MAT 3777
Échantillonnage et sondages

Chapitre 5
Conception de questionnaires et collecte automatisée

P. Boily (uOttawa)

Session d'hiver – 2022

Aperçu

5.1 – Conception de questionnaires

> Personne n'aime être recensé, mais donnez-moi une page de profil et je passerai toute la journée à vous dire qui je suis.
>
> – Max Berry, *Lexicon*, 2013

Un **questionnaire** est une suite de questions visant à **obtenir de l'information sur un sujet** auprès de répondant.e.s.

Les principes de conception varient en fonction du **sujet** et du **mode de collecte des données**; il demeure prudent de tâter le terrain en testant une variété de questionnaires sur une **population pilote aléatoire**.

5.1.1 – Principes fondamentaux

En général, un questionnaire devrait:

- être aussi bref que possible, sans questions inutiles;
- être accompagné d'instructions claires et concises;
- garder les intérêts de la personne interrogée en tête;
- mettre l'accent sur la confidentialité;
- garder un ton sérieux et courtois;

- être exempt d'erreurs et présentés de manière attrayante;
- être formulé de façon claire et précise;
- être conçu de manière à ce qu'on puisse y répondre avec précision, et
- ordonné avec soin.

La qualité des données recueillies dépend en grande partie de la qualité du questionnaire – **c'est un aspect pratique de la discipline sur lequel on devrait passer beaucoup plus de temps que sur l'analyse des données**.

Les maisons de sondage réputées emploient des **équipes spécialisées** afin d'y veiller.

5.1.2 – Types de questions

L'unité de base du questionnaire est, bien entendu, la **question**, qui se présente sous deux formes:

- la question **fermée**, avec un nombre fixe de choix de réponses prédéterminés, mutuellement exclusifs, et collectivement exhaustifs (et qui devrait toujours inclure une catégorie "Autre (veuillez préciser)" afin de contrecarrer la perte d'expressivité), et

- la question **ouverte**, qui sert surtout à identifier les choix de réponses communs à utiliser dans les questions fermées d'un questionnaire ultérieur. Toute question fermée devrait avoir étée à un point une question ouverte.

Dans une conversation de tous les jours, les questions fermées sont contre-indiquées.

> Le fait de poser des questions ouvertes est un moyen sympathique d'aborder les autres dans le cadre de discussions. Connaitre la différence entre des questions ouvertes et des questions fermées vous sera d'une aide inappréciable pour votre carrière et votre vie sociale.
>
> – Comment poser des questions ouvertes, $WikiHow$

Dans un sondage, ce sont les question ouvertes qui le sont: les questions fermées demandent un **effort moindre** de la part des répondant.e.s, et elles sont en général plus **simple à quantifier**, ce qui permet de poser plus de questions pour un **laps de temps** et un **budget** donnés.

Question ouverte: Quel est le problème le plus important auquel le Canada fera face en 2022?

Question fermée: Lequel de ces problèmes représente le plus important défi pour le Canada en 2022?

- l'économie et le chômage
- les incidences de la COVID-19
- la réconcilliation avec les communautés indigènes
- les impôts
- le déficit budgétaire
- l'environnement
- le crime organisé
- la violence des gangs
- le racisme
- autre (veuillez préciser)

Cependant, les questions fermées peuvent aussi mener à

- la **perte d'une occasion de tâter le terrain** afin d'obtenir des précisions supplémentaires;
- l'**introduction d'un biais dans la réponse** en présentant des alternatives auxquelles les répondant.e.s n'auraient peut-être même pas pensé, et
- la **perte d'intérêt** si le choix des réponses ne correspond pas aux idées des répondant.e.s.

L'ajout de questions ouvertes au questionnaire peut réduire ces risques. L'utilisation de l'analyse textuelle et de méthodes de traitment de language naturel peut aussi permettre d'extraire les grandes lignes ou sentiments d'une réponse à une question ouverte (en anglais, surtout, et même là...).

5.1.3 – Considérations relatives à la formulation

Il est bien connu que la **formulation des questions** peut influencer les réponses d'un questionnaire; il est bon de garder les **considérations de formulation** suivantes en tête lors de l'élaboration de questionnaires:

- éviter les **abréviations** et le **jargon**: "Votre organisation utilise-t-elle des pratiques TTWQ?";
- éviter d'utiliser des **termes complexes** quand des termes plus simples font l'affaire: "Combien de fois avez-vous été défenestré?" vs "Combien de fois vous a-t-on jeté par la fenêtre?";
- veiller à ce que toutes les personnes interrogées puissent répondre aux questions, en posant des questions **pertinentes** et de **niveau approprié**;

- précisez le **cadre de référence**: "Quel est votre revenu annuel?" vs "Quel était le revenu total de votre ménage, toutes sources confondues, avant impôts et déductions, en 2017?";

- rendre la question aussi **précise** que possible: "Combien de carburant votre compagnie de déménagement a-t-elle utilisé l'an dernier?" (réponses reçues: 2500 litres, 800 gallons, $13,500, plus que l'année précédente, etc.) vs "Combien votre compagnie de déménagement a-t-elle dépensée en carburant l'an dernier?";

- éviter les questions à **double volet**: "Prévoyez-vous laisser votre voiture à la maison et prendre le train léger afin de vous rendre au travail?" vs "Prévoyez-vous laisser votre voiture à la maison? Si oui, prévoyez-vous prendre le train léger afin de vous rendre au travail?", et

- éviter les **questions tendancieuses**: le toujours excellent *Yes, Prime Minister* donne un exemple pas si facétieux que ça, en fin de compte.

