



Séance 2: La collecte de données

BIO 500 - Méthodes en écologie computationnelle

Dominique Gravel
Laboratoire d'écologie intégrative



Séance 2

- ✓ Ces diapositives sont disponibles en **version web** et en **PDF**.
- ✓ L'ensemble du matériel de cours est disponible sur la page du portail **moodle**.

Projet de session

Rappel du problème

Est-ce que le réseau de collaboration entre les étudiants est différent des réseaux écologiques ?

Rappel du problème

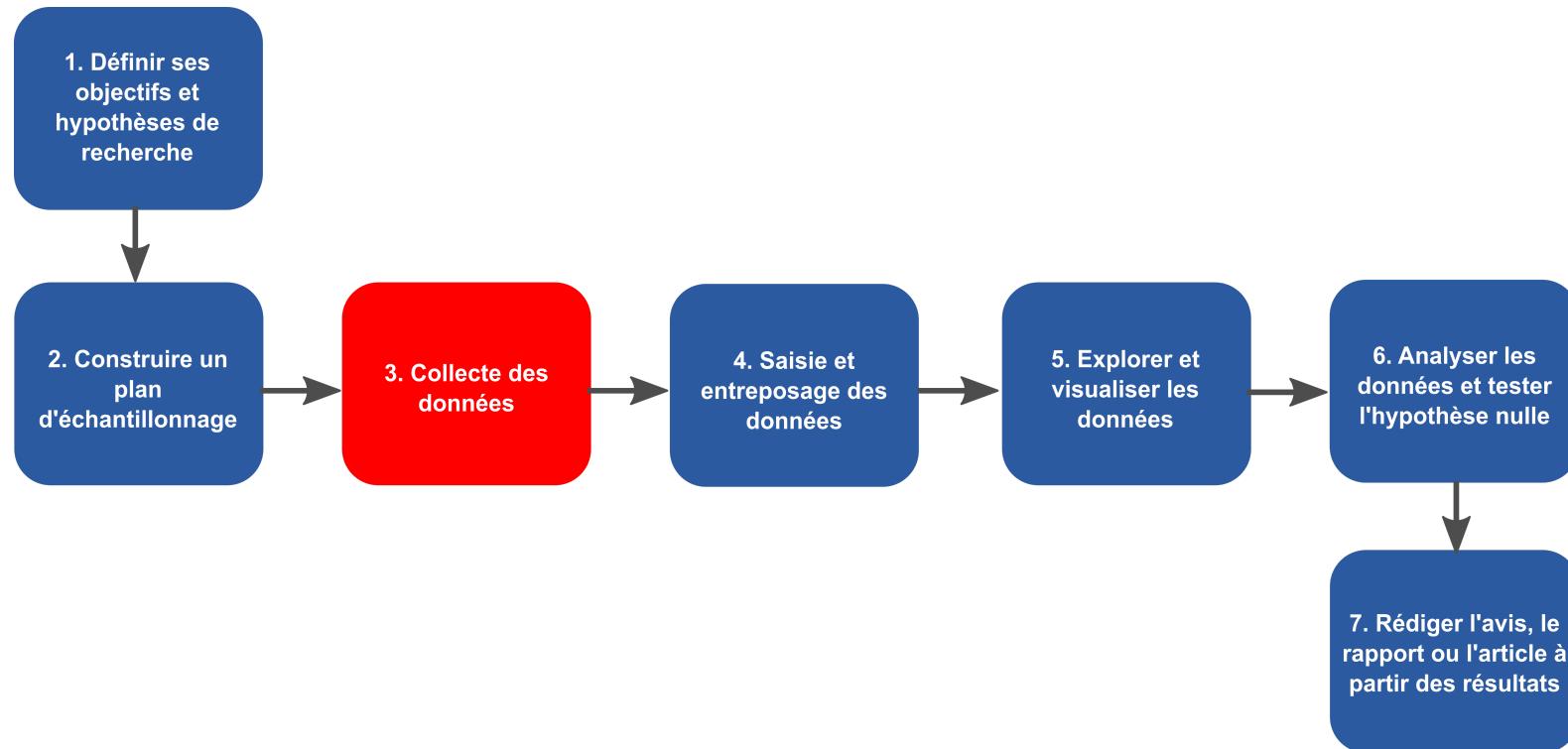
Est-ce que l'on parvient à expliquer une collaboration à partir de caractéristiques des noeuds ?

