

Ouverture des données et plan de gestion des données de recherche

Journée de rentrée Élico

Bernard Jacquemin

25 septembre 2026



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Introduction – Et la communauté scientifique inventa... la donnée

- **Apparition : collectée, observée, générée ou simulée**
- **Finalité : sert de preuve dans un processus de recherche**
- **Indice pour connaître le réel**
- **Grande variété en SIC**
 - entretiens transcrits, corpus de presse, relevés d'enquête, code source, *notebooks*, traces d'usage, annotations, données bi(bli)ométriques, corpus...
 - « non-numérique nativement » (archives, entretiens) vs. « numérique natif » (traces de plateformes, réseaux sociaux, *logs*)
 - âges de la donnée : donnée brute / donnée traitée / donnée dérivée
 - fraîcheur de la donnée : données primaires / données secondaires

Introduction – Contexte de la science ouverte

- À l'origine : les archives ouvertes (arXiv, Paul Ginsparg 1991 ; BOAI, 2002)
- Étape 2 : les données de recherche (mouvement FAIR, 2016)
- Étape 3 : codes et logiciels, évaluation (Unesco, 2021)
- Plans nationaux pour la science ouverte (2018-2021 ; 2021-2024)
- Principes généraux FAIR (2014, 2016)
 - Findable (métadonnées, identifiant, entrepôt)
 - Accessible (en ligne, métadonnées d'accès)
 - Interoperable (formats, vocabulaires, jeux de métadonnées standards)
 - Reusable (documentées)
- Concrétisation notamment dans un PGD (H2020 / ANR)

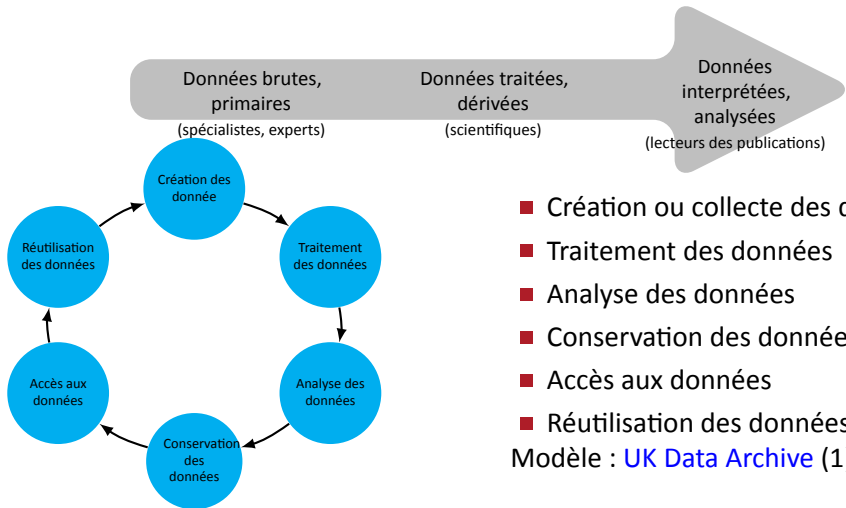
Principe du cycle de vie de la donnée



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Principe du cycle de vie de la donnée



- Création ou collecte des données
- Traitement des données
- Analyse des données
- Conservation des données
- Accès aux données
- Réutilisation des données

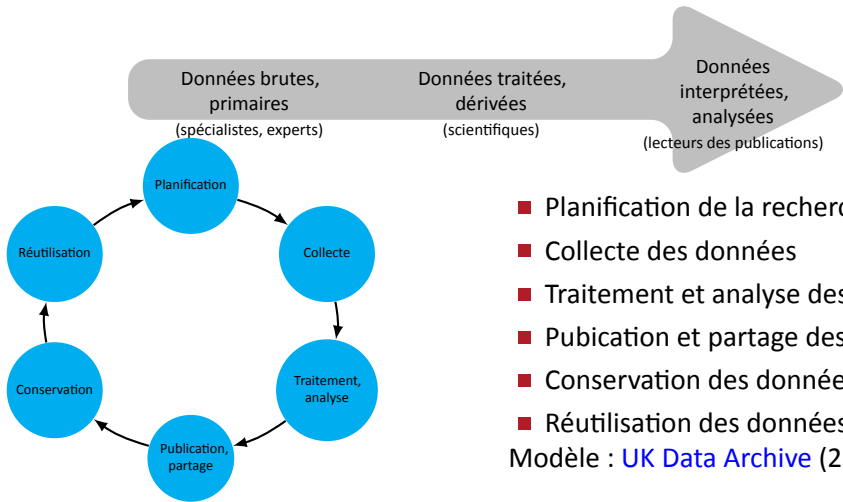
Modèle : [UK Data Archive](#) (1)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Principe du cycle de vie de la donnée