S04xE02 – Leading Questions, *The Ministerial Broadcast*,
`https://www.youtube.com/watch?v=G0ZZJXw4MTA`

D'une part, Sir Humphrey démontre que le fait de poser des questions tendancieuses dans un ordre particulier peut inciter une personne interrogée à soutenir la ré-introduction du service national:

- Est-ce que le nombre de jeunes sans emploi vous inquiète?
- Est-ce que l'augmentation de la criminalité chez les adolescents vou inquiète?
- Pensez-vous qu'il y a un manque de discipline dans nos écoles polyvalentes?
- Pensez-vous que les jeunes apprécieraient un peu de leadership dans leur vie?
- Pensez-vous qu'ils répondraient à un défi?
- Seriez-vous en faveur de la ré-introduction du service national au Royaume-Uni?

Les cinq premières questions sont conçues et présentée de manière à susciter l'adhésion – la réponse évidente à chacune d'elles est "oui".

Après ce schéma de concordance, Sir Humphrey lance la question cruciale, formulée de telle sorte qu'elle propose le service national comme une solution supposée à tous les problèmes sus-mentionnés.

Dans la deuxième partie de de l'échange, Sir Humphrey démontre qu'une autre série de questions tendancieuses peut amener la personne interrogée à s'opposer à la ré-introduction du service national:

- Est-ce que le danger présenté par la guerre vous inquiète?
- Est-ce que la course aux armements vous inquiète?
- Pensez-vous qu'il soit dangereux d'armer les jeunes et de leur apprendre à tuer?
- Est-ce une bonne idée de forcer les gens à prendre les armes contre leur gré?
- Seriez-vous opposé à la ré-introduction du service national?

Les quatre premières questions de Sir Humphrey sont délibérément conçues pour produire un accord.

Conformément à la conception de l'enquête, la cinquième question fait de même: une personne qui répond "oui" à chacune de ces questions est forcément opposée à la ré-introduction du service national. [Selon une idée de Nagesh Belludi].

5.1.4 – Ordre des questions

L'**ordre** dans lequel les questions sont présentées est tout aussi important que leur formulation. Les questionnaires doivent être conçus de manière à se **dérouler sans heurts** et à **suivre une démarche logique** (pour la personne interrogée):

1. commencer avec une **introduction** qui fournit le titre, le sujet et l'objectif de l'enquête;

2. demander la **coopération** des répondant.e.s et expliquer l'importance de l'enquête et la manière dont les résultats seront utilisés;

3. indiquer le degré de **confidentialité** et fournir une date limite et une adresse de contact;

4. enchaîner avec une série de questions **faciles** et **intéressantes** afin d'établir la confiance des répondant.e.s;

5. grouper les questions semblables sous une **même rubrique**;

6. n'introduire les **sujets sensibles** que lorsque un rapport de confiance avec les répondant.e.s est susceptibles de s'être développé;

7. laisser un peu d'espace et/ou de temps pour les **commentaires supplémentaires**, et

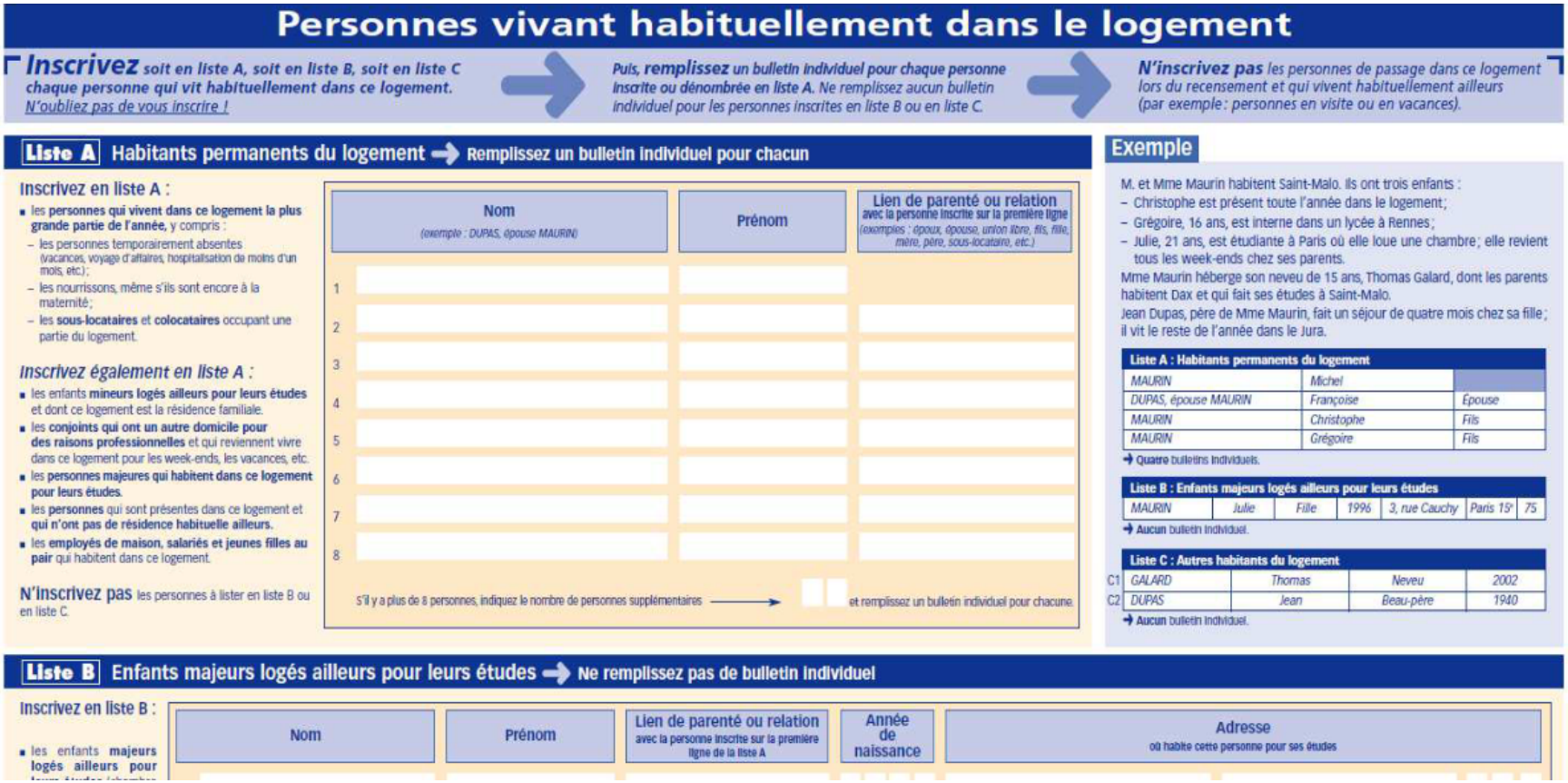
8. **remercier** les répondant.e.s de leur participation.

De nombreux ouvrages discutent du design des questionnaires, comme:

- Hidiroglou, M., Drew, J. and Gray, G. [1993], "A Framework for Measuring and Reducing Non-Response in Surveys," *Survey Methodology*, v.19, n.1, pp.81-94
- Gower, A. [1994], "Questionnaire Design for Business Surveys," *Survey Methodology*, v.20, n.2, pp.125-136
- *Méthodes et pratiques d'enquête*, Statistique Canada, catalogue 12-587-X

Il est bon de se rappeller que sans **plan d'échantillonnage solide**, les données recueillies, quelles qu'elles soient, peuvent être de telle piètre qualité qu'il devient impossible d'en tirer des conclusions exploitables.

Il est également essentiel de capter les **renseignement démographiques** permettant la classification des unités en **strates** (chapitre 3) ou en **grappes** (chapitre 6).



Personnes vivant habituellement dans le logement

Inscrivez soit en liste A, soit en liste B, soit en liste C chaque personne qui vit habituellement dans ce logement. N'oubliez pas de vous inscrire !

Puis, **remplissez** un bulletin individuel pour chaque personne inscrite ou dénombrée en liste A. Ne remplissez aucun bulletin individuel pour les personnes inscrites en liste B ou en liste C.

N'inscrivez pas les personnes de passage dans ce logement lors du recensement et qui vivent habituellement ailleurs (par exemple : personnes en visite ou en vacances).

Liste A Habitants permanents du logement → Remplissez un bulletin individuel pour chacun

Inscrivez en liste A :

- les **personnes qui vivent dans ce logement la plus grande partie de l'année,** y compris :
  - les personnes temporairement absentes (vacances, voyage d'affaires, hospitalisation de moins d'un mois, etc.) ;
  - les nourrissons, même s'ils sont encore à la maternité ;
  - les **sous-locataires** et **colocataires** occupant une partie du logement.

Inscrivez également en liste A :

- les enfants **mineurs logés ailleurs pour leurs études** et dont ce logement est la résidence familiale.
- les **conjoints qui ont un autre domicile pour des raisons professionnelles** et qui reviennent vivre dans ce logement pour les week-ends, les vacances, etc.
- les **personnes majeures qui habitent dans ce logement pour leurs études.**
- les **personnes** qui sont présentes dans ce logement et **qui n'ont pas de résidence habituelle ailleurs.**
- les **employés de maison, salariés et jeunes filles au pair** qui habitent dans ce logement.

N'inscrivez pas les personnes à lister en liste B ou en liste C.

	Nom (exemple : DUPAS, épouse MAURIN)	Prénom	Lien de parenté ou relation avec la personne inscrite sur la première ligne (exemples : époux, épouse, union libre, fils, fille, mère, père, sous-locataire, etc.)
1			
2			
3			
4			
5			
6			
7			
8			

S'il y a plus de 8 personnes, indiquez le nombre de personnes supplémentaires → et remplissez un bulletin individuel pour chacune.

Exemple

M. et Mme Maurin habitent Saint-Malo. Ils ont trois enfants :
- Christophe est présent toute l'année dans le logement ;
- Grégoire, 16 ans, est interne dans un lycée à Rennes ;
- Julie, 21 ans, est étudiante à Paris où elle loue une chambre ; elle revient tous les week-ends chez ses parents.

Mme Maurin héberge son neveu de 15 ans, Thomas Galard, dont les parents habitent Dax et qui fait ses études à Saint-Malo.
Jean Dupas, père de Mme Maurin, fait un séjour de quatre mois chez sa fille ; il vit le reste de l'année dans le Jura.

Liste A : Habitants permanents du logement

MAURIN	Michel	
DUPAS, épouse MAURIN	Françoise	Épouse
MAURIN	Christophe	Fils
MAURIN	Grégoire	Fils

→ **Quatre** bulletins individuels.

Liste B : Enfants majeurs logés ailleurs pour leurs études

MAURIN	Julie	Fille	1996	3, rue Cauchy	Paris 15e	75

→ **Aucun** bulletin individuel.