Pour commencer

En équipe de 4, on vous demande de commencer à planifier une campagne de collecte de données. Commencer par discuter des types de données que vous souhaitez récolter, faites la liste des informations nécessaires pour répondre à la question. Ensuite, établissez un protocole afin de récolter ces données.

Les données en biologie

La collecte de données



La collecte de données

En biométrie, il existe plusieurs grandes familles de données:

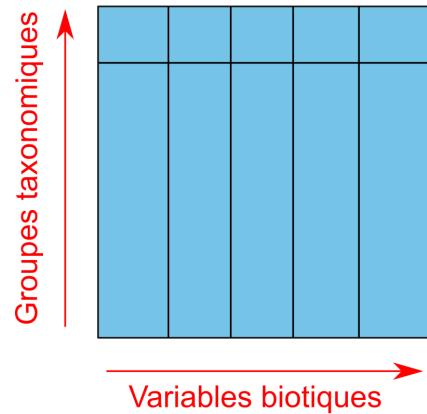
1. Quantitative (variables continues)
2. Semi-quantitative (variables discrètes)
3. Qualitatives (variables de rang)

Le type de données collectées conditionne les analyses statistiques que l'on pourra réaliser sur les données.

La collecte de données en biologie

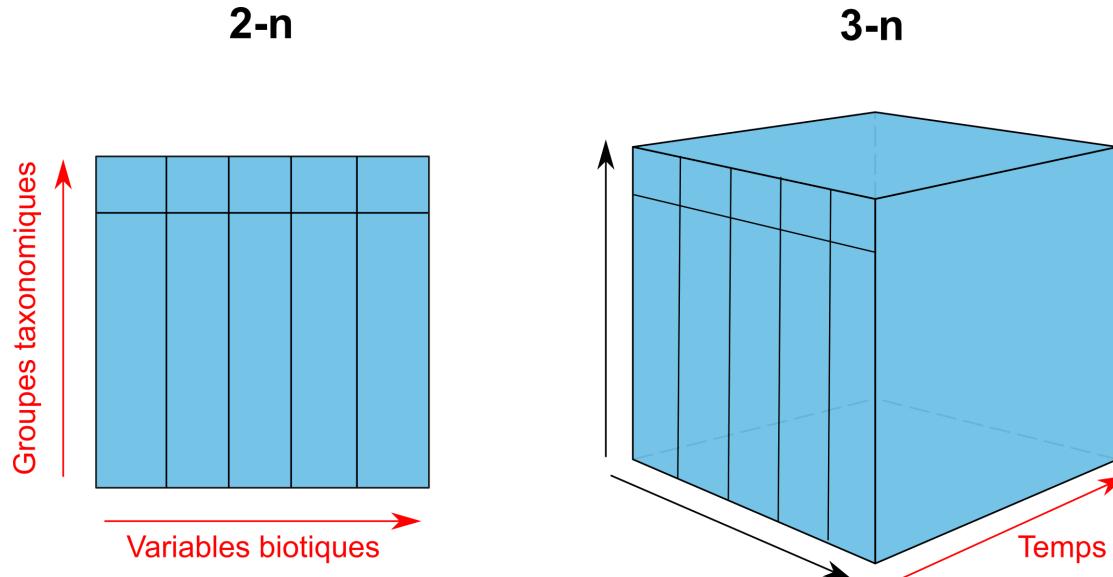
Alors, qu'en est-il d'une donnée biologique ?

2-n



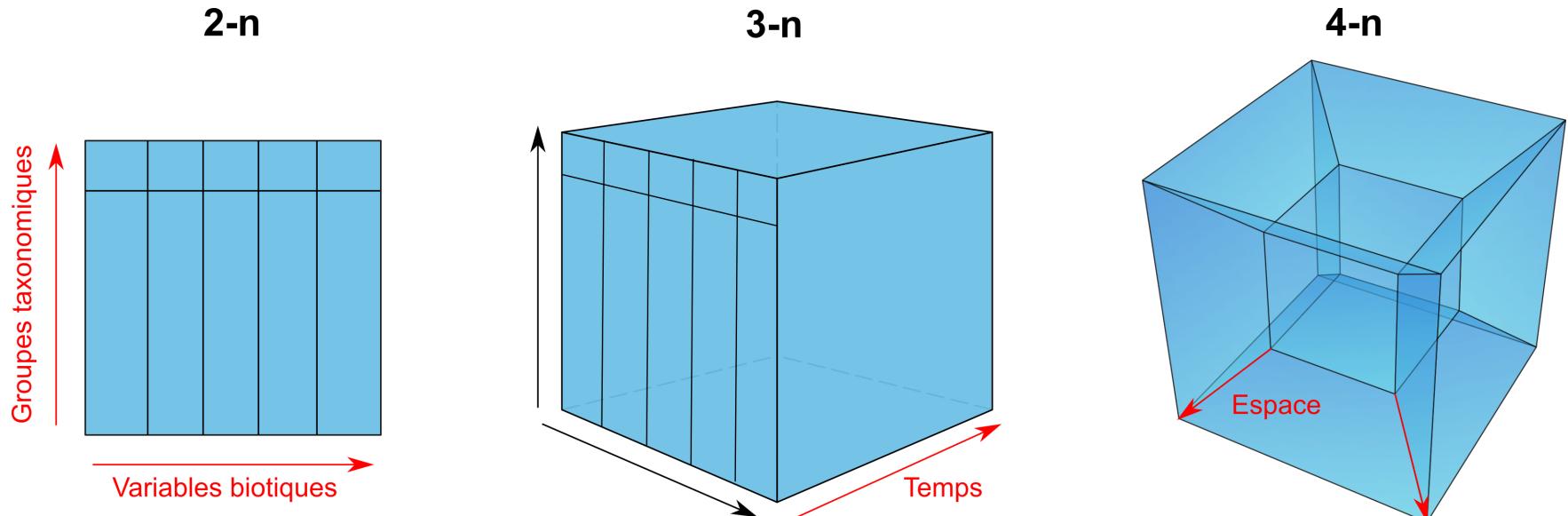
La collecte de données en biologie

Le problème de multidimensionnalité



La collecte de données en biologie

Le problème de multidimensionnalité



Note: Pour la prise de données de facteurs environnementaux (abiotiques), on retrouverait une forme de type 3n.

La collecte de données en biologie

En biologie, on classifie les données selon 4 dimensions/classes d'information:

1. Biotique/abiotique
2. Taxonomique
3. Temporelle
4. Spatial

Au sein de ce cours, nous nous attarderons à la façon de structurer ses données. J'aborderais les spécificités propres à chacune de ces dimensions. Nous attarderons d'abord au format des données, puis aux types de données.

Le format des données

Le format des données

Format large

ids	esp	2010	2011	2014
567-1	acsa	460	NA	NA
567-2	acsa	100	NA	NA
567-3	acsa	120	NA	NA
598	piru	NA	380	NA
876	abba	NA	NA	160

- ✓ Privilégier le format long
- ✓ Une ligne = une observation

Format long

ID	esp	annees	dhp_mm
567-1	acsa	2010	460
567-2	acsa	2010	100
567-3	acsa	2010	120
598	piru	2011	380
876	abba	2014	160

- ✓ Nom de colonnes court, sans accent, sans espace et explicite.
- ✓ Si possible, attachez les unités au nom de la colonne.

Le format des données: tableaux

Garder l'approche un tableau doit contenir un type d'information:

ID_plot	ID_arbre	ID_multi	esp	annees	dhp_mm
A	567	1	acsa	2010	460
A	567	2	acsa	2010	100
A	567	3	acsa	2010	120
B	598	NA	piru	2011	380
B	876	NA	abba	2014	160

ID_plot	annees	variable	valeur
A	2010	pp_tot_mm	880
B	2011	pp_tot_mm	560
B	2014	pp_tot_mm	900
A	2010	temp_max_deg	24
B	2011	temp_max_deg	26
B	2014	temp_max_deg	28

- ✓ Si l'on veut ajouter des données sur le climat, on ouvrira un nouveau tableau.

Le format des données: colonnes

Ne pas agréger l'information dans une seule colonne

ID_plot	ID_arbre	ID_multi	esp	annees	dhp_mm
A	567	1	acsma	2010	460
A	567	2	acsma	2010	100
A	567	3	acsma	2010	120
B	598	NA	piru	2011	380
B	876	NA	abba	2014	160

ID	ID_multi	esp	annees	dhp_mm
567	1	acsma	2010	460
567	2	acsma	2010	100
567	3	acsma	2010	120
598	NA	piru	2011	380
876	NA	abba	2014	160

✓ Une colonne = une information

Le format des données: colonnes

Important: votre fichier brut de données (destinée au stockage à long terme) ne doit pas contenir de champ calculé (c.a.d. une nouvelle colonne avec une moyenne, etc..)

Les types de données

Les données biotiques et abiotiques

En informatique, on distingue plusieurs types de données:

Appellation	Type	Valeurs	Taille
BOOLEAN	Boléen	vrai/faux	1 octet
INTEGER	Entiers	-998, 123	1 à 4 octets
DOUBLE, FLOAT	Nombres réels	9.98, -4.34	4 à 8 octets
CHAR, VARCHAR	Chaine de caractères	lapin	n x 1 à 8 octets
TIMESTAMP, DATE, TIME	Dates et heures	1998-02-16	4 à 8 octets

- ✓ Ce sont ces types qui seront utilisés pour entreposer nos données biotiques et abiotiques.
- ✓ Le choix d'un type approprié permet de réduire la taille du fichier de données.

Les données temporelles

La plupart des langages/programmes disposent d'un type **TIMESTAMP**, **DATE** et **TIME** pour représenter une donnée temporelle.

On utilisera préférablement la norme **ISO8601** pour représenter ces données.

- ✓ **TIMESTAMP** (Heure et temps): On utilisera la notation **YYYY-MM-ddThh:mm:ss**. ex.

1977-04-22T01:00:00-05:00

- ✓ **DATE**: On utilisera la notation **YYYY-MM-dd**. ex. **1997-04-22**

- ✓ **TIME**: On utilisera la notation **HH:mm:ss** dans un système de 24 heures. ex. **01:30:00**.

Les données temporelles

Garder à l'esprit que vos données pourraient être réutilisées à travers le Monde. Les dates ne sont pas représentées de la même manière que l'on soit en Amérique du Nord ou en Europe. **Il est donc important de normaliser la saisie de ce type d'information.**

Les données temporelles

Une autre représentation de la date du jour peut-être basé sur le calendrier Julien.

Day-of-Year Table for Non-Leap Years

DATE	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1	1	32	60	91	121	152	182	213	244	274	305	335
2	2	33	61	92	122	153	183	214	245	275	306	336
3	3	34	62	93	123	154	184	215	246	276	307	337
4	4	35	63	94	124	155	185	216	247	277	308	338
5	5	36	64	95	125	156	186	217	248	278	309	339
6	6	37	65	96	126	157	187	218	249	279	310	340
7	7	38	66	97	127	158	188	219	250	280	311	341
8	8	39	67	98	128	159	189	220	251	281	312	342
9	9	40	68	99	129	160	190	221	252	282	313	343
10	10	41	69	100	130	161	191	222	253	283	314	344
11	11	42	70	101	131	162	192	223	254	284	315	345
12	12	43	71	102	132	163	193	224	255	285	316	346
13	13	44	72	103	133	164	194	225	256	286	317	347
14	14	45	73	104	134	165	195	226	257	287	318	348
15	15	46	74	105	135	166	196	227	258	288	319	349
16	16	47	75	106	136	167	197	228	259	289	320	350
17	17	48	76	107	137	168	198	229	260	290	321	351
18	18	49	77	108	138	169	199	230	261	291	322	352
19	19	50	78	109	139	170	200	231	262	292	323	353
20	20	51	79	110	140	171	201	232	263	293	324	354
21	21	52	80	111	141	172	202	233	264	294	325	355
22	22	53	81	112	142	173	203	234	265	295	326	356
23	23	54	82	113	143	174	204	235	266	296	327	357
24	24	55	83	114	144	175	205	236	267	297	328	358
25	25	56	84	115	145	176	206	237	268	298	329	359
26	26	57	85	116	146	177	207	238	269	300	330	360
27	27	58	86	117	147	178	208	239	270	300	331	361
28	28	59	87	118	148	179	209	240	271	301	332	362
29	29	88	119	149	180	210	241	272	302	333	363	
30	30	89	120	150	181	211	242	273	303	334	364	
31	31	90	151	212	243	304	365					

Day-of-Year Table for Leap Years

DATE	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1	1	32	61	92	122	153	183	214	245	275	306	336
2	2	33	62	93	123	154	184	215	246	276	307	337
3	3	34	63	94	124	155	185	216	247	277	308	338
4	4	35	64	95	125	156	186	217	248	278	309	339
5	5	36	65	96	126	157	187	218	249	279	310	340
6	6	37	66	97	127	158	188	219	250	280	311	341
7	7	38	67	98	128	159	189	220	251	281	312	342
8	8	39	68	99	129	160	190	221	252	282	313	343
9	9	40	69	100	130	161	191	222	253	283	314	344
10	10	41	70	101	131	162	192	223	254	284	315	345
11	11	42	71	102	132	163	193	224	255	285	316	346
12	12	43	72	103	133	164	194	225	256	286	317	347
13	13	44	73	104	134	165	195	226	257	287	318	348
14	14	45	74	105	135	166	196	227	258	288	319	349
15	15	46	75	106	136	167	197	228	259	289	320	350
16	16	47	76	107	137	168	198	229	260	290	321	351
17	17	48	77	108	138	169	199	230	261	291	322	352
18	18	49	78	109	139	170	200	231	262	292	323	353
19	19	50	79	110	140	171	201	232	263	293	324	354
20	20	51	80	111	141	172	202	233	264	294	325	355
21	21	52	81	112	142	173	203	234	265	295	326	356
22	22	53	82	113	143	174	204	235	266	296	327	357
23	23	54	83	114	144	175	205	236	267	297	328	358
24	24	55	84	115	145	176	206	237	268	298	329	359
25	25	56	85	116	146	177	207	238	269	299	330	360
26	26	57	86	117	147	178	208	239	270	300	331	361
27	27	58	87	118	148	179	209	240	271	301	332	362
28	28	59	88	119	149	180	210	241	272	302	333	363
29	29	60	89	120	150	181	211	242	273	303	334	364
30	30	90	121	151	182	212	243	274	304	335	365	
31	31	91	152	213	244	305	366					

- ✓ **Inconvénient:** Le jour julien doit toujours être accompagné de l'année (YYYY).
- ✓ **Avantage:** simplifie les analyses temporelles intra-annuelles.

Les données taxonomiques

Un exemple avec l'érable à sucre

Selon vous quelle option est la meilleure?

Option	Exemple
1. Code spécifique à l'étude	ACSA
2. Code du ministère	ERS
3. Genre et espèce Option	Acer saccharum Exemple
4. Nom vernaculaire	Érable à sucre
5. Numéro Taxonomique (TSN - ITIS)	28731



Les données taxonomiques

Un exemple avec l'érable à sucre

Option	Exemple
1. Code spécifique à l'étude	ACSA
2. Code du ministère	ERS
3. Genre et espèce	Acer saccharum
4. Nom vernaculaire	Érable à sucre
5. Numéro Taxonomique (TSN - ITIS)	28731

- ✓ **Option 1 et 2:** Doit être associé à des métadonnées. Risque de perte du fichier attaché.
- ✓ **Option 3:** Le genre et l'espèce peuvent changer à travers le temps.
- ✓ **Option 4:** Le nom vernaculaire des espèces est le pire choix. Le nom vernaculaire est propre à un pays, à une région géographique, à une culture/dialecte.

Les données taxonomiques

Un exemple avec l'érable à sucre

Option	Exemple
Code spécifique à l'étude	ACSA
Code du ministère	ERS
Genre et espèce	<i>Acer saccharum</i>
Nom vernaculaire	Érable à sucre
Numéro Taxonomique (TSN - ITIS)	28731

✓ **Option 4:** Cette option couplée à l'option 3, est le meilleur choix.

Les données taxonomiques

On privilie^{ge} g  n  ralement, l'utilisation de code esp  ce standardis  e:

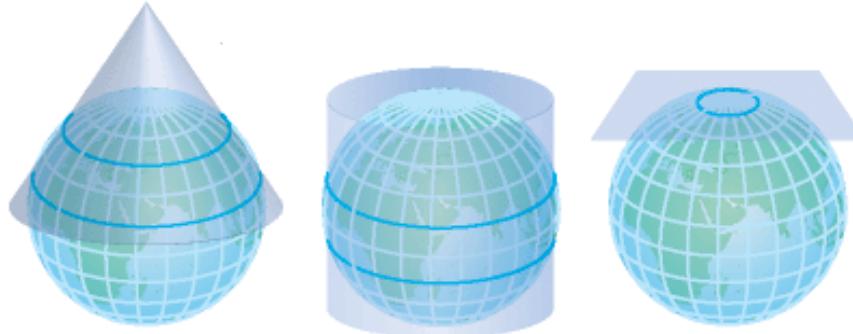
1. ITIS
2. VASCAN (Plantes vasculaires du Canada)
3. NCBI

Avantage: Chacune de ces institutions/infrastructures, nous permettent de valider et retirer l'ensemble de la classification taxonomique d'une esp  ce    partir de son code. M  me si l'identifiant change (nouvelle classification), nous serons en mesure de trouver le nouvel identifiant taxonomique    partir de l'ancien.

Exemple: [https://www.itis.gov/servlet/SingleRpt/SingleRpt?
search_topic=TSN&search_value=28731#null](https://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_value=28731#null)

Les données spatiales

Il existe plus de **65 familles de projections géographiques** pour représenter des coordonnées sur la planète, en voici 3 des plus connues:



- ✓ Il est important de choisir un bon système de projection pour minimiser la déformation spatiale (surtout à nos latitudes)
- ✓ À nos latitudes, on privilégiera l'utilisation d'une projection conique. Les ministères du Québec conseillent généralement l'utilisation d'une **projection conique conforme de Lambert**.

Les données spatiales

- ✓ **Ce qu'il est important de savoir:** des coordonnées spatiales sans système de projection ne veulent strictement rien dire.
- ✓ Ainsi, lorsque l'on entrepose des données spatiales, trois colonnes doivent être représentées:
 - La coordonnée en X
 - La coordonnée en Y
 - La projection écrite en texte (voir votre GPS), ou préférablement l'identifiant unique de la projection.

Les données spatiales

Deux bases de données connues permettent de fournir des identifiants uniques:

1. **EPSG**: *European Petroleum Survey Group.*
2. **SRID**: *Spatial reference system.*

Ces deux identifiants sont généralement identiques et peuvent être trouvés à cette adresse:

<http://spatialreference.org/>

Exemple: **<http://spatialreference.org/ref/epsg/2138/>**

L'absence de données

On peut représenter l'absence de données de plusieurs façons:

- ✓ Laisser la cellule vide (**NULL**)
- ✓ Mettre un **NA** (*Not Available*)
- ✓ Mettre un 0
- ✓ Mettre **-9999** dans une colonne numérique

Selon vous, quelle est l'action la plus appropriée ?

Le format des données

On peut représenter l'absence de données de plusieurs façons:

- ✓ Laisser la cellule vide: montre que l'information n'a pas été saisie (un oubli)
- ✓ Mettre un **NA** (*Not Available*): Montre que l'information est réellement indisponible (car le NA est saisie par un humain).
- ✓ ~~Mettre un 0~~: **JAMAIS** (empêche la distinction entre un vrai d'un faux 0, influence la moyenne)
- ✓ Mettre **-9999** dans une colonne numérique: Ce choix peut être utilisé seulement pour les jeux de données très importants (centaine de Megas-octet), et doit être référencé dans les métadonnées.

Choisir le bon type et format de données

Si l'on ne choisit pas le type de données approprié, cela aura diverses conséquences:

- ✓ Des problèmes de performance (ex. : il est plus rapide de faire une recherche sur un nombre que sur une chaîne de caractères)
- ✓ Un comportement contraire à celui attendu (ex. : trier sur un nombre stocké comme tel, ou sur un nombre stocké comme une chaîne de caractères ne donnera pas le même résultat)
- ✓ L'impossibilité d'utiliser des fonctionnalités propres à un type de données (ex. : stocker une date comme une chaîne de caractères vous prive des nombreuses fonctions temporelles disponibles).

Finalelement...

Pourquoi prendre soin de ces données ?

La saisie des données dans LibreOffice Calc

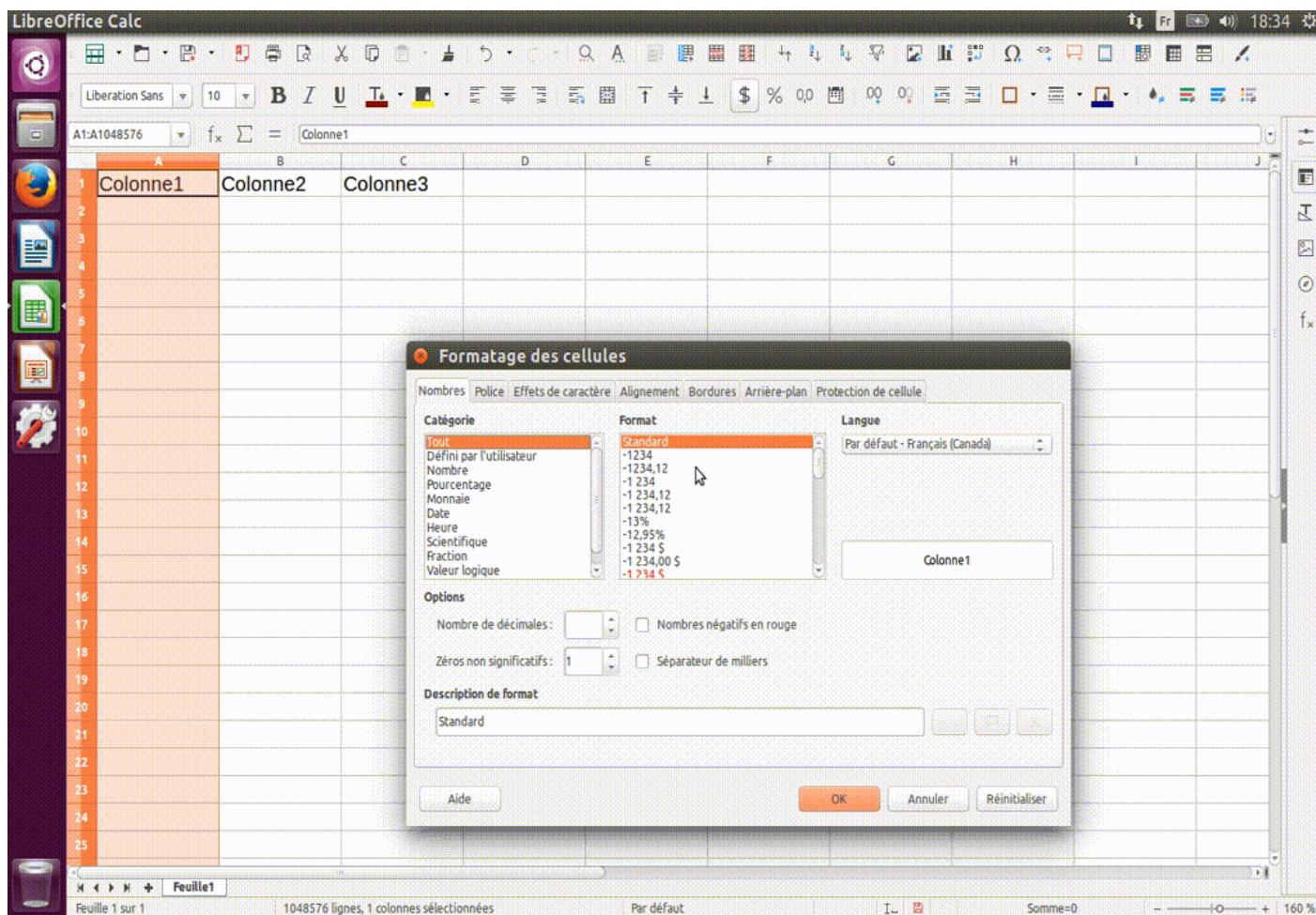
LibreOffice Calc



LibreOffice Calc est la version gratuite et *open-source* de Microsoft Excel. L'ensemble des fonctionnalités présenté dans ce cours peuvent être retrouver dans Microsoft Excel.

- ✓ Ouvrez LibreOffice Calc, en vous servant de la barre de lancement à votre gauche.

Déterminez vos colonnes et le type de données



Consolidez vos données à l'aide de la validation

The screenshot shows a LibreOffice Calc spreadsheet window titled "Sans nom 1 - LibreOffice Calc". The status bar indicates "VMBIO-500 [Running]" and the time "19:09". The toolbar at the top includes various icons for file operations, text styling, and data manipulation. The formula bar shows the cell reference "A2". The main worksheet has three columns labeled "ID_arbre", "annees", and "dhp_mm". Row 1 contains the column headers, and row 2 contains the first data entry. The sidebar on the left lists icons for various applications, and the bottom right corner shows a zoom level of "160%".

Saisir l'information

The screenshot shows the LibreOffice Calc application interface. A data table is displayed in the main area, and a data entry dialog box is overlaid on it.

Data Table:

	A	B	C
1	ID_arbre	annees	dhp_mm
2	8	2020	110
3	9	2017	114
4	7	2017	116
5	47	2017	860

Formulaire de données (Data Entry Form):

This dialog box contains three input fields corresponding to the selected row in the table:

- ID_arbre: 8
- annees: 2020
- dhp_mm: 110

On the right side of the dialog, there are several buttons:

- Nouveau (New)
- Supprimer (Delete)
- Restaurer (Restore)
- Enregistrement précédent (Previous Record)
- Enregistrement suivant (Next Record)
- Aide (Help)
- Fermer (Close)

At the bottom of the dialog, it says "1 / 4".

Application Status:

- Top bar: LibreOffice Calc, Liberation Sans font, cell A1:CS, formula ID_arbre.
- Toolbar: Standard Calc tools.
- Bottom status bar: Feuille 1 sur 2, 5 lignes, 3 colonnes sélectionnées, Par défaut, Somme=9342, zoom 160%.

Retour sur le projet de session

Maintenant que vous en savez plus sur le format des données et sur les règles de saisie, élaborez votre formulaire et commencez la récolte. Vous devez collecter les données pour le début de la séance de la semaine prochaine.