- Planification de la recherche
- Collecte des données
- Traitement et analyse des données
- Publication et partage des données
- Conservation des données
- Réutilisation des données

Modèle : [UK Data Archive](#) (2)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Principe du cycle de vie de la donnée

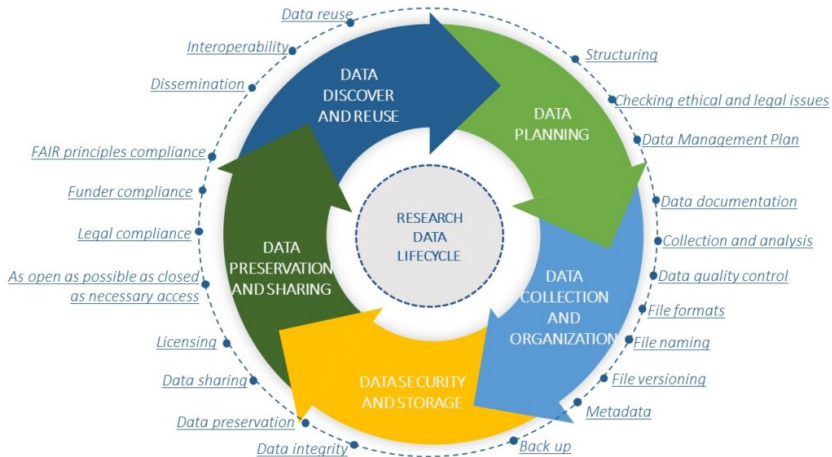


Figure – Source : [Université de Milan](#)



elico

Équipe de recherche de Lyon en sciences de l'information et de la communication

Plan de gestion de... Données? Projet!

Planifier la gestion des données = Anticiper la gestion du projet

1. Quelles données allez-vous collecter ou créer?
2. Comment ces données vont-elles être collectées ou créées?
3. Quelles sont la documentation et les métadonnées qui vont être adjointes à ces données?
4. Comment prenez-vous en compte les enjeux d'éthique?
5. Comment prenez-vous en compte les enjeux de *copyright* et de propriété intellectuelle?
6. Comment les données vont-elles être stockées et sauvegardées durant la recherche?
7. Comment prenez-vous en compte l'accès et la sécurité des données?
8. Quelles sont les données qui devront être protégées, partagées et/ou conservées?
9. Quel est le programme de conservation des jeux de données sur le long-terme?
10. Comment allez-vous partager les données?
11. Y a-t-il des restrictions réglementaires de partages sur des données?
12. Qui est responsable de la gestion des données?
13. Quelles sont les ressources nécessaires pour mettre en œuvre ce programme?

Un formulaire, un document administratif... vivant !

1. Description des données et collecte

- Types de données (texte, image, séquençages génétiques, logiciels, enregistrements audio, enquêtes)
- Origine (créées, réutilisées, combinées)
- Formats (plutôt ouverts, plutôt simple, plutôt qualitatifs)
- Volumes (anticiper les besoins structurels)

2. Documentation et qualité des données : rendre les données compréhensibles

- Métadonnées standards (Dublin Core, EXIF...)
- Documentation : README, structure et contexte, objectifs, démarche
- Contrôle qualité : suivi des données (méthode et outils de création et de saisie)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Un formulaire, un document administratif... vivant !

3. Stockage et sécurité **pendant** le projet

- Lieux de stockage : disques et clefs USB (*non !*), serveurs ou cloud institutionnels (ou pas)... sécurisé !
- Modalités de sauvegarde (fréquence, automatisation ; règle du 3 copies-2 supports-1 ailleurs)
- Droits d'accès (identification des utilisateurs, droits, chiffrement et confidentialité)

4. Exigences légales, éthiques et conformité [réglementation]

- RGPD et protection de la personne (anonymisation, pseudonymisation, consentements)
- Propriété intellectuelle (chercheurs, institutions, partenaires privés)
- Comités d'éthique et comités de protection des personnes (protection + légitimité) : avant ! – CNIL



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Un formulaire, un document administratif... vivant !