Liste C : Autres habitants du logement

C1	GALARD	Thomas	Neveu	2002
C2	DUPAS	Jean	Beau-père	1940

→ **Aucun** bulletin individuel.

Liste B Enfants majeurs logés ailleurs pour leurs études → Ne remplissez pas de bulletin individuel

Inscrivez en liste B :

- les enfants **majeurs logés ailleurs pour**

Nom	Prénom	Lien de parenté ou relation avec la personne inscrite sur la première ligne de la liste A	Année de naissance	Adresse où habite cette personne pour ses études

Première page du formulaire de recensement de la France, 2018

Recensement de 2021 – Comment puis-je remplir le questionnaire?
https://www.youtube.com/watch?v=2o09PWizNPw

Retranscription du vidéo

En mai, votre ménage recevra une lettre vous invitant à remplir le questionnaire du Recensement de 2021. Vous trouverez dans cette lettre un code d'accès sécurisé qui vous permettra de remplir le questionnaire en ligne. Une fois en ligne, vous pourrez remplir le questionnaire en trois étapes simples. Il vous suffit d'ouvrir une session à l'aide de votre code d'accès sécurisé, de remplir le questionnaire et de sélectionner "Soumettre". Si vous avez besoin d'aide ou d'une version papier, veuillez communiquer avec l'Assistance téléphonique du recensement. Pour obtenir plus de renseignements ou pour remplir le questionnaire du Recensement de 2021, visitez le `recensement.gc.ca`. C'est sécuritaire, rapide et simple.

Message du Statisticien en chef du Canada

Je vous remercie de prendre quelques minutes pour participer au Recensement de 2021. Les renseignements que vous fournissez sont convertis en statistiques pour que les collectivités [...] et les gouvernements planifient des services et prennent des décisions éclairées relatives à l'emploi, à l'éducation, aux soins de santé et au développement du marché.

Vos réponses sont recueillies en vertu de la Loi sur la statistique et gardées de façon strictement confidentielle. Selon la loi, tous les ménages doivent remplir un questionnaire du Recensement de la population de 2021. Les exploitants agricoles doivent également remplir un questionnaire du Recensement de l'agriculture.

Statistique Canada utilise les sources d'information existantes telles que les données sur l'immigration, l'impôt sur le revenu et les avantages sociaux pour alléger le plus possible le fardeau imposé aux ménages.

Il se peut que Statistique Canada utilise vos renseignements à d'autres fins statistiques et de recherche ou qu'il les combine avec ceux provenant d'autres enquêtes ou sources de données administratives.

Soyez du nombre dans le portrait statistique du Canada et remplissez votre questionnaire du recensement dès aujourd'hui.

Merci,

Anil Arora, Statisticien en chef du Canada

Questionnaire de satisfaction

Dans le cadre de notre politique de qualité et d'amélioration continue, nous souhaiterions recueillir votre avis. Accordez-nous quelques minutes pour répondre à ce questionnaire.

Adresse email _____ **Nom et prénom** _____

N° de téléphone _____ **Age** _____

Quelle est votre appréciation sur chacun des critères suivants concernant la qualité de la restauration :

Les qualités gustatives	Mauvais	◯	◯	◯	◯	◯	Très bien
La quantité servie	Mauvais	◯	◯	◯	◯	◯	Très bien
La variétés des plats	Mauvais	◯	◯	◯	◯	◯	Très bien
Le rapport qualité/prix	Mauvais	◯	◯	◯	◯	◯	Très bien
La présentation des plats	Mauvais	◯	◯	◯	◯	◯	Très bien
La carte des vins	Mauvais	◯	◯	◯	◯	◯	Très bien

Concernant le personnel, quelle est votre appréciation sur chacun des critères suivants :

Courtoisie	Mauvais	◯	◯	◯	◯	◯	Très bien
Efficacité	Mauvais	◯	◯	◯	◯	◯	Très bien
Apparence globale	Mauvais	◯	◯	◯	◯	◯	Très bien
Professionnalisme	Mauvais	◯	◯	◯	◯	◯	Très bien

Quelle est votre appréciation sur chacun des critères suivants concernant le service :

Attente et rapidité de la prise en charge	Mauvais	◯	◯	◯	◯	◯	Très bien
Amabilité et écoute	Mauvais	◯	◯	◯	◯	◯	Très bien
Accueil	Mauvais	◯	◯	◯	◯	◯	Très bien
Qualité des renseignements	Mauvais	◯	◯	◯	◯	◯	Très bien

Avez-vous des remarques ou des suggestions ?

_____

_____

5.2 – Collecte automatisée

> On dit que les rues du Web sont pavées de données qui n'attendent que la collecte, mais la quantité de déchets qu'on y retrouve est ... surprenante.
>
> – Patrick Boily, 2020

Les façons dont les données sont **partagées**, **recueillies** et **publiées** ont changé récemment en raison de l'omniprésence de la **toile** (WWW).

Les **entreprises privées**, les **gouvernements** et les **utilisateurs individuels** publient et partagent toutes sortes de données et d'informations. À chaque instant, des mécanismes engendrent de grandes quantités de données.

La rareté et l'inaccessibilité des données a longtemps constitué un problème pour les chercheurs et les décideurs. Ce n'est **définitivement** plus le cas.

L'abondance des données pose toutefois certains problèmes, notamment

- des masses de données **enchevêtrées**, et
- des méthodes traditionnelles de **collecte et d'analyse de données** qui ne sont plus à la hauteur en raison de leur **manque d'efficacité**.

La popularité et la puissance croissantes des **logiciels libres**, tels que R et Python (dont le code source peut être inspecté, modifié, et amélioré par quiconque), rendent la collecte automatisée de données très attrayante.

⚠ Les modules et les librairies de code deviennent **désuets** en un clin d'œil.

Si l'analyste est incapable (ou refuse) de **maintenir leur programme d'extraction et d'analyse** et de **surveiller les sites** desquels les données sont extraites, le choix du logiciel ne fera pas une grande différence.

Alors pourquoi prendre la peine d'automatiser la collecte des données? Voici quelques considérations courantes:

- la faiblesse des ressources financières;
- le manque de temps ou le désir de recueillir les données manuellement;
- le désir de travailler avec des sources de données actualisées, et
- la nécessité de documenter le processus analytique du début (collection) jusqu'à la fin (publication).

La collecte **manuelle**, en revanche, tend à être

- **encombrante** et **sujette aux erreurs**;

les approches **non reproductibles** sont de plus propices à la "**mort par l'ennui**", alors que les **solutions programmatiques** sont généralement plus

- **fiables**,
- **reproductibles**,
- **rapides**, et
- produisent des données de **meilleure qualité** (en supposant que des données cohérente existent au départ).

5.2.1 – Liste de contrôle pour la collecte automatisée

Le **raclage de la toile** (“web scraping”) n’est pas toujours recommandé.

En premier lieu, il est possible qu’aucune source de données en ligne et librement accessible **ne réponde aux besoins de l’analyse**, auquel cas on devrait privilégier une **approche d’échantillonnage**.

Cependant, si la réponse à la plupart des questions suivantes est positive, une approche automatisée peut s’avérer être un choix judicieux.

- Est-il nécessaire de répéter la tâche de façon périodique (par exemple pour mettre à jour une base de données)?
- Des sources de données en ligne sont-elles fréquemment utilisées?

- Est-il nécessaire que d'autres analystes soient en mesure de reproduire le processus de collecte?
- La tâche est-elle non triviale en termes de portée et de complexité?
- Si la tâche peut être effectuée manuellement, les ressources financières nécessaires manquent-elles?

L'**objectif** est simple: la collecte automatique de données devrait permettre d'obtenir **des données non structurées à un prix raisonnable**.

Et il n'est pas dit que la collecte automatisée doit demeurer **incompatible** avec l'échantillonnage probabiliste (cf. EUVC), mais on doit tenir compte des remarques du chapitre 1.

5.2.2 – Considérations d'ordre éthique

Question: les données disponibles en ligne sont-elles vraiment libres?

Un **spider** est un programme qui parcourt rapidement le web à la recherche de données. Il saute d'une page à l'autre, en s'emparant de tout leur contenu.

Le **raclage** ("scraping") consiste à recueillir des renseignements spécifiques sur des sites spécifiques – en quoi ces deux concepts sont-ils différents?

> Le raclage implique intrinsèquement la **copie** d'information; l'une des revendications contre le raclage est la violation des droits d'auteur.
>
> – Munzert, Rubba, Meissner, Nyhuis.

Que peut-on faire pour minimiser le risque?

- travailler de manière aussi transparente que possible;
- documenter les sources de données à tout moment;
- remettre le mérite à ceux qui ont collecté et publié les données au départ;
- demander l'autorisation de reproduire les informations (si vous ne les avez pas recueillies) et, surtout
- ne commettre aucun acte illégal.

Les tribunaux n'ont pas encore trouvé leur rythme dans ce dossier (consulter, par exemple, *eBay* vs *Bidder's Edge*, *Associated Press* vs *Meltwater*, *Facebook* vs *Pete Warden*, etc.).

Il y a plusieurs questions juridiques à étudier, mais il semble en général que les **grandes entreprises/organisations** sortent généralement victorieuses de ces batailles légales.

La question est floue par ce qu'il n'est pas évident de différentier les actions de grattage illégales de celles qui sont légales.

Il y a des lignes directrices approximatives: la **re-publication de contenu à des fins commerciales** est considérée plus problématique que le **téléchargement de pages pour la recherche et l'analyse**, par exemple.

Le fichier `robots.txt` ("Robots Exclusion Protocol") de chaque site indique aux gratteurs quelles informations peuvent être recueillies sur un site web avec le consentement de ses auteurs – au minimum, il faut en tenir compte... **quoique cela n'offre pas une protection absolue**.

Un bon programme de grattage doit

1. se comporter "convenablement";
2. fournir des données utiles, et
3. être efficace.

En cas de doute, contactez les propriétaires du site afin de vérifier s'ils accordent l'accès aux bases de données ou aux fichiers.

```
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
# Files
Disallow: /CHANGELOG.txt
Disallow: /cron.php
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /INSTALL.sqlite.txt
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /LICENSE.txt
Disallow: /MAINTAINERS.txt
Disallow: /update.php
```

Fichier `robots.txt` du site `cqads.carleton.ca` (extrait).

Règles de bienséance du grattage:

1. **demeurer identifiable**;

2. **réduire le trafic** – accepter les fichiers comprimés, vérifier qu'un fichier a été modifié avant d'y accéder à nouveau, etc.;

3. **ne pas déranger le serveur avec des requêtes multiples** – de nombreuses requêtes par seconde peuvent entraîner des pannes de serveur, ce qui peut mener les webmestres à vous bloquer si votre gratteur est trop gourmand (quelques requêtes par seconde suffisent);

4. **écrire des grattoirs efficaces et polis** – il n'y a aucune raison de gratter les pages quotidiennement ou de répéter la même tâche sans cesse. . . il est préférable de sélectionner des ressources spécifiques et de laisser le reste intact.

5.2.3 – Qualité des données de la toile

Le question de la **qualité des données** est incontournable.

Il n'est pas rare de voir des organisations dépenser des milliers de dollars selon une stratégie de collecte de données (automatique ou manuelle) mal concue, pour ensuite insister que leurs analystes se servent de données défectueuses **puisque ce sont les seules données disponibles**.

Ce problème peut être escamoté dans une certaine mesure lorsque les analystes participent au **design du plan d'échantillonnage** et de la **stratégie de collecte des données**.

En particulier, on doit pouvoir répondre aux questions suivantes:

- quel **type de données** est le mieux adapté pour répondre aux question de l'organisation?
- les données disponibles sont-elles de **qualité suffisante** pour donner des reponses utiles aux questions du client/organisation/etc.?
- les informations disponibles sont-elles **systématiquement erronées**?

Sur la toile, les données peuvent provenir de **sources directes** (un tweet ou un article d'actualité), ou de **sources indirectes** (copiées d'une source hors ligne/grattées à partir d'un autre site, ce qui rend le traçage difficile).

Le **recoupement de données** ("cross-referencing") est une pratique courante lorsque l'on compose avec des données secondaires.

La qualité des données dépend également de leur **usage** et des **objectifs** d'analyse – le **contexte** de la collecte joue également un rôle.

Un échantillon de "tweets" recueilli un jour quelconque peut être utilisé afin d'analyser l'utilisation de "hashtags" spécifiques, mais cet ensemble de données peut s'avérer pratiquement inutile si l'échantillon est recueilli le jour d'une élection fédérale (en raison du **bias de collecte**).

Que sont les **pièges** et les **défis**?

Exemple: un client commandite une enquête téléphonique afin d'apprendre ce que les gens pensent d'une nouvelle éplucheuse de patate.

Une telle approche comporte un certain nombre de risques:

- **échantillon non représentatif** – l'échantillon sélectionné peut ne pas représenter la population visée;
- **non-réponse systématique** – les gens qui n'aiment pas les enquêtes téléphoniques peuvent être moins (ou plus) susceptibles d'aimer la nouvelle éplucheuse;
- **erreur de couverture** – les gens sans ligne téléphonique fixe ne sont pas rejoignables, et
- **erreur de mesure** – les questions de l'enquête peuvent ne pas fournir l'information requise.

Les solutions classiques à ces problèmes nécessitent le recours à l'**échantillonnage probabiliste**, à l'**élboration de questionnaires**, aux **enquêtes omnibus**, à des **systèmes de récompense**, etc.

Ces solutions peuvent s'avérer **coûteuses** et **inefficaces**, quoiqu'il demeure difficile (voir même impossible) de **cerner l'erreur de sondage** autrement.

L'utilisation de "**proxies**" peut aussi être utile – il s'agit d'indicateurs fortement liés à la popularité d'un produit, telles les statistiques de vente sur un site web commercial.

Le **classement des éplucheuses** sur `Amazon.ca` (ou un site web similaire) peut dresser un portrait bien plus complet du marché des éplucheuses que ne pourrait le ferait une enquête traditionnelle (en supposant, bien sûr, que l'on fasse confiance à ce dernier).

L'information recherchée pourrait donc être obtenue en élaborant un gratteur compatible avec l'**interface de programmation** (API) d'Amazon afin de recueillir les données appropriées – mais il peut y avoir des problèmes:

- **représentativité** des **produits listés** – est-ce que les éplucheuses sont toutes répertoriées? Si ce n'est pas le cas, est-ce parce que ce site web ne les vend pas? Y a-t-il une autre raison?

- **représentativité** des **clients** – y a-t-il des groupes spécifiques qui achètent (ou non) de produits en ligne? Y a-t-il des groupes spécifiques qui achètent sur des sites spécifiques? Y a-t-il des groupes spécifiques qui laissent (ou non) des critiques de produits?

- **fiabilité** des clients et des critiques – comment distingue-t-on les fausses critques des critiques réelles?

Le scraping est généralement bien adapté à la collecte de données sur les produits, mais il existe plusieurs situations pour lesquelles il est nettement plus difficile d'imaginer où trouver des données en ligne:

- quelles données pourriez-vous utiliser afin de déterminer la popularité d'une politique gouvernementale?
- quelles données pourriez-vous utiliser afin de déterminer la consommation d'essence journalière des canadien.ne.s?
- quelles données pourriez-vous utiliser afin de déterminer le salaire moyen des scientifiques des données au pays?

⚠ La facilité avec laquelle on peut avoir accès à certaines données en ligne n'offre pas de garantie au sujet de leur utilité.

5.2.4 – Technologies du web: premiers pas

En ligne, les données se retrouvent sous forme de **textes**, de **tableaux**, de **listes**, de **liens**, et autres structures, mais elles ne sont pas présentées dans les fureteurs de la même manière qu'elles sont stockées en format HTML/XML.

De plus, lorsque les pages web sont **dynamiques**, il y a un "coût" associé à la collecte automatisée.

Par conséquent, une connaissance de base du web et de ses technologies est cruciale. On peut facilement trouver des renseignements à ce sujet sur la toile et dans les bouquins de Munzert, Rubba, Meissner, Nyhuis [*Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*] et Mitchell [*Web Scraping with Python*], entre autres.

