Guide de bonnes pratiques de partage et de valorisation des données linguistiques

Thomas Gervais d'Aldin

18 juin 2024

Table des matières

Introduction

1. Contexte au LT2D
   (a) Le LT2D développe des projets de recherche dans le domaine des lexiques, des textes, des discours et des dictionnaires.
   (b) La CPJ a pour objectif le renforcement des travaux autour des Humanités Numériques au LT2D par la création de nouvelles ressources et par
   (c) Un objectif plus spécifique : la création et la diffusion de ressources visant à **valoriser les travaux de recherche du labo**
   (d) Parmi ces nouvelles ressources : le guide
2. Objectifs à l'issue de l'étude du guide
   (a) Comprendre les enjeux de la Science Ouverte et l'intérêt du chercheur à porter attention aux bonnes pratiques dans la constitution et dans l'usage de ses données
   (b) Contexte éthique et légal à définir en amont
   (c) Stratégies de planification avant la constitution d'un jeu de données
   (d) Différents leviers de valorisation des jeux de données
   (e) Re-direction vers diverses ressources pour approfondir en fonction de ses besoins

Liens `https://lt2d.cyu.fr/version-francaise/politique-scientifique-et-themes-de-recherche/themes-de-rech`
`https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte-pnso/`

Mots clé Science Ouverte, données de la recherche

Enjeux de la valorisation des données de la recherche

(Introduction partie 1)

1. Définir les données de la recherche
   (a) Définition officielle de l'OCDE : ensemble des informations collectées, produites et utilisées dans le but d'un travail scientifique.
   (b) Mais définition large > Elle est à adapter à chaque domaine. (Rebouillat 2019)
   (c) En linguistique, on les définit par un ensemble d'enregistrements écrits, ou oraux
   (d) Les volumes croissants de données linguistiques induisent le besoin d'organisation et de gestion des corpus (Minel 2017)
2. Pourquoi diffuser les données de la recherche ?
   (a) Poser le contexte, faire référence au récent Plan National pour la Science Ouverte (Open data)
   (b) Le partage et l'ouverture des données de la recherche favorisent leur réutilisation par vous-même et par les autres : des membres de l'équipe de votre projet, de votre équipe de recherche, communauté scientifique dans son ensemble. (CF : Ouvrir la science)
   (c) Des données bien constituées = plus de **visibilité** (articles scientifiques basés sur des données ouvertes sont 25% plus cités)
   (d) Des données bien constituées = plus **reproductibles**, donc atteste de l'intégrité scientifique, en les rendant les hypothèse plus facilement vérifiables. Si les données bien constituées ne sont pas utilisées entièrement ou du tout dans le travail du chercheur, elles pourront être reprises par d'autres, possiblement dans d'autres disciplines.

Liens `https://www.ouvrirlascience.fr/wp-content/uploads/2024/03/24-02-22-Donnees-FR-WEB.pdf`
`https://shs.hal.science/tel-02447653v1`
`https://shs.hal.science/halshs-01590750`

Mots clé données de la recherche

1.1 Sauvegarder les travaux

1. Contexte
   (a) La préservation de l'information est une phase essentielle de la gestion des données. Le dépôt des données dans un entrepôt assure leur sauvegarde et leur disponibilité ce qui facilite leur partage et leur réutilisation.
   (b) Jusqu'à récemment, les données de la recherche avaient une durée de vie très courte, car des standards de conservation n'étaient pas encore définis.
   (c) Conserver les données dans un environnement sécurisé
   (d) Gestion des modalités de partage par l'attribution de licences de diffusion
2. Degrés de préservation 3 grands concepts : le stockage de la sauvegarde et de l'archivage (graduel du - pérenne au + pérenne)

Le stockage Le dépôt des données sur un support numérique accessible, physique (disque personnel ou partagé, serveur) ou dématérialisé (cloud) dans un but d'exploitation à court terme

↓

La sauvegarde La duplication des données sur un support externe à celui sur lesquelles elles sont stockées. Elle sécurise les données à moyen terme, permettant de les restaurer en cas de dégradation ou de perte ou d'inaccessibilité au support de stockage. La sauvegarde fait l'objet d'une stratégie définissant la fréquence des sauvegardes en fonction de la criticité des données pour le projet.

↓

L'archivage L'archivage consiste à ranger un document dans un lieu où il sera conservé pendant une période plus ou moins longue et d'y associer les moyens pour réutiliser les données : la réutilisation se faisant en ajoutant de l'intelligence à la sauvegarde. Le contenu des documents archivés n'est pas modifiable. Par contre le contenant (format) des documents archivés peut être modifié (pour éviter l'obsolescence logicielle)

3. Les bonnes pratiques à mettre en place
   (a) Avant de commencer la constitution, prévoir multiples supports de sauvegarde : sur l'ordinateur de travail directement, sur un disque externe pour un back-up, sur une plate-forme cloud (de préférence spécialisée, éviter les outils GAFAM.)
   (b) Conserver un historique des sauvegardes (type git)
   (c) Privilégier sur les dépôts spécialisés pour pérenniser les données. (NAKALA, ORTHOLANG ...)