5. Partage, diffusion et réutilisation [concrétisation de FAIR]

- Modalités de partage (ouverture et limites, plateformes Nakala, RDG, Zenodo...)
- Calendrier de mise à disposition (embargo ?)
- Choix des licences (CC, licences *Open Data* comme Etalab LO...)
- Attribution d'un identifiant pérenne (DOI, Handle, ARK...), « citabilité »

6. Conservation à long terme et archivage

- Critère de choix de la conservation (et de l'élimination – anticipation nécessaire !)
- Modalités d'archivage (infrastructure, institution garante)
- Coût et responsabilité (frais de conservation, gestion pendant l'archivage à long terme)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Annexes

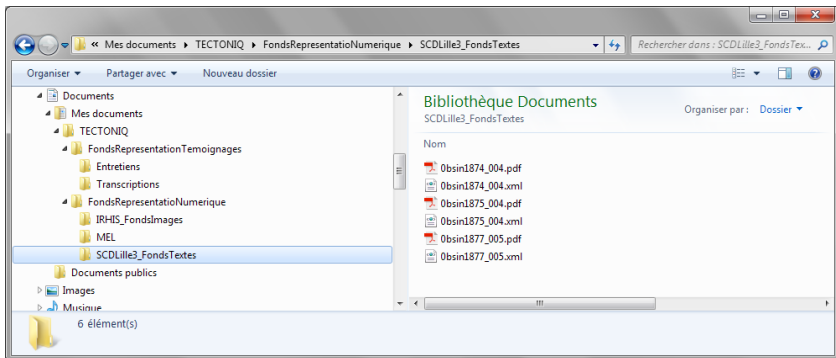


elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Concevoir une arborescence de fichiers / dossiers
- Utiliser des noms de fichiers / dossiers informatifs et clairs



elico

Equipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux

```
questionnaire brut.doc  
questionnaireSujet1  
accordéon/flûte.flac  
image:cuivres & percussions.png  
exemple.doc-2
```

- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux

Nom correct

Nom erroné

Danger

questionnaire brut.doc
questionnaireSujet1
accordéon/flûte.flac
image:cuivres & percussions.png
exemple.doc-2

- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux

Nom correct	Nom erroné	Danger
questionnaireBrut.doc	questionnaire brut.doc	
questionnaireSujet1.txt	questionnaireSujet1	
accordeon_flute.flac	accordéon/flûte.flac	
img_cuivresPercussions.png	image:cuivres & percussions.png	
exemple-2.doc	exemple.doc-2	

- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux

Nom correct	Nom erroné	Danger
questionnaireBrut.doc	questionnaire brut.doc	questionnaire%20brut.doc
questionnaireSujet1.txt	questionnaireSujet1	pas d'application associée
accordeon_flute.flac	accordéon/flûte.flac	accordÃ©on/flÃ»te.flac
img_cuivresPercussions.png	image:cuivres & percussions.png	image:cuivres%20&esp;%20percussions
exemple-2.doc	exemple.doc-2	pas d'application associée

- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu

`mon_tableau.xlsx`

`2.wav`

- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
 - mon_tableau.xlsx → reponsesQuestionnaireSujet7.xlsx
 - 2.wav → entretien_2_ARM.wav
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout

- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
 - entretienARM-1.wav
 - interview-2-ARM.wav
 - 3ARM-echange.wav
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
 - entretienARM-1.wav → interview-1-ARM.wav
 - interview-2-ARM.wav → interview-2-ARM.wav
 - 3ARM-echange.wav → interview-3-ARM.wav
- noms courts : 256 caractères maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : **256 caractères** maximum pour **tout le chemin**



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**

`./TECTONIQ/FondsRepresentationTemoignages/Transcriptions/Entretien 2 ARM.txt`



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Standardisation du nommage des fichiers

- noms simples : pas d'accents, d'espace ni de caractères spéciaux
- noms informatifs : il faut que l'on puisse identifier facilement le contenu
- noms standardisés : même présentation partout
- noms courts : 256 caractères maximum pour **tout le chemin**

./TECTONIQ/FondsRepresentationTemoignages/Transcriptions/Entretien 2 ARM.txt
./TECTONIQ/FondsR~1/Transc~1/Entret~3.TXT → abrègement automatique si trop long

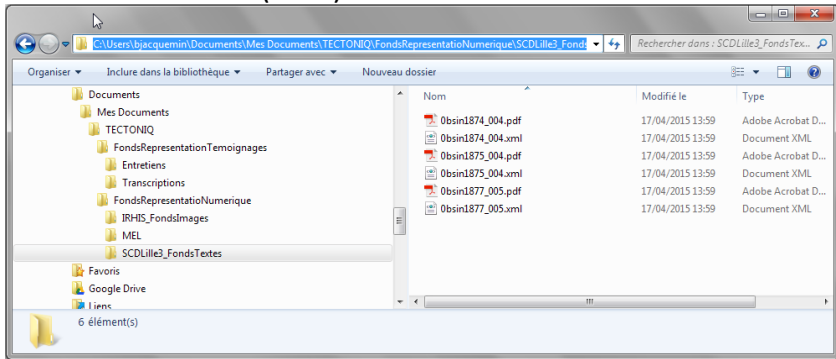


elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

■ Visualiser le « chemin » (*Path*)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)

```
                                README.TXT
*****
Organisation des données du corpus test (textes, images et sons)
accompagnant le projet TECTONIQ pour la valorisation des données liées
au patrimoine industriel textile.
*****
./section_12_creation_evaluation_de_voix :
| Rassemble des échantillons sonores originaux et synthétiques pour
| chacun des 49 dictionnaires de synthèse impliqués dans l'évaluation
| perceptive de la section 12.
|
|-----/originaux :
| Echantillons originaux des voix enregistrées. Les noms de
| fichiers sont de la forme "NomVoix_NomScript_nat.wav"
|
|-----/phoauto_segauto :
| Exemples de synthèse obtenus avec une segmentation phonétique
| entièrement automatique des données acoustiques. Les noms de
| fichiers sont de la forme "NomVoix_NomScript_synth_pa_sa.wav"
```



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)

Sexe	Prénom	Année de naissance
M	Alphonse	1932
F	Béatrice	1964
F	Charlotte	1970

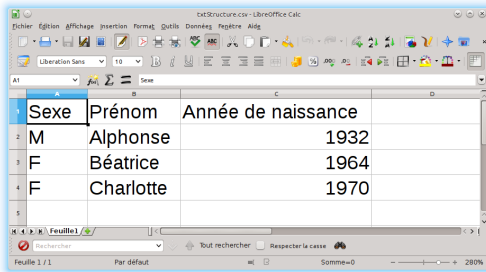


elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)



	A	B	C	D
	Sexe	Prénom	Année de naissance	
1	M	Alphonse	1932	
2	F	Béatrice	1964	
3	F	Charlotte	1970	
4				
5				

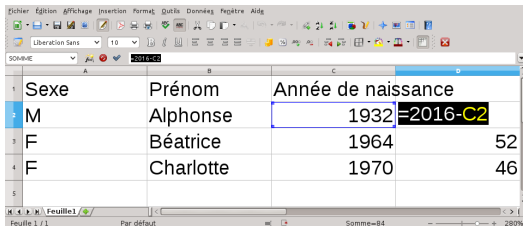


elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)



The screenshot shows a Microsoft Excel spreadsheet with the following data:

	A	B	C	D
1	Sexe	Prénom	Année de naissance	
2	M	Alphonse	1932	=2016-C2
3	F	Béatrice	1964	52
4	F	Charlotte	1970	46
5				

The formula bar at the top shows the formula `=2016-C2` for cell D2. The status bar at the bottom indicates 'Somme=84' and '280%' zoom.



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)

```
<HTML>
  <HEAD>
    <TITLE>Le titre de la fenêtre</TITLE>
  </HEAD>
  <BODY BGCOLOR="CCCCCC">
    <CENTER><IMG SRC="geriico2.png" ALIGN="BOTTOM"> </CENTER>
    <HR>
    <a href="http://www.google.fr">Google</a> est un lien vers un moteur de recherche.
    <H1>Un titre de premier niveau</H1>
      <H2>Puis un titre de deuxième niveau</H2>
        Envoyez-moi un E-mail à <a href="mailto:Bernard.Jacquemin@univ-lille3.fr">mon
          adresse</a>.
        <P> Un paragraphe
        <P> <B>Et un autre paragraphe</B>
        <BR> <B><I>Une phrase sans changement de paragraphe,
          en gras italiques.</I></B>

    <HR>
  </BODY>
</HTML>
```



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

- Un format de fichier pérenne, ouvert, de référence (pour la conservation et la diffusion)

```
<?xml version="1.0" encoding="UTF8"?>
<annuaire>
  <personne>
    <nom>HEUTE</nom>
    <prenom>Thomas</prenom>
    <email>webmaster@xmlfacile.com</email>
  </personne>
  <personne>
    <nom>CANTAT</nom>
    <prenom>Bertrand</prenom>
    <email>noir@desir.fr</email>
  </personne>
</annuaire>
```



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

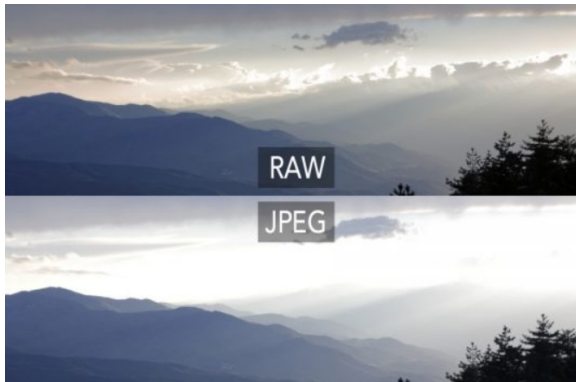
De bonnes pratiques !

- Privilégier les formats non propriétaires (.odt plutôt que .doc, .svg plutôt que .ai) ou largement répandu (.pdf pour un texte à stabiliser)
- Privilégier les formats sans perte (.raw, .tiff plutôt que .jpg – et même .png)



De bonnes pratiques !

- Privilégier les formats non propriétaires (.odt plutôt que .doc, .svg plutôt que .ai) ou largement répandu (.pdf pour un texte à stabiliser)
- Privilégier les formats sans perte (.raw, .tiff plutôt que .jpg – et même .png)



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

De bonnes pratiques !

Une description des données telles qu'elles sont préservées – et collectées

- Un README . txt pour comprendre cette structure et les pratiques associées

De bonnes pratiques !

Une description des données telles qu'elles sont préservées – et collectées

- Un README.txt pour comprendre cette structure et les pratiques associées

```
                                README.TXT
*****
Organisation des données du corpus test (textes, images et sons) accompagnant le projet
TECTONIQ pour la valorisation des données liées au patrimoine industriel textile.
*****
./section_12_creation_evaluation_de_voix :
| Rassemble des échantillons sonores originaux et synthétiques pour chacun des
| 49 dictionnaires de synthèse impliqués dans l'évaluation perceptive de la
| section 12.
|
|-----/originaux :
| Echantillons originaux des voix enregistrées. Les noms de fichiers sont
| de la forme "NomVoix_NomScript_nat.wav"
|
|-----/phoauto_segauto :
| Exemples de synthèse obtenus avec une segmentation phonétique
| entièrement automatique des données acoustiques. Les noms de fichiers
| sont de la forme "NomVoix_NomScript_synth_pa_sa.wav"
|
|-----/phomanu_segauto :
| Exemples de synthèse obtenus avec une transcription phonétique révisée
| par un expert et une segmentation automatique. Les noms de fichiers
```

Décrire les données de la recherche : Dublin Core

Nom de l'élément en français	Décrit	Nom de l'élément en anglais
Titre Sujet Description Source Langue Relation Couverture	contenu	Title Subject Description Source Language Relation Coverage
Créateur Éditeur Contributeur Droits	propriété intellectuelle	Creator Publisher Contributor Rights
Date Type Format Identifiant	instance particulière	Date Type Format Identifier



elico

Équipe de recherche de Lyon en sciences
de l'information et de la communication

Décrire les données de la recherche : Dublin Core

Element refinement	Element refined
Abstract	description
Access Rights	rights
Alternative	title
Available	date
Bibliographic Citation	identifier
Conforms To	relation
Created	date
Date Accepted	date
Date Copyrighted	date
Date Submitted	date
Audience Education Level	audience
Extent	format
Has Format	relation
Has Part	relation
Has Version	relation
Is Format Of	relation

Element refinement	Element refined
Is Part Of	relation
Is Replaced By	relation
Issued	date
Is Version Of	relation
License	rights
Mediator	audience
Medium	format
Modified	date
References	relation
Replaces	relation
Requires	relation
Spatial	coverage
Table Of Contents	description
Temporal	coverage
Valid	date



elico

Equipe de recherche de Lyon en sciences de l'information et de la communication

En bonus

Pour trouver des entrepôts de données de recherche (cf. [Dorandum](#) ou [University of Alabama](#)).

Des entrepôts génériques

- Re3data : <https://www.re3data.org/>
- Datacite <https://repositoryfinder.datacite.org/>
- Fairsharing <https://fairsharing.org/search/?q=repository>
- CatOpidor https://cat.opidor.fr/index.php/Entrep%C3%B4t_de_donn%C3%A9es

Des entrepôts plus orientés vers les sciences humaines et sociales

- Nakala <https://www.nakala.fr/>
- Zenodo <https://www.zenodo.org/>
- Research Data Gouv <https://recherche.data.gouv.fr/>
- UK Data Archive <http://www.data-archive.ac.uk/deposit>
- ICPSR - Inter-university Consortium for Political and Social Research
<http://www.icpsr.umich.edu/icpsrweb/>
- ADA - Australian Data Archive <http://www.ada.edu.au/>