Il existe trois domaines d'importance pour la collecte de données sur le web:

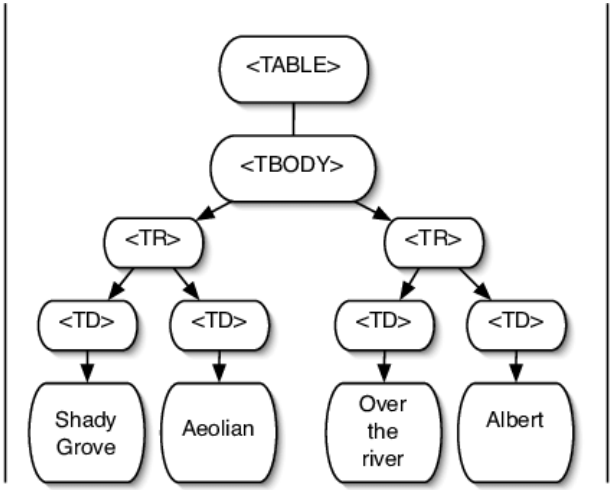
- les technologies de **diffusion de contenu** (HTTP, HTML/XML, JSON, texte brut, etc.);
- les technologies d'**extraction d'information** (Python, R, XPath, parser JSON, Beautiful Soup, Selenium, regexps, etc.), et
- les technologies de **stockage des données** (R, Python, SQL, formats binaires, formats de texte brut, etc.).

Le contenu d'une page Web se répartit en trois grandes catégories: le **langage de balisage hypertexte** (HTML; contenu et code web), les **feuilles de style en cascade** (CSS; style des pages web), et le **JavaScript** (JS; interactivité avec la page web).

5.2.5 – Boîte à outils de grattage de la toile

En quelque sorte, la partie HTML est la **plus fondamentale**; c'est en comprenant la **structure arborescente** des documents HTML, par exemple, qu'on apprend à utiliser pleinement la **boîte à outils de grattage**.

```
<table border=1>
  <tbody>
    <tr>
      <td> Shady Grove </td>
      <td> Aeolian </td>
    </tr>
    <tr>
      <td> Over the river </td>
      <td> Albert </td>
    </tr>
  </tbody>
</table>
```



Shady Grove	Aeolian
Over the river	Albert

(Rosiello, Kirda, Kruegel, Ferrandi)

Par expérience, un certain nombre d'outils peuvent faciliter le processus de collecte automatisé des données, notamment

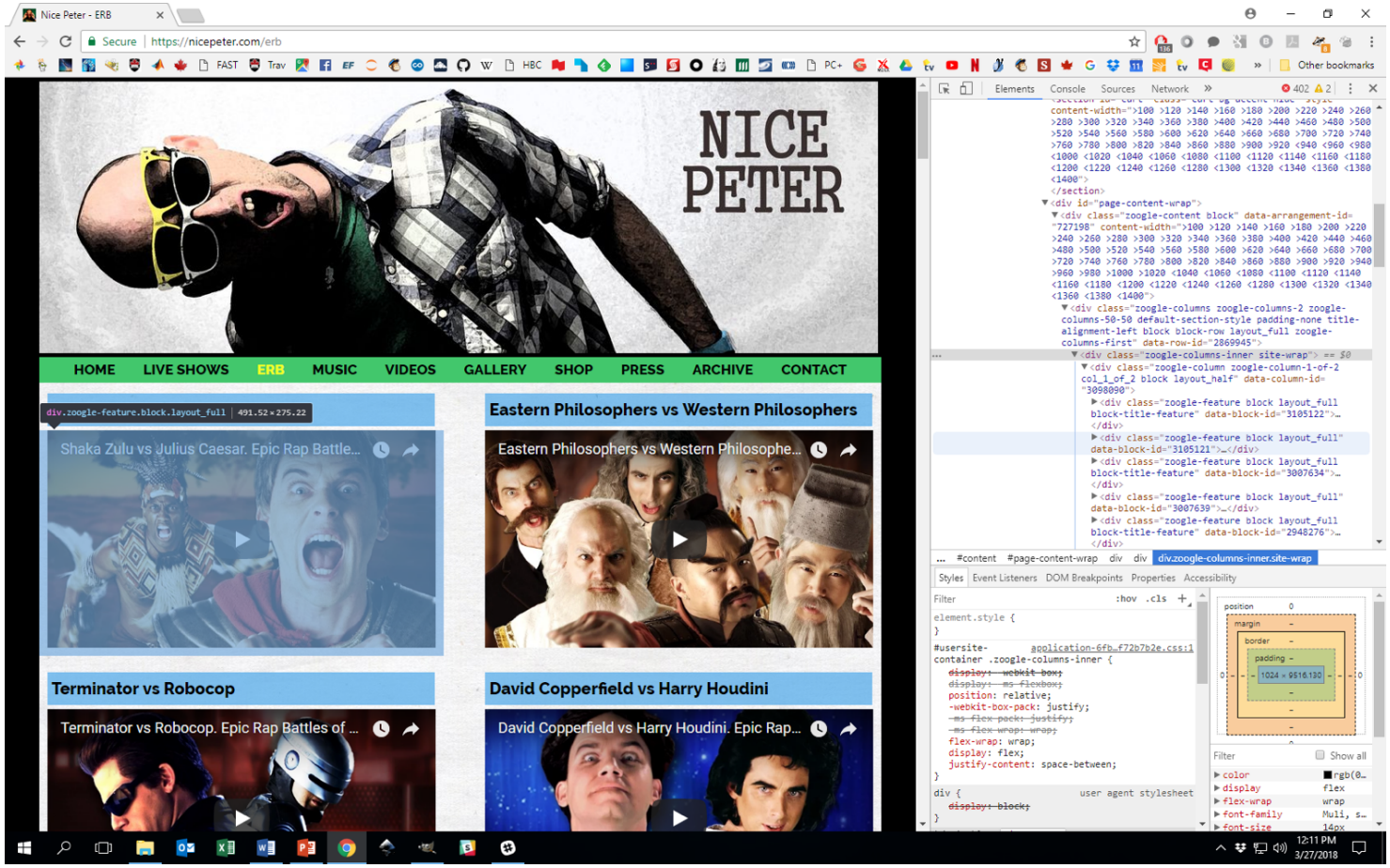
- les outils de développement ("Developer Tools"),
- XPath,
- Beautiful Soup,
- Selenium, et
- les expressions régulières ("regexps").

Mais les technologies changent du jour au lendemain, et la boîte à outil doit rester à jour.

Les **outils de développement** affichent la correspondance entre le code HTML d'une page et la version présentée par le navigateur.

Contrairement à l'option "View Source", les outils de développement affichent la version **dynamique** du contenu HTML (c'est-à-dire que le code HTML est affiché avec toutes les modifications apportées par JavaScript depuis la première réception de la page).

L'inspection des différents éléments d'une page et la découverte de leur emplacement dans le fichier HTML est une étape **cruciale** afin d'obtenir un grattage efficace – prenez la peine d'explorer le site avant de commencer à écrire du code pour le "scraper".



Inspection des éléments de la page YouTube `https://nicepeter.com/erb` à l'aide des *Developer Tools* de Chrome.

Nous ne dirons ici que:

- **XPath** est un langage de requête que l'on utilise afin de sélectionner des informations spécifiques dans des documents balisés tels que HTML, XML, etc. Les requêtes XPath nécessitent à la fois un **chemin** et un **document** à rechercher.

- On peut utiliser les **expressions régulières** ("regexps") afin d'extraire des informations pertinentes parmi une multitude de données, dans lesquelles se cachent des **éléments systématiques** qui peuvent être employés afin de faciliter le processus d'automatisation, surtout si des méthodes quantitatives seront éventuellement appliquées aux données raclées.

  Les structures systématiques comprennent des **numéros**, des **noms** (pays, etc.), des **adresses** (courrier, e-mail, URL, etc.), des **chaînes de caractères spécifiques**, etc.

- **Beautiful Soup** est un module Python qui permet d'extraire des données de fichiers HTML et XML.

  Beautiful Soup ne se contente pas de convertir le mauvais HTML en code X/HTML valide; il permet également à un utilisateur d'inspecter la structure HTML (corrigée) qu'il produit dans son ensemble, de manière programmatique.

  La **soupe** qui en résulte est une API qui permet de **parcourir**, **rechercher**, et **lire** les éléments du document.

  Elle fournit essentiellement des moyens de navigation, de recherche et de modification **idiomatique** de l'arbre d'analyse du fichier HTML, ce qui permet de gagner un temps considérable.

- **Selenium** est un outil Python utilisé pour automatiser les interactions avec des fureteurs.

  Il permet à l'utilisateur d'ouvrir un navigateur et d'agir "naturellement", c'est-à-dire comme le ferait un être humain:

  – en cliquant sur les boutons;
  – en saisissant des informations dans les formulaires;
  – en recherchant des informations spécifiques sur une page, etc.

  Selenium contrôle automatiquement un navigateur, y compris le "rendering" des documents web et l'exécution de code JavaScript.

  Selenium peut programmer des actions comme "cliquez sur ce bouton" ou "tapez ce texte" pour donner accès au HTML dynamique ou à l'état actuel de la page (mais le processus peut désormais être entièrement automatisé).

Résumé du **processus décisionnel relatif à la collecte automatisée**:

1. Il faut bien comprendre **quel type d'information le client requiert**, que ce soit **spécifique** (le PIB des pays de l'OPEC au cours des 10 dernières années, les ventes des 10 premières marques de thé en 2017, etc.) ou **vague** (l'opinion des gens sur la marque de thé X, etc.).

2. Il faut prendre la peine de **découvrir s'il existe des sources de données web qui pourraient fournir des informations directes ou indirectes sur le problème** – il est plus facile d'y parvenir pour des faits spécifiques (la page web d'un magasin de thé fournira des informations sur les thés en demande, par exemple) que pour des faits vagues.

   Les tweets et les plateformes de médias sociaux peuvent contenir de l'information au sujet des tendances d'opinion; les plateformes commerciales sur la satisfaction relative à un produit spécifique, etc.

3. Il peut s'avérer utile de **développer une théorie du processus de génération de données lors de l'examen des sources de données potentielles** – Quand les données ont-elles été générées? Quand ont-elles été téléchargées sur la toile? Qui les a téléchargé? Y a-t-il des aspects qui ne sont pas couverts, cohérents ou précis? (etc.).

4. N'oubliez pas de **peser le pour ou le contre des sources de données potentielles** – lors de la validation de la qualité des données, peut-on trouver d'autres sources indépendantes qui fournissent des informations similaires à recouper, ou la source originale des données secondaires?

5. Finalement, on doit **prendre une décision** – choisissez les sources de données qui vous semblent les plus appropriées et justifiez les raisons de cette décision.