Liens (Hadrossek et al. 2023) https://bu.univ-lille.fr/chercheurs-doctorants/science-ouverte/donnees-de-recherche/sauvegarde-et-stockage-des-donnees

Mots clé données de la recherche, Entrepôt de données de la recherche, criticité,sauvegarde,archives

1.2 Donner de la visibilité à ses travaux

1. Respect des bonnes pratiques = plus de visibilité
   (a) Identité numérique cohérente
   (b) DOI lié à chacun de ses travaux
   (c) Importance de rendre les références à ses travaux accessibles et lisibles par les machines, au delà de simplement ses travaux
2. Meilleure garantie de transparence de la recherche
   (a) CF la réutilisation plus loin : si les travaux sont dispo et bien référencés, ils seront plus probablement repris pour d'autres travaux ou pour reproduction de l'experience
   (b) Valoriser le travail de l'équipe en matière de production, de gestion, de description et de partage des données de recherche. L'attribution d'identifiants pérennes tels que l'identifiant ORCID pour les chercheuses et chercheurs et le DOI pour les jeux de données facilite la citation et la mise en visibilité des données produites.
3. Modes de publication
   (a) Privilégier la publication en revue en accès ouvert
   (b) Dépôt dans une archive ouverte
4. Autres leviers de visibilité
   (a) Réseaux sociaux académiques
   (b) Réseaux spécialisés

Liens https://libguides.biblio.usherbrooke.ca/valorisation/recherche/visibilite
https://oa100.snf.ch/wp-content/uploads/2020/09/augmenter-visibilite-impact-publication-scientifiq
pdf
https://www.ouvrirlascience.fr/wp-content/uploads/2023/10/Livret_pour_impression.pdf

Mots clé données de la recherche, bibliométrie, h-index

1.3 Assurer la réutilisation

1. Pourquoi les réutiliser ?
   (a) Trop de données ne sont pas visibles / réutilisées
   (b) Coûts financiers et humains importants à la production de nouvelles données
   (c) Nouvelle analyse sur un autre projet de recherche
2. Question économique
   (a) La réutilisation dans le cadre d'appels à projets peut être fortement encouragée voire requise pour le financement
   (b) La vérification de la conformité étique et légale des données est déjà faite.
3. Exploitation du plein potentiel scientifique des données
   (a) Faire un comparatif avec un nouveau jeu de données (complètement différent ou l'ancien jeu mis à jour)
   (b) L'évolution des outils d'évaluation peut apporter un nouveau regard sur les recherches passées
   (c) La reproductibilité d'une théorie est essentielle à sa validation.
4. Les bonnes pratiques à mettre en place
   (a) Choisir une licence de diffusion adaptée à ses données, de façon à ce qu'elle respectent le principe *" le plus ouvert possible, aussi fermé que nécessaire "*

Liens `https://sciencespo.libguides.com/donnees-de-la-recherche/endetails/Reutilisables`
`http://...`

Mots clé Données de la recherche, Licence de diffusion

1.4 Sécuriser le financement

Financement public Beaucoup d'infos sur le site de l'ANR

1. Appels à projet
   (a) Financement public > Crédits alloués sur la base d'appels à projets compétitifs
   (b) Grands piliers des processus de selections : **Intégrité, déontologie, confidentialité des informations, et transparence des processus**
2. Respect des principes de la science ouverte > favorise le financement.
   (a) Petit-a
   (b) Petit-b
   (c) Petit-c
3. En pratique
   (a) Planification type DMP fortement recommandée ou imposée par la tutelle ou l'agence de financement d'un projet. (plus loin)
   (b) Mise à disposition des données, pouvant être planifiée dès l'appel à projet (via un DMP ou autre)
   (c) Petit-c

Liens https://anr.fr/fr/lanr/nous-connaitre/processus-de-selection/
http://...
http://...

Mots clé PGD,Science Ouverte

1. Conditions
   (a) Planification type DMP fortement recommandée ou imposée par la tutelle ou l'agence de financement d'un projet.
   (b) Crédits alloués sur la base d'appels à projets compétitifs

Préparatifs à la constitution d'un jeu de données

(Introduction partie 2)

1. Idée de faire lma
   (a) Petit-a
   (b) Petit-b
   (c) Petit-c
2. Grand-2
   (a) Petit-a
   (b) Petit-b
   (c) Petit-c
3. Grand-3
   (a) Petit-a
   (b) Petit-b
   (c) Petit-c

Liens http://...
http://...
http://...

Mots clé

2.1 Principes étiques et légaux

2.1.1 Les principes FAIR

1. Faciliter la découverte des données
   (a) Les données ont un identifiant unique pérenne (PID ou DOI) pour assurer l'accès à la ressource.
   (b) Les données sont décrites par des métadonnées scientifiques et documentaires
   (c) Les données, ou au moins leurs métadonnées, sont indexées ou enregistrées dans un outil de recherche, par exemple à travers le dépôt des données
2. Interopérabilité des données aux différents environnements informatiques utilisés par les humains et les machines.
   (a) Élaboration d'un vocabulaire contrôlé (glossaire, lexique, liste de mots clