

Carte de départ Synopsis



Vous faites partie d'une jeune équipe de recherche d'un institut en tourisme en Suisse et prévoyez de vous lancer dans un projet de recherche sur l'hôtellerie en Suisse. Votre collègue, avec son équipe, a déjà mené une étude sur le sujet. Contrairement à elle, vous n'êtes pas au clair sur les besoins liés à la gestion de vos données, car vous n'avez pu participer à la formation de la Bibliothèque sur la thématique de la gestion des données de la recherche.

Votre collègue part en vacances en Amazonie cet après-midi. Juste avant son départ, elle rédige rapidement l'ébauche d'une feuille de route avec les grandes étapes pour mener à bien votre projet et assurer une gestion adéquate de vos données. Comme son avion va bientôt décoller, elle n'a pas le temps de terminer et vous propose de vous installer à son bureau, le numéro 59, où vous trouverez toutes les informations pour compléter la liste qu'elle a commencée, à condition de fouiller un peu dans ses affaires et celles de son équipe...

Heureusement, vous pouvez aussi compter sur Santiago, le spécialiste en information documentaire qui travaille dans la bibliothèque de votre institution, Inès, l'informaticienne, ainsi que Jack le juriste, qui vous aideront pendant votre quête.

Mais attention, il ne reste qu'une heure avant la fermeture de l'institut pour les vacances de Noël, il vous faut donc sortir avant cela si vous ne voulez pas y rester coincés !

1

1


La **3.3 visualisation des données** permet de

- donner de la valeur aux données
- communiquer des informations clairement et efficacement à travers des moyens graphiques
- représenter visuellement des valeurs numériques

Top 10 des communes selon le total des nuitées

 Zurich
1,1 mio

 Zermatt
1,0 mio




 Genève
0,7 mio

4.	Davos	0,7 mio
5.	St. Moritz	0,6 mio
6.	Bâle	0,5 mio
7.	Lucerne	0,5 mio
8.	Lausanne	0,4 mio
9.	Grindelwald	0,4 mio
10.	Arosa	0,4 mio

36,1%
Taux d'occupation des chambres



23,7 mios
de nuitées dans l'hôtellerie

 Suisse	16,4 mios
 Europe	6,0 mios
 Asie	0,6 mio
 Amérique	0,6 mio
 Afrique et Océanie	0,1 mio

4646
hôtels recensés



60,2
lits disponibles par établissement en moyenne



Évolution des prix à la consommation dans l'hôtellerie (2019/2020)

Nuitées par grandes régions dans l'hôtellerie

Suisse orientale	6,3 mios
Région lémanique	5,8 mios
Espace Mittelland	4,3 mios
Suisse centrale	2,3 mios
Zurich	2,0 mios
Tessin	1,8 mios
Suisse du Nord-Ouest	1,1 mios

2,2
nuits

Durée de séjour moyenne dans un hôtel



3

3

Grâce aux différents critères listés sur le site re3data.org, l'équipe de recherche a pu **5.1 choisir un dépôt de données**. Elle a hésité entre plusieurs dépôts...



#1 SWISSUbase

SWISS  base

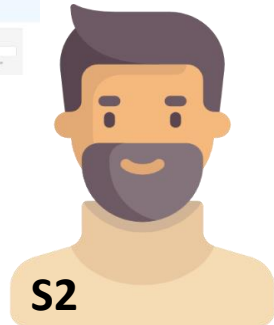


#2 OLOS



#3 Zenodo

zenodo



4

4



Les 4 post-it vous ont permis de trouver différents types de données et ainsi **1.3 Réfléchir aux données produites ou réutilisées** que vous serez amené à produire dans le cadre de votre projet.

**Données
d'observation**

Ex. questionnaires,
entretiens,
neuroimagerie

**Données de
simulation**

Ex. données
météorologiques,
simulation sismique

**Données
expérimentales**

Ex.
chromatogrammes,
puces à ADN, essais

**Données dérivées
ou compilées**

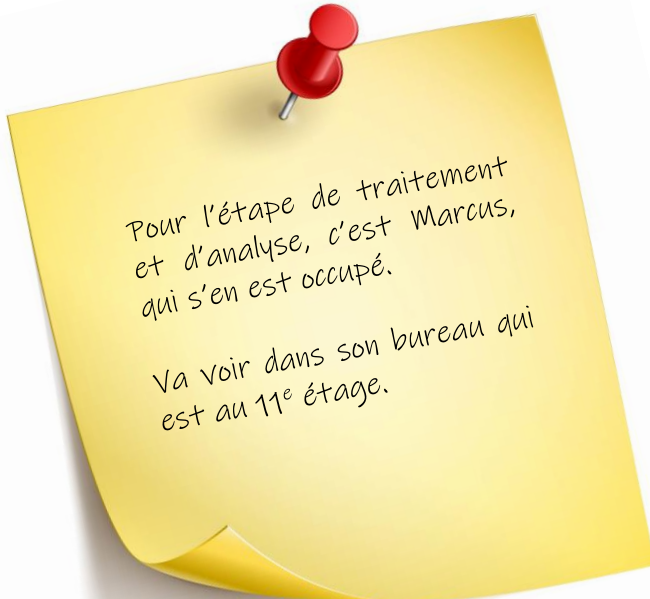
Ex. fouille de texte,
imagerie IRM, bases de
données compilées

6

6

92

Après avoir passé une dernière fois le bureau de votre collègue en revue, vous trouvez encore ce post-it.



Pour l'étape de traitement
et d'analyse, c'est Marcus,
qui s'en est occupé.

Va voir dans son bureau qui
est au 11^e étage.

7

7

96 19

Aïe aïe aïe, l'équipe de recherche s'est emmêlé les pinceaux au moment du **5.3 choix d'un identifiant pérenne**.

Entre les identifiants des jeux de données et ceux des contributeurs, saurez-vous retrouver celui des jeux de données ?



8

8

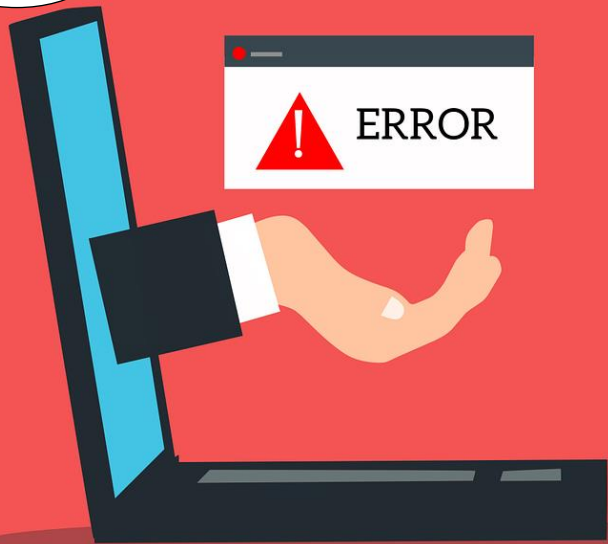
Voilà un petit café bien mérité.

Le temps de cette pause, vous **comptez** les post-it annotés par votre collègue...



9

9



Mauvaise réponse !

1 minute de pénalité

Avez-vous complété
intégralement la feuille de route
pour l'étape en cours ?
Si oui retournez cette carte

11

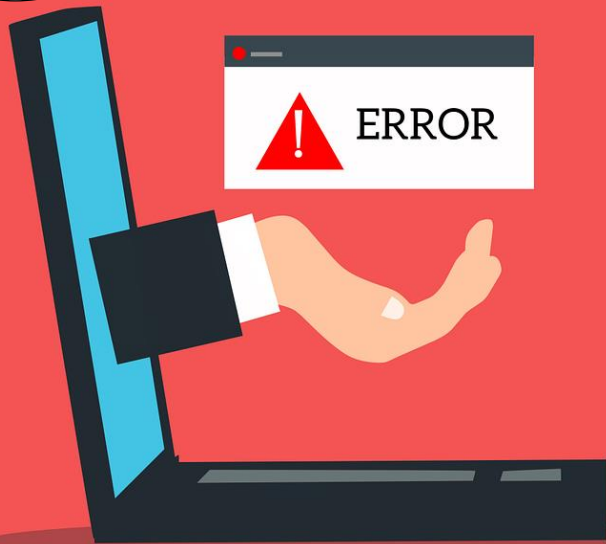
11

Etape 3 : traitement et analyse

~~6 28 31 S1~~

12

12



Mauvaise réponse !

1 minute de pénalité

13

13



Mauvaise réponse !

1 minute de pénalité

15

15

	A	B	C	D	E	F
1	GEONR	GEONAME	CLASS_HAB	VARIABLE	VALUE	STATUS
353	6153	Monthey	3	arr_etr_p	35.7	A
354	3443	Gossau (SG)	3	arr_etr_p	35.9	A
355	616	Münsingen	2	bet_t	36	A
356	4001	Aarau	4	nuit_etr_p	36	A
357	5721	Gland	2	zim_t	36	A
358	5822	Payerne	2	arr_etr_p	36.7	A
359	5938	Yverdon-les-Bains	4	nuit_etr_p	36.8	A
360	5586	Lausanne	6	etab_t	37	A
361	1201	Altdorf (UR)	1	nuit_etr_p	37	A
362	3001	Herisau	3	arr_etr_p	37.3	A
363	3851	Davos	2	arr_etr_p	37.5	A
364	942	Thun	4	nuit_etr_p	37.8	A
365	5401	Aigle	2	bet_t	38	A
◀ ▶		Données	Variables	-	+	

... + celui-ci ?

16

	Remarques
Villes suisses	Les villes considérées dans cette publication sont les 162 villes statistiques suisses
geonr/genoame	- geonr sont les numéros officiels de l'OFS pour chaque ville - genoame sont les noms des régions
class_hab	Explications taille de la commune - classe 6: Ville de 100000 habitants et plus, - classe 4: 20000-49999 habitants, - classe 2: 10000-14999 habitants, - Suisse: pas de classe de taille - classe 5: 50000-99999 habitants, - classe 3: 15000-19999 habitants, - classe 1: moins de 10000 habitants,
variables	<div> arr_t : Arrivées total arr_etr_p : Arrivées de l'étranger (%) nuit_etr_t : Nuitées - hôtes de l'étranger sej_moy : Durée moyenne de séjour (nuits) etab_t : Établissements ouverts (nombre) bet_t : Lits disponibles (nombre) </div> <div> arr_etr_t : Arrivées hôtes de l'étranger nuit_t : Nuitées - total nuit_etr_p : Nuitées - de l'étranger (%) nui_1000 : Nuitées pour 1000 habitants zim_t : Chambres disponibles (nombre) </div>
status	Explications - A: Valeur normale - N: Non significatif - Q: Non indiqué : protection des données
Titre	Hôtellerie, établissements, arrivées, nuitées, en 2020
Période	2020
Source	OFS – statistique de l'hébergement touristique

Que peut bien contenir ce fichier... ?

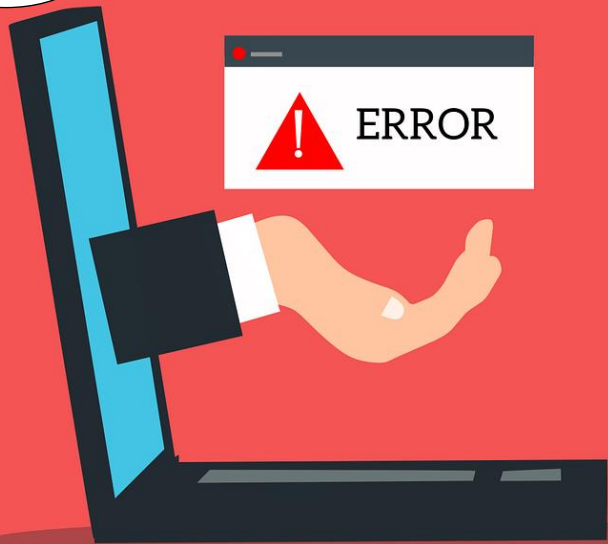
17

17



18

18



Mauvaise réponse !

1 minute de pénalité

19

















19

Logiciel d'analyse des données (3.2)

Ajoutez le fichier à analyser

Sélect. fichiers Aucun fichier choisi

21

	Découverte de la Médiathèque santé.pptx	01.09.2021 12:59	Présentation Microso...
	Demande d'articles dans swisscovery copies nu...	15.03.2021 08:43	Adobe Acrobat Docu...
	Déroulé du séminaire.docx	23.11.2020 18:32	Document Microsoft ...
	DLCM_Policy.pdf	12.05.2021 14:31	Adobe Acrobat Docu...
	Dossier_Numero vingt-six.zip	06.09.2021 12:48	Dossier compressé
	Early Detection of Mild Cognitive ImpairmentW...	07.10.2020 16:51	Adobe Acrobat Docu...
	ERES_HIRSC.pdf	19.10.2020 15:17	Adobe Acrobat Docu...
	etat_des_lieux_niveau_preuve_gradation.pdf	30.11.2020 15:16	Adobe Acrobat Docu...
	exercice_solution.tflx	05.06.2021 15:31	Tableau Flow File
	export.ris	25.11.2020 10:09	RIS Formatted File
	file.png	24.08.2021 20:00	Fichier PNG
	FHDF_EX.doc	15.07.2021 08:32	Document Microsoft ...
	folder.svg	24.08.2021 18:24	SVG Document
	future-science-laboratory-human-genetics-rese...	24.08.2021 20:39	Dossier compressé
	gazette_Mise_au_Point_PhD_Ginkgo.pdf	20.07.2021 20:11	Adobe Acrobat Docu...
	googlepage.html	03.06.2021 12:26	Chrome HTML Docu...

22

22

79

Résultat de l'analyse (extrait)

Hôtellerie: offre et demande dans les établissements ouverts, chiffres provisoires

2020	Etablissements ouverts	Nuitées	Variation des nuitées en %*
Janvier-juillet	3 860	15 018 745	12,2
> Juillet	4 345	3 633 744	6,0

*par rapport à la même période de l'année précédente

Source: Statistique de l'hébergement touristique (HESTA)

23

En résumé, lorsque les équipes de recherche se préparent à archiver leurs données, elles doivent s'assurer de choisir des formats

- ✓ non propriétaires
- ✓ non cryptés
- ✓ non compressés
- ✓ utilisés et reconnus dans leur domaine
- ✓ interopérables

25

25



C'est faux, la carte 25 ne correspond pas au meilleur nommage de fichier.

Il faut éviter d'utiliser des caractères spéciaux ainsi que les espaces entre les mots.

1 minute de pénalité

26

26

Vous avez récupéré le dossier sur la recherche préliminaire de votre collègue. Mais quelle pagaille! Lequel de ces fichiers respecte le mieux les règles de nommage?

21



labtox_recent_110**820**_old_version.sps



FFTX_Méta_3**77**6438656.sps



OFS Ville&Hotel **25**-12-2019.docx



OFS_Continental_202006**28**_V01.png



Avez-vous complété
intégralement la feuille de route
pour l'étape en cours ?
Si oui retournez cette carte

28


28

4 26 29 59 S10 S15



Le dossier «Villes suisses 2020 : hôtellerie et hébergement» contenait cette image de l'Hôtel Continental

29



Les informations contenues sur ce cahier sont très importantes pour la réussite d'un projet.



1.1 Rm d×gM □ IM MR

- A. Informations administratives
- B. Collecte de données
- C. Documentation et métadonnées
- D. Éthique et conformité légale
- E. Stockage et sauvegarde
- F. Migration et conversion
- G. Restauration de données
- H. Responsabilités et ressources

30

30

~~17~~

Il semblerait que vos collègues aient enregistré plusieurs versions des données récoltées à partir du questionnaire pendant le processus de nettoyage.

Prenez les cartes **56**, **60** et **69**.

Laquelle correspond au fichier le mieux nettoyé, qu'ils ont ensuite pu **ajouter au logiciel d'analyse** ?

31

31

~~15 16 36~~

Le fichier contenait **des données d'observation**, sous forme de réponses au questionnaire des villes suisses (carte 15) ainsi que **la documentation des données** (carte 16) qui vise à décrire les données collectées pour en faciliter l'utilisation, la récupération et la gestion.



En effet, il est important de **2.3 Documenter son projet** le plus tôt possible puis de compléter cette documentation au fur et à mesure.

32

32

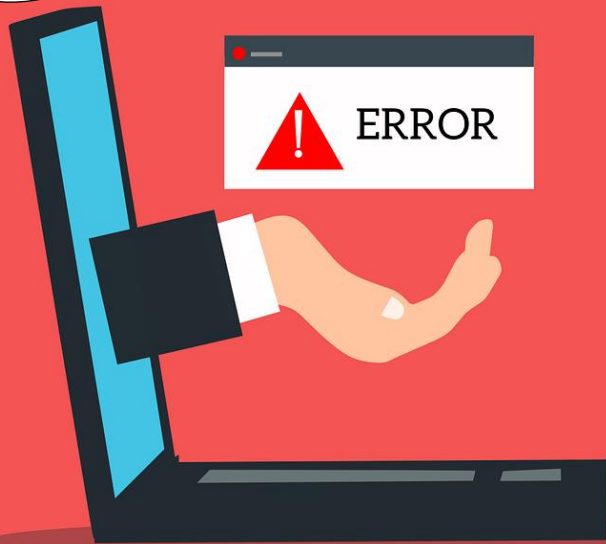


Mauvaise réponse !

1 minute de pénalité

35

35



Mauvaise réponse !

1 minute de pénalité

36

Questionnaire villes suisses 2020

Hôtellerie, établissements, arrivées, nuitées

Ville

Commune

Nbre habitants

.....

.....

.....

.....

Les établissements

Nbre d'établissements ouverts

16

Nbre de chambres disponibles

Nbre de lits disponibles

Les arrivées

Nbre total d'arrivées – hôtes de l'étranger

15

Nbre total d'arrivées de l'étranger en %

Nbre total d'arrivées

Les nuitées

Nbre de nuitées – hôtes de l'étranger

Nbre de nuitées de l'étranger en %

Total des nuitées

Nuitées pour 1000 habitants

Envoyer

No de sauvegarde

Sauvegarder

La réception de l'hôtel a dû remplir un questionnaire en 2020 sur l'hôtellerie et l'hébergement dans son établissement afin de **2.2 Collecter/acquérir des données** pour le projet des chercheurs.

37



37
FIN!







48 85 S8

Bravo ! Toutes vos notes en main, vous passez la porte du bureau in extremis. Votre feuille de route sera votre guide pour mener à bien votre propre projet.

Vous avez bien travaillé, les vacances de Noël vous attendent. Vous reviendrez en forme en janvier pour attaquer votre projet de recherche et assurer une bonne gestion de vos données.

44

En consultant les fichiers, vous observez que les données archivées sont enregistrées sous des formats bien spécifiques...

Nom	Type
 OFS_Villes_Suisses_2020_Dataset.csv	Fichier CSV Microsoft Excel
 OFS_Villes_Suisses_2020_Infographie.pdf	Foxit Reader PDF Document
 OFS_Villes_Suisses_2020_Logo.png	Fichier PNG
 OFS_Villes_Suisses_2020_Metadata.xml	Document XML
 OFS_Villes_Suisses_2020_Questionnaire.odt	Texte OpenDocument
 README.txt	Document texte

En effet, pour assurer la réutilisation, l'interopérabilité et la pérennité des données, il est nécessaire de **4.1 utiliser des formats de fichiers ouverts**.

Mais pourquoi ? Qu'est-ce qu'un format ouvert ? Comment choisir le bon format ?



Avez-vous complété
intégralement la feuille de route
pour l'étape en cours ?
Si oui retournez cette carte

47

47

23 51
99

Etape 5 : partage des données



Avez-vous complété
intégralement la feuille de route
pour l'étape en cours ?
Si oui retournez cette carte

48

48

47 84

Etape 6 : Réutilisation des données



51

51

~~61~~ ~~68~~
~~528~~

47 messages non lus

Je viens d'embarquer!

J'ai trop hâte 😊

J'ai oublié de te

prévenir : pour les

infos sur le partage

des données,

consulte l'ordinateur

de Rachel, il doit être

dans son bureau.

Bonnes vacances!

P.S: N'oublie pas de

rentrer chez toi avant

la fermeture ;)

Effacer

Répondre

56

56

GEONR	GEONAME	CLASS_HAB	VARIABLE	VALUE	STATUS
	Allschwil	4	etab_t	1	A
6248	Sierre	3	sej_moy	1.5	A
6248	Sierre	3	sej_moy	1.5	A
247	Schlieren	3	etab_t	1	A
5938	Yverdon-les-Bains	4	sej_moy	1.6	A
141	Thalwil	3	etab_t	1	A
4012	Suhr	2	etab_t	2	A
6436	Le Locle	2		2	A
1407	Sarnen	2	sej_moy	1.8	A
3443	Gossau (SG)	3	sej_moy	5	A
616	Münsingen	2	etab_t	1	A
155	Männedorf	2	etab_t	2	
4280	Oftringen	2	etab_t	1	A
177		2	etab_t	1	A
5589	Prilly	2	etab_t	1	A
1707	Risch	2	etab_t	1	A
118	Rüti (ZH)	2	etab_t	1	A
1708	Steinhausen		etab_t	1	A

59

59

Synopsis

21

29 janvier

Rédiger le DMP

A.....

B.....

C.....

D.....

E.....

F.....

G.....

H.....

Etape 1 : Elaboration et planification du projet

60

60

GEONR	GEONAME	CLASS_HAB	VARIABLE	VALUE	STATUS
1201	Altdorf (UR)	1	arr_etr_p	39.5	A
1201	Altdorf (UR)	1	arr_etr_t	3819	A
1201	Altdorf (UR)	1	arr_t	9672	A
1201	Altdorf (UR)	1	bet_t	115	A
1201	Altdorf (UR)	1	etab_t	3	A
1201	Altdorf (UR)	1	nui_1000	2132	A
1201	Altdorf (UR)	1	nuit_etr_p	37	A
1201	Altdorf (UR)	1	nuit_etr_t	7411	A
1201	Altdorf (UR)	1	nuit_t	20041	A
1201	Altdorf (UR)	1	sej_moy	2.1	A
1201	Altdorf (UR)	1	zim_t	59	A
3251	Altstätten	2	arr_etr_p	27.6	A
3251	Altstätten	2	arr_etr_t	1713	A
3251	Altstätten	2	arr_t	6204	A
3251	Altstätten	2	bet_t	163	A
3251	Altstätten	2	etab_t	3	A
3251	Altstätten	2	nui_1000	848	A

61

61

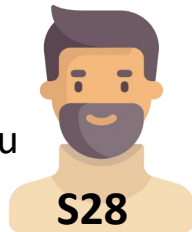
Il est nécessaire de **4.3 ajouter des métadonnées** adéquates aux données stockées, afin de permettre leur recherche et leur découverte. *Les métadonnées sont généralement présentées comme des «données à propos des données».*

~~81 S13 S25~~

Il existe **plusieurs types de métadonnées** :

- Métadonnées descriptives : titre, sujet, créateurs des données...
- Métadonnées administratives : formats de fichiers, versions, droits de réutilisation...
- Métadonnées structurelles : relations avec d'autres entités, liste de variables...

Prenez la carte **68**, il s'agit des métadonnées du projet de vos collègues au format XML.



64

64

~~1~~ ~~22~~

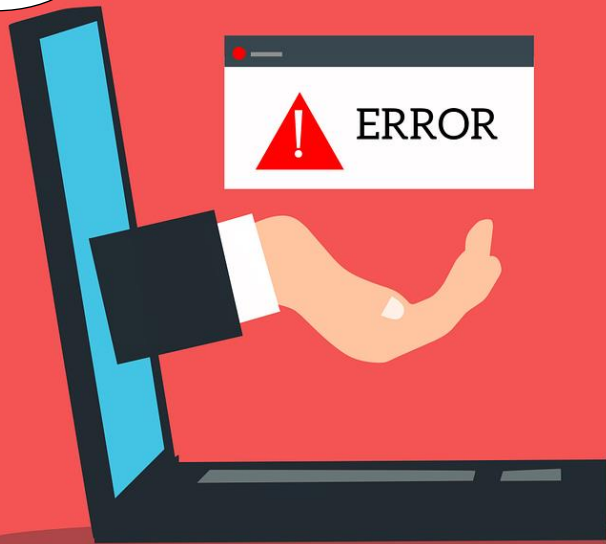
TO DO

- ✓ Nettoyer les données
- ✓ Analyser les données
- ✓ Visualiser les données
- ✓ RDV au service informatique pour discuter archivage des données
(n° tel interne 0099)

En voyant un extrait de la liste écrite par Marcus, vous décidez de vous rendre vous aussi au service informatique.

65

65



Mauvaise réponse !

1 minute de pénalité

68

```

<title>Hôtellerie, établissements, arrivées, nuitées, en 2019</title>
</>
<creator>
  <creatorName>Office fédéral de la statistique OFS</creatorName>
</creator>
</>
<publisher>Opendata.swiss</publisher>
<publicationYear>2021</publicationYear>
<resourceTypeGeneral="Dataset">Dataset</>
<subject>Tourism</subject>
<subject>Switzerland</subject>
</>
<date dateType="Available">2021-04-20</date>
</>
<carte carteAPrendre="pourEtapeSuivante">51</carte>
</>
<relatedIdentifier relatedIdentifierType="URL" relationType="IsIdenticalTo">https://www.bfs.admin.ch/bfsstatic/dam/assets/16564305/master</relatedIdentifier>
</>
<description descriptionType="TableOfContents">Statistiques des villes suisses 2021 - Tourisme: Établissements ouverts (nombre); Chambres disponibles; Arrivées - total; Arrivées - hôtes de l'étranger; Arrivées - de l'étranger (%); Nuitées - total; Nuitées - hôtes de l'étranger; Nuitées (nuits); Nuitées pour 1000 habitants</description>
</>
<language>de, fr</language>
<alternateIdentifier Type de l'identifiant alternatif="Opendata.swiss">ts-b-ssv-10.03.01-2021</alternateIdentifier>
</>
<undefined>CSV</undefined>
</>

```


69

69

GEONR	GEONAME	CLASS_HAB	VARIABLE	VALUE	STATUS
_2762	Allschwil	4	etab_t	1	A
53	Bülach	4	etab_t	1	a
6248	Sierre	3	sej_moy	1.5	A
1702	Cham	3	etab_t	1	A
247	Schlieren	3	etab_t	1	A
5938	Yverdon-les-Bains	4	sej moy	1.6	A
141	Thalwil	3	etab_t	N/A	A
_199x	Volketswil	3	etab_t	1	A
6436	Le Locle	2	etab_t	2	A
156	Meilen	2	etab_t		A
3443	Gossau (SG)	3	etab_t	5	A
616	Münsingen	2	etab t	1	A
356	Muri bei Bern	2	etab_t	N.A.	A
4280	Oftringen	2	etab_t	1	A
177	Pfäfflikon	2	etab_t	NA	a
5589	Prilly	2	etab_t		A
1707	Risch	2	etab_t	1	A
118	Rüti (ZH)	2	etab_t	1	A

71

71

Lors du dépôt des données, l'équipe a dû **5.2 déterminer les conditions d'accès**. Quelle option a-t-elle choisie, à votre avis ?

3

S2 S5



Données
fermées



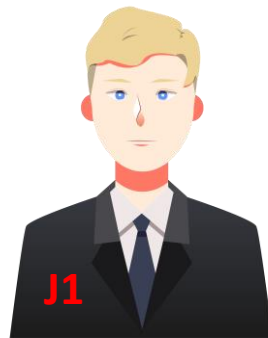
Données sur
demande



Données
protégées par
un embargo



Données
ouvertes



72

72

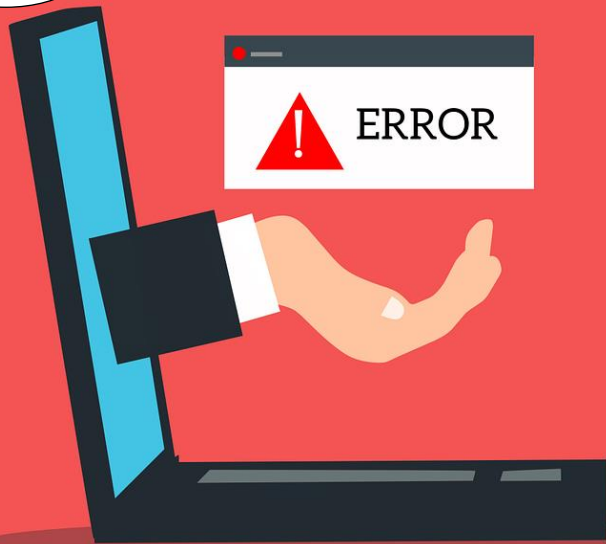


Mauvaise réponse !

1 minute de pénalité

73

73



Mauvaise réponse !

1 minute de pénalité

75

75



C'est faux, la carte 56 ne correspond pas à la meilleure version des données nettoyées.

Pour cause : les données sont complètement mélangées, il y a des doublons et certaines cellules sont vides, sans explication.

1 minute de pénalité

77

77



C'est faux, la carte 77 ne correspond pas au meilleur nommage de fichier.

Il n'est pas recommandé d'utiliser de la ponctuation dans le nom des fichiers

1 minute de pénalité

79

79

~~19~~ ~~30~~ ~~56~~

~~60~~ ~~69~~

~~i67~~ ~~i68~~

Bravo ! Vous avez bien identifié le fichier le mieux nettoyé et le plus structuré, qui a été utilisé pour l'analyse des données.



Vous pouvez maintenant observer le résultat de cette analyse.



81

81

Vous vous rendez compte que les données archivées ne comportent pas l'ensemble des données que vos collègues vous avaient présentées pendant leur projet.

 OFS_Villes_Suisses_2020	Dossier de fichiers
 OFS_Villes_Suisses_2020_ARCHIVAGE	Dossier de fichiers

Cela vous étonne, vous décidez donc de demander conseil à Santiago.



82

82



C'est faux, la carte 82 ne correspond pas au meilleur nommage de fichier.

Il existe une meilleure manière d'indiquer les versions d'un fichier

1 minute de pénalité

84

Une fois les données publiées en Open Access, l'équipe de recherche s'est intéressée à savoir si elles étaient consultées et réutilisées par d'autres équipes de recherche ou par le grand public.



85

85


Vous consultez les altmetrics du set de données de vos collègues, afin de **6.1 évaluer la consultation et la réutilisation** de leurs données.

Eh bien ! Autant de **vues**, et encore **plus de téléchargements**, ça en fait du monde qui a consulté leurs données ! Ils vont aller loin, avec ces chiffres...






152,377

 views

110,640

 downloads

[See more details...](#)

	All versions	This version
Views 	152,377	0
Downloads 	110,640	0
Data volume 	219.2 TB	0 Bytes
Unique views 	86,653	0
Unique downloads 	27,800	0

[More info on how stats are collected.](#)

88

88



Mauvaise réponse !

Ici, les données sont non standardisées, certaines cellules sont vides, il y a plusieurs variantes pour noter les cellules vides (N/A) et des erreurs au niveau du codage (p. ex. des a au lieu de A)

1 minute de pénalité

92

Les données récoltées ont dû être sauvegardées et stockées convenablement tout au long du projet afin d'éviter les mauvaises surprises. Mais quelles sont les **2.1 Stratégies de sauvegarde et de stockage** à mettre en place durant le projet?

Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 Ordinateur professionnel	★★★★★ Sujet au piratage informatique, aux détériorations et pannes	★★★★★ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud...)	★★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter les données confidentielles et sensibles
 Support externe	★★★★★ - Sujet au vol, à la perte du support - Durée de vie limitée (dégradation du matériel)	★★★★★ Facilement transportable, il permet de transférer les données vers un autre ordinateur	★★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter ou de sécuriser physiquement les données confidentielles et sensibles
 Serveur institutionnel	★★★★★ Stockage fiable, durable et sécurisé (contre le vol, le piratage, les incendies...)	★★★★★ La connexion au serveur institutionnel ne facilite pas le travail avec des personnes extérieures	★★★★★ Coût assez important mais pas forcément répercuté sur l'utilisateur	- Pour un stockage plus pérenne - Adapté pour le stockage de données sensibles et des versions « stables » de vos données - Toutes les institutions ne proposent pas ce service
 Serveur Cloud	★★★★★ On ne sait pas vraiment où sont stockées les données, ni ce qu'elles deviennent	★★★★★ Permet un travail synchronisé avec toutes les personnes ayant été autorisées au partage	★★★★★ Payant à partir d'une certaine limite de stockage	- Pour un partage avec des personnes externes à l'institution - Ne pas y mettre de données sensibles ou confidentielles - Pas de contrôle sur la procédure de sauvegarde des données

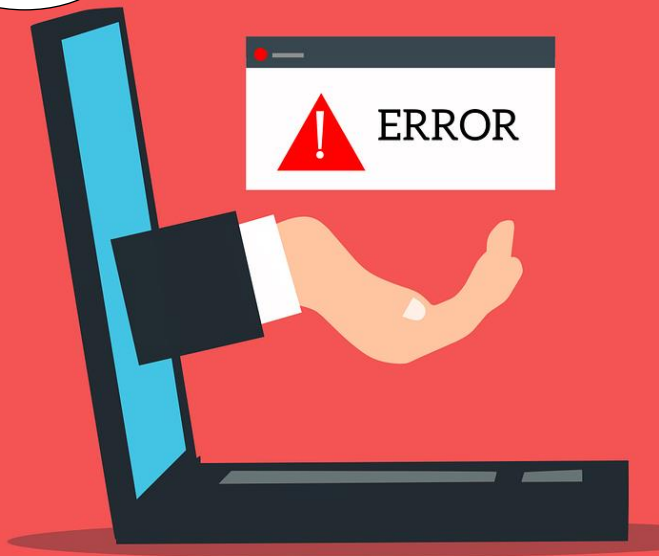
Conseil pour ne pas perdre ses données:

La règle du 3+2+1

3 exemplaires sur 2 supports différents dont au moins 1 copie dans un autre emplacement physique

93

93



Mauvaise réponse !

Dans notre cas, les données doivent être accessibles, en plus de la publication de l'article scientifique. En effet, l'équipe de recherche a souhaité être conforme aux bonnes pratiques de l'Open Access.

1 minute de pénalité

94

94



Mauvaise réponse !

Dans notre cas, l'équipe de recherche ne peut pas rendre les données accessibles uniquement sur demande. Cette pratique ne respecte pas les recommandations de l'Open Access.

1 minute de pénalité

95

95



Mauvaise réponse !

Dans notre cas, les données ne sont pas protégées par un embargo. En effet, un embargo est utilisé dans le cas où la publication se fait avant le partage des données. Cela peut être le cas quand les données sont liées à des enjeux commerciaux, comme un brevet par exemple.

1 minute de pénalité

96

96

71 11

En effet, comme les données ont pu être publiées sans embargo et ne contenaient pas de données sensibles, elles ont pu être déposées sans restriction d'accès. Vos collègues ont donc souhaité que leur **réutilisation** soit possible, à condition qu'ils soient **crédités pour leur travail**.

Pour cela, ils ont dû leur **5.4 attribuer une licence de réutilisation**. Selon vous, laquelle ont-ils choisie pour répondre à ces critères ?



7



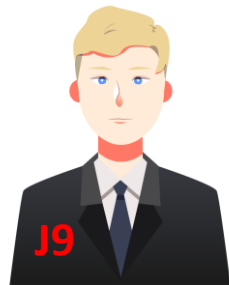
12



35



13



Avez-vous complété
intégralement la feuille de route
pour l'étape en cours ?
Si oui retournez cette carte

99

99

~~11~~ 64

Etape 4 : préservation et archivage



i3

i3

Il existe deux types d'identifiants :

- identifiants objet pour les publications et les données
- identifiants contributeur pour les auteur-e-s et les institutions

Par exemple, Handle et DOI sont des identifiants d'objets numériques. Alors que ORCID et ArXiv author ID sont des identifiants de contributeurs.

Voici des exemples d'identifiants

Handle 20.1000/100

DOI doi:10.1000/182

ORCID 0000-0002-1825-0097

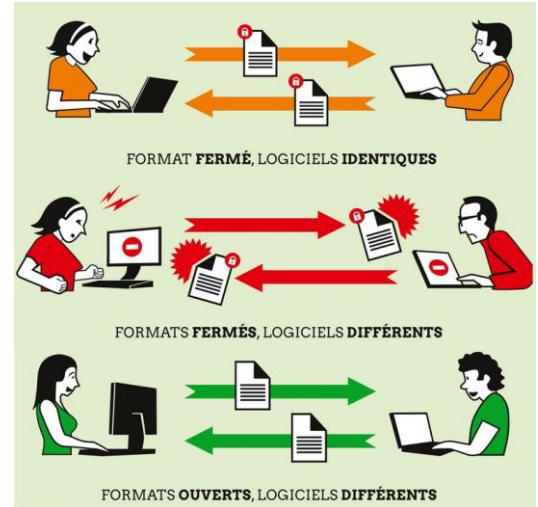
ArXiv author ID arXiv:0706.0001



i22

Un format ouvert est un format de fichier pouvant être lu par plusieurs logiciels. L'utilisation de formats ouverts a deux objectifs principaux :

- Assurer un accès à vos documents sans vous soucier du logiciel utilisé par les personnes qui les consulteront dans le futur.
- Assurer la pérennité de vos documents, afin de pouvoir les ouvrir même si le logiciel utilisé pour les créer n'est pas disponible ou n'existe plus.



i51

i51

Additionnez **tous** les chiffres
des **formats appropriés** pour
trouver la prochaine carte



File formats extensions for reusability / preservation

TYPE OF DATA	APPROPRIATE	ACCEPTABLE	DEPRECATED
Tabular (extensive metadata)	CSV – HDF5	TXT – HTML – TEX – FASTQ – POR	
Tabular (minimal metadata)	CSV – TAB – ODS – SQL – TSV	XML (if appropriate DTD) – XLSX	XLS – XLSB
Textual / Presentation	TXT – PDF – ODT – ODM – TEX – MD – HTM – XML – EXTXYZ – ODF	PPTX – RTF – DOCX – PDF (with embedded forms) – EPS – IPF	DOC – PPT – DVI – PS
Code / Computation	M – R – PY – IYPNB – RSTUDIO – RMD – NETCDF – AIML	SDD	MAT – RDATA
Image & Spectroscopy	TIF – PNG – SVG – JPEG – FITS	JCAMP – JPG – JP2 – TIF – TIFF – PDF – GIF – BMP – DM3 – OIR – LSM	INDD – AIT – PSD – SPC
Audio	FLAC – WAV – OGG – MXL – MIDI – MEI – HUMDRUM	MP3 – AIF	
Video	MP4 – MJ2 – AVI – MKV	OGM – MP4 – WEBM	WMV – MOV – QT
Geospatial	NETCDF – tabular GIS attribute data – SHP – SHX – DBF – PRJ – SBX – SBN – POSTGIS – TIF – TFW – GEOJSON	MDB – MIF	
3D structures & images	X3D – X3DV – X3DB – PDF3D – POV – PDBML	DWG – DXF – PDB	PXP
Generic	XML – JSON – RDF		

i67

i67



Pourquoi nettoyer les données ?

- Pour éliminer les champs superflus
- Pour faciliter leur analyse en assurant leur standardisation

Quels logiciels utiliser ?

Cela dépend du type de données, de leur quantité, des ressources à disposition, des connaissances de l'équipe.

Quelques exemples d'outils gratuits :



OpenRefine

i68

i68

Comment nettoyer les données ?

- Identifier les données essentielles (p. ex. supprimer certains champs ajoutés automatiquement par le système de gestion utilisé pour la collecte)
- Identifier et éliminer les doublons
- Résoudre les valeurs vides
- Assurer la cohérence des données (toutes les données de même type doivent avoir la même forme)

Attention : gardez toujours une version brute de votre fichier, et notez toutes les étapes de nettoyage pour assurer la transparence et la reproductibilité de votre démarche !



J1

J1

Selon Jack, voilà les points à retenir concernant les conditions d'accès aux données



Données fermées

La description des données est publiée, mais l'accès aux données n'est pas accordé.

Ex. le dataset contient des données sensibles non anonymisées /pseudonymisées



Données sur demande

La description des données est publiée mais pour accéder aux données il est nécessaire d'en faire la demande au propriétaire.

Ex. Les données du dataset ont un fort risque de réidentification. Elles sont partagées sous certaines conditions.



Données sous embargo

La description des données est publiée. Mais les données sont inaccessibles une certaine période. A la fin de l'embargo, les données seront soit librement accessibles soit sur demande.

Ex. Le chercheur veut déposer un brevet avant de publier les données.



Données ouvertes

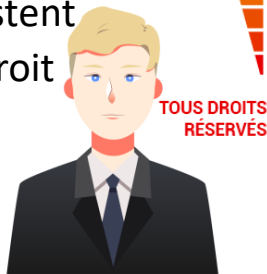
Les données sont accessibles librement pour tout le monde



J9

J9

Les licences creative commons sont des licences de diffusion qui permettent d'accorder à l'avance certains droits d'utilisation sur les œuvres diffusées mais elles restent complémentaires au droit d'auteur.



DOMAINE PUBLIC



BY

ATTRIBUTION (BY)

Le titulaire des droits autorise toute exploitation de l'œuvre à condition de l'attribuer à son auteur en citant son nom.



NC

PAS D'UTILISATION COMMERCIALE (NC)

Le titulaire des droits autorise l'utilisation de l'œuvre à des fins non commerciales.



ND

PAS DE TRAVAUX DÉRIVÉS (ND)

Le titulaire des droits autorise toute utilisation de l'œuvre originale mais n'autorise pas la création d'œuvres dérivées.



SA

PARTAGE À L'IDENTIQUE (SA)

Le titulaire des droits autorise toute utilisation de l'œuvre originale ainsi que la création d'œuvres dérivées, à condition qu'elles soient distribuées sous une licence identique à celle qui régit l'œuvre originale.

Combinaisons possibles



S1

S1

Cette documentation peut être sous différentes formes:

- Fichier Readme.txt
- Codebook (dictionnaire de variables)
- Cahier de laboratoire électronique, aussi appelé *electronic lab notebook* ou ELN)
- Carnet de terrain
- Fichier texte (.docx, .odt, .pdf)
- Intégrée directement dans le fichier de données

Elle doit renseigner sur l'étude en elle-même, l'échantillonnage, les fichiers, la structure et le détail des données au sein du fichier ainsi que sur le texte des questions et réponses.



S2

S2

Comparaison de trois dépôts

Nom	SWISSUbase	OLOS	Zenodo
Numéro dépôt	1	2	3
Organisation	FORS, Université de Lausanne et de Zurich,	DLCM	CERN & OpenAire
Serveur	Suisse	Suisse	Suisse, Hongrie
Discipline	Sciences sociales, linguistiques, en développement pour d'autres disciplines	Multidisciplinaire	Multidisciplinaire (Spécialisation en physique)
Qui peut archiver	Chercheurs affiliés à des institutions de recherche et des universités suisses	Chercheurs affiliés à des institutions de recherche et des universités suisses	Pas de limitation
Format	Tabulaire, textuel, image, audio, vidéo	Large choix de formats	Aucune indication
Licence	Licences au choix (CC)	Licences au choix (CC)	Licences au choix (CC)
Identifiants pérennes	DOI	DOI	DOI
Coûts	Gratuit	Payant	Gratuit
Avantages	<ul style="list-style-type: none"> Architecture basé sur le modèle OAIS Validation des métadonnées et de la documentation par un expert Contrats de dépôt et d'utilisation Métadonnées détaillées sur le projet, les jeux de données et les documents 	<ul style="list-style-type: none"> Architecture basé sur le modèle OAIS Possibilité de rajouter des métadonnées personnalisées Choix du niveau de sensibilité à l'aide de DataTags Choix d'une durée de conservation 	<ul style="list-style-type: none"> Altmetrics Possibilité de créer des collections personnelles Simple d'utilisation
Désavantages	<ul style="list-style-type: none"> Processus de soumission plus complexe Liste formats limités Métadonnées ciblées sciences sociales 	<ul style="list-style-type: none"> Jeunesse du dépôt Payant Stockage non chiffré 	<ul style="list-style-type: none"> Aucun contrôle qualité des données et des métadonnées Taille limitée des jeux de données Seule la personne qui dépose le dataset peut le modifier

Enfin,
l'équipe a choisi
un **dépôt gratuit et**
ayant ses serveurs
en Suisse.

Ajoutez 70
au numéro
du dépôt
choisi.



S5

S5

Comment partager ses données ?

Le dépôt de données sur un entrepôt est important car il permet de stocker ses données dans un endroit distinct, différent de celui de son article scientifique.

Il existe plusieurs types de dépôts : disciplinaires, multidisciplinaires, institutionnels, propres à un éditeur, spécifiques à un projet,...

Pour déterminer son choix, l'équipe de recherche a consulté re3data.org, un annuaire de dépôt de données qui présente ces critères :



S8

S8

Les altmetrics (*article-level metrics* ou *alternative metrics*) sont des indicateurs bibliométriques.

Ils évaluent l'impact sur internet d'une publication ou d'un élément d'information, en observant sa diffusion, les actions et interactions qu'elle engendre en temps réel, par exemple le nombre de téléchargements, de citations, de partage sur les réseaux sociaux...



S10

S10

Il est important de définir, dès la phase préliminaire du projet, des **1.2 Règles de nommage et d'organisation des fichiers**

Règles de nommage

Choisir un nom de fichier court et significatif. <ul style="list-style-type: none"> Max 32 caractères. Élément important en premier. Utiliser des abréviations pour réduire le nbre de caractères. Contient sujet, date, version 	Reunion_20210902_OJ_V01
Éviter les caractères spéciaux, la ponctuation et les caractères accentués	+, =, « », [], < >, \$, %, &, é, è, ê, ï, ç, ?, !, ., ;, ;
Éviter les espaces entre les mots <ul style="list-style-type: none"> Utiliser le tiret souligné "_" (underscore) Utiliser une majuscule 	Regles_Noms_fichiers ReglesNomsFichiers
Indiquer les dates dans le bon format	AAAA-MM-JJ ou AAAA-MM-JJ 20210902 ou 2021-09-02
Indiquer les versions des documents	V01 VP pour version provisoire ou VF pour version finale
Numérations des fichiers <ul style="list-style-type: none"> Utiliser des zéros à gauche pour assurer leur bon ordre séquentiel lors de leur affichage et tri 	Pour une séquence de 1-10 : 01-10 Pour une séquence de 1-100 : 001-010-100

Exemples de conventions de nommage

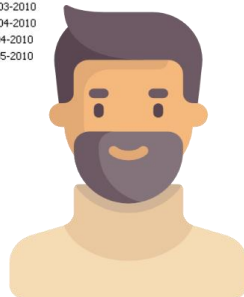
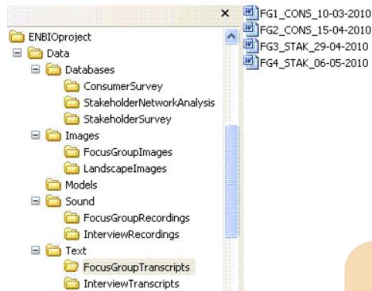
[enqueteur]_[methode]_[sujet]_[YYYYMMDD]_[version].[ext]

[projet#]_[methode]_[YYYYMMDD].[ext]

[YYYYMMDD]_[version]_[sujet]_[collection de donnees].[ext]

[type de fichier]_[auteur]_[YYYYMMDD].[ext]

ou autres conventions à décider au début du projet



S13

S13

Pendant un projet de recherche, de nombreuses données sont créées et il n'est pas possible de toutes les archiver/conserver, pour plusieurs raisons :

- Coûts liés à l'archivage
- Plus le volume de données archivées est élevé, plus leur découverte est compliquée
- Coûts liés au maintien et à la création de métadonnées

Il est donc nécessaire de **4.2 identifier la valeur des données et les données à conserver**



\$15

S15



Le DMP ou Plan de gestion de données est un document qui est rédigé dans la phase préliminaire d'un projet puis complété au fur et à mesure de l'avancée de la recherche. Il **aide à organiser et anticiper toutes les étapes du cycle de vie de la donnée**. Il explique comment les données du projet sont gérées, de la création jusqu'au partage et à l'archivage.

Les étapes que vous allez découvrir dans la suite de vos investigations vont vous permettre de comprendre comment bien remplir chaque partie du DMP.

S25

S25



Vous vous demandez peut-être comment évaluer les données et sélectionner lesquelles doivent être archivées ? Cela dépend de nombreux critères :

Critères liés à la mission de recherche (valeur des données)

- Exigences du bailleur de fonds (ex: FNS)
- Exigences légales
- Exigences de l'éditeur-trice
- Exigence de son institution de rattachement
- Les données soutiennent une publication et des résultats de recherche
- Les données ont un caractère unique
- Les données disposent d'un caractère lié à la notion de patrimoine culturel immatériel
- Originalité des données

Critères liés à la nature de la donnée

- Données brutes
- Données traitées
- Données qui soutiennent une publication et des résultats de recherche
- Données qui synthétisent la recherche

Critères liés aux types de données

- Données d'observation
- Données d'expérimentation
- Données secondaires
- Données négatives

Critères liés aux matériaux qui complètent les données

- Echantillons physiques
- Métadonnées et documentation
- Logiciels utilisés

S28

S28



A l'inverse de la documentation des données qui a pour objectif d'être interprétable par des humains uniquement, les métadonnées doivent pouvoir être lues par des machines afin de permettre leur découverte sur le web, c'est pourquoi elles sont souvent décrites en XML.

Des outils peuvent aider les équipes à rédiger leurs métadonnées au format XML, par exemple le Datacite Metadata Generator.

Ce sont des renseignements standardisés.

Il en existe de nombreux schémas, disciplinaires ou généralistes.