dans les

questions fermées, il y a 2 ou 3 voix possibles et tu résumes les réponses. Dans les questions ouvertes de Garance, c'était impossible.

Une question comme, 'Qu'est-ce qui était dur pendant le stage d'autodéfense?', amenait à des réponses qui dépendaient de l'expérience de chacune. Une des réponses était par exemple, 'taper quelqu'un en sachant qu'on peut lui casser la jambe, parce que je me rends compte que je peux faire mal'.

Il y avait de la richesse dans leurs questionnaires. Elles demandaient deux courtes réponses ouvertes à tout le monde, c'est à dire un petit peu à beaucoup. Dans leur cas, à mon avis, ça pourrait être mieux de demander beaucoup à un petit nombre, par exemple en faisant des interviews individuels avec quelques participantes, plutôt qu'un questionnaire pour tout le monde, parce que l'exploitation d'un questionnaire est forcément simplificatrice.. Mais c'est toujours des choix délicats, elles veulent aussi avoir des variables qui indiquent la satisfaction de chaque participante... d'où le choix d'avoir des questions ouvertes dans un questionnaire général quand-même.  Je n'ai presque jamais de questions ouvertes dans mon boulot.

**An : Des questions ouvertes ne permettent donc pas de narration collective?**

Anne-Laure : Si. Elles permettent la narration chorale, comme Svetlana Alexievitch le fait dans son livre *La Supplication* ou *Voices from Tsjernobil*. L'auteur a interviewé des centaines de personnes sur la tragédie de Tsjernobil, et elle en fait quelque chose qui n'est

plus journalistique, c'est chorale : tu entends toutes les voix de toutes les personnes, réécrites. Les perspectives, ce qui est grave ou pas, dépendent des expériences. Tu suis une personne ou l'autre, mais tu es plongée dans une situation collective.

**An : Tu travailles donc exclusivement avec le résultat de questionnaires personnelles à choix multiples, qui me posent souvent de grands problèmes... Comment ne pas remplir le case 'varia' ? L'exemple type : est-ce que ton genre c'est masculin ou féminin? Comment toi en tant que personne tellement sensible à ça, comment tu te positionnes par rapport à ça ?**

Anne-Laure : C'est compliqué. Il y a un livre très intéressant qui est écrit par rapport à cela, *Sorting things out* de Susan Leigh Star.

Ici il s'agit plutôt d'un problème de nomenclature. Tu retrouves toute la législation sur le site de l'INAMI et tu verras que beaucoup d'actes sont définis de façon floue. Les limites sont floues. Dès que tu as une faille dans la nomenclature, souvent une partie des prestataires en profitera…

L'autre problème, c'est le problème de l'identification. On décrit des événements dans la facturation et chaque individu doit avoir un numéro, sinon, on ne peut pas l'identifier. Quand les prestataires ou les patients n'ont pas de numéro, c'est le bordel. Par exemple, il y a des tabacologues qui n'ont pas de numéro parce qu'ils ne sont pas médecins, mais psychologues. Du coup, j'ai une catégorie 'fourre-tout'

qui dépense 200.000 euro avec des gens que je ne peux pas identifier, ça ne me va pas du tout !

Pour pouvoir être précis dans les calculs, il faudrait couvrir tout le terrain, numéroter, identifier tout le monde ! Et là, on touche à une discussion problématique, comparable à celle de la surveillance, qui a le même problème de législation.

L'assurance maladie génère un milliard de lignes de facturation par an ! On a donc beaucoup de données, d'une richesse énorme, mais qui sont sous-utilisées en analyse de données. Pour le moment, ces données sont beaucoup utilisées pour des analyses de dépenses, mais elles pourraient servir pour ouvrir des analyses plus complexes, et donner du sens et du contenu... Progressivement, les utilisateurs sont plus nombreux : le KCE (Kennis Centrum Expertise), l'IMA (Agence Intermutualiste) développent des analyses ciblées, des atlas, etc... mais c'est encore un work in progress !

**An : Tu travailles avec les données personnelles, nom, adresse, numéro de registre national... Qu'en est-il de la confidentialité ?**

Anne-Laure : On signe un contrat qu'on ne peut pas poser des questions qui ressortent du domaine personnel, parce que même quand le nom et l'adresse sont enlevés et que les données sont soi-disant anonymes, il y a toujours le numéro de registre national dans lequel tu retrouves la date de naissance, le genre, et qui peut permettre éventuellement d'identifier la personne. Je trouve qu'on n'est pas

encore assez vigilant sur ces questions de vie privée, mais que c'est en train de changer. La conscience de la manière dont on doit protéger les données que l'on utilise est plus présente maintenant.

**An : Et comment tu te positionnes vis-à-vis de la pratique de normalisation qui est quand-même la base de ton travail ? Ce qui est hors norme, saute aux yeux, est sujet de suspect.**

Anne-Laure : On est tout le temps confronté à la normalité, en effet. Mais cela ne veut pas dire que les exceptions sont toutes hors loi. Il y a quelques années, on avait regardé les montants de prescription moyenne en médicament par patient, pour les généralistes. Le médecin avec le plus de dépenses en Belgique, c'est quelqu'un qui a un patient hémophile, qui doit prendre beaucoup de médicaments qui sont très chers, donc il arrive à des budgets disproportionnées, qui paraissent anormaux, mais qui sont tout à fait légitimes !

Par ailleurs, notre travail est conditionné par le contexte. Les données et les recherches sont liées au système d'assurances maladies tel qu'il est constitué maintenant. Les données sont liées à l'acte, et non à des heures prestées par exemple,  comme en Angleterre où le service médical est un service gratuit. Ça a des désavantages, mais aussi des avantages : cela permet de 'connaître' aussi la réalité des actes médicaux faits.

**An : L'accès facile à des grandes bases de données de la part des institutions d'Etat donne à penser ces jours-ci, que les activités de**

**contrôle du NSA sont révélées. Qu'en penses-tu en tant qu'activiste ?**

Anne-Laure : J'ai gardé un article pour en parler lors des cours à l'ERG. Il s'agit d'une femme qui a cherché 'cocotte minute' sur Google, le même jour que son mari a cherché 'sac-à-dos'. Ils ont vu débarquer le FBI, qui fait une centaine de descentes pareilles à l'an. Bref, tu peux tout voir dans les données, mais quand tu prépares un attentat terroriste, tu ne vas pas chercher comment le faire dans Google, ni acheter tes outils par internet. Tu le fais avec une fausse carte d'identité dans un café internet, ou bien tu vas dans un magasin et tu payes en cash !

C'est-à-dire que les données ne sont pas toujours suffisantes, elles ne sont pas une photo complète de la réalité. Ceux qui n'échappent pas à la NSA ne sont pas des terroristes ! Les systèmes de contrôles veulent qu'il n'y ait plus de marge, mais il y aura toujours des zones non-couvertes.

Le grand problème pour moi, au niveau politique, plus que l'analyse des données, c'est le Patriot Act. La possibilité de retenir des gens sans preuve est beaucoup plus grave que la surveillance permanente des données. Mais peut être que c'est mon métier qui me déforme ?

D'un autre coté, j'adore l'Ecosse parce qu'il y a plein de régions non-couvertes, sans chemins d'accès, pas sur la carte. Et en même temps on crée toujours des espaces invisibles, aussi sur internet. Je tiens un blog avec un mot de passe dont personne ne connaît

l’adresse...

**An : Est-ce que tu as des fantaisies sur des données que tu aimerais bien traiter ? Ou des types de narration que tu pourrais développer avec cela ?**

Anne-Laure : Je crois que toutes les nouvelles données sont toujours excitantes. C’est comme quand tu ouvres un nouveau livre. Après, c’est vrai que tout ce qui est lié au quantitatif et qualitatif. J’ai une amie qui fait une thèse en psycho et quand je vois comment elle utilise l’analyse des textes, l’analyse des interviews, pour essayer de retrouver les sous-textes, de ce que racontent les  interviewés sur leur histoire personnelle... Tu peux des fois utiliser des techniques quantitatives pour révéler des choses qui ne sont pas forcément formulées de façon claire, parce qu’elles font des choix de mots qu’elles ne savent pas nécessairement à ce moment-là.

Quand j’étais monitrice des colonies de vacances, on faisait des rapports d’activités sur les adultes qu’on avait avec nous. Les gens écrivent ce qu’il croient être une description objective, mais quand tu la lis, ce n’est jamais une description objective. En fait, quand tu racontes ta vie, ce que tu dis, ce que tu ne dis pas, les mots que tu utilises, tu choisis une formulation parmi cent possibles. Quand je dis, par exemple, j’ai un chien, je ne dis pas comment il s’appelle, comment il est arrivé chez moi, pourquoi je le garde, mais je choisi de donner cette information-là. C’est donc que c’est important de partager cette information avec toi, ce

qui veut sans doute dire que mon chien est important, et que j'essaye de partager quelque chose avec toi...

Cette amie identifie des chaînes. Elle travaille avec des logiciels de travail de textes qui te permettent de reclasser les éléments du parcours de vie qu'ils racontent, en les recatégorisant. Mais aussi, à cet endroit là, la personne parle à la première personne..., après tu compares entre les personnes aussi.

Pour répondre à ta question, il y a un côté très messy, très bordélique dans l'analyse des données. Il y a toujours trop de données, jamais on n'arrive à mettre fin à ça... Donc, soit c'est arriver à faire une analyse de données sur des données très circonscrites, des choses toutes petites, ... où tu peux aller très dans le détail.

J'ai souvent un problème avec raconter des histoires parce qu'il y a question du vrai et du faux, de l'inversion et du subjectif. Comme statisticienne, j'essaie de faire remonter le maximum d'information de la réalité et d'être le plus juste possible sur l'aspect de la vérité. Je suis bonne pour trouver les histoires des autres et de les mettre en lumière, mais à condition que je sois totalement absente finalement de la narration. Ou que j'ai cette sensation, parce qu'évidemment ce n'est pas le cas. J'essaie de faire des choix scientifiques et pas personnels.

Et donc, ce serait un projet sur des données... de la vie privée, la vie à moi. Pas sur la vie de tous les jours, sur combien t'as marché, ou combien de courrier que tu as

lu... créer de l'information sur les choses à garder, par exemple, sur la mémoire, la trace des gens, des liens.

**An : Des souvenirs ?**

Anne-Laure : Oui.

**An : Comment tu ferais ça ?**

Anne-Laure : Je ne sais pas, c'est pour cela que je ne le fais pas d'ailleurs.

Il y a un côté d'objectivation. Quand j'étais ado, j'avais des cahiers et j'écrivais, j'écrivais... avec cette fonction-là, de garder la mémoire. Et en fait, ça n'a aucun sens. Les choses importantes tu ne les notes pas, parce qu'elles prennent toute la place et tu t'en rappelles, même vingt ans après. Et tu écris des trucs, genre, j'étais amoureuse de ce mec, mais vingt ans après, tu relis et tu te demandes : c'était qui ?

Mais par contre, ce qui faudrait c'est de se souvenir des choses précieuses.

Si tu ne fictionnes pas tes souvenirs, une fois que tu les auras oubliés, ils n'auront plus de valeur. Si tu décris objectivement, de façon neutre, un souvenir important au moment où tu le vis, finalement quand tu le relis dix ans après, cela n'évoque rien. Cela ne réveille aucune émotion.

Pereira Prétend de Tabucchi, par exemple, c'est un roman écrit comme une déposition. C'est prétendument objectif, neutre et policier. C'est une histoire qui se passe à Lisbonne pendant la période fasciste. Et Pereira prétend qu'il ne connaissait pas ce

monsieur qui a fait l'attentat, qu'il n'avait rien à voir avec lui, ... C'est construit à cette fin, mais avec la capacité d'écrire de la fiction. C'est soit disant neutre, c'est soit disant objectif, mais en fait, quand tu le lis, tu es totalement envahie par ce qui est écrit. C'est très sensible, très fort. C'est de la littérature.

J'ai écrit 3 poèmes dans ma vie qui sont très loin de la réalité mais qui par contre arrivent à évoquer la grande émotion que je vivais au moment où je les ai écrit, beaucoup plus que des tonnes et des tonnes de journaux intimes exhaustifs et détaillés !

**An : Et comment tu aborderais donc ce projet ?**

Anne-Laure : La question des mots et des données, c'est aussi une représentation différente de la réalité. Ce serait une base de données qui en soi serait inintéressante mais qui, quand on l'analyse, régénère l'histoire et l'émotion. C'est possible.  C'est remettre tous les faits, tous les côtés factuels, objectifs etc et que tu ne puisses pas ne pas re-raconter, revivre l'histoire quand tu l'analyses. J'avais pensé à un truc comme ça avec des timelines... Je ne pourrai pas écrire le roman, mais je pourrai faire la base de données qui permettrait de refaire le roman !

C'est une question de revoir tes comportements et ton être comme un résultat de choses extérieures (influence famille) et pas comme un choix, des choses dont tu découles et que tu peux te reconstituer, côté clonage peut-être aussi.

**An : On fait quand-même notre propre sélection des données.**

Anne-Laure : Oui, mais on n’a pas toutes les variables.

**An : Ce qu’on pourrait imaginer, c’est de choisir une situation particulière qui t’a marqué et aller à la recherche de toutes les variables que tu aurais pu avoir, de toutes les choses qui étaient à ce moment offertes, et tous les choix que tu aurais pu faire.**

Anne-Laure :  Et puis, ce que tu en as fait, qui a fait que tu t’es déplacée d’un point vers un autre.

Constant Verlag, Bruxelles (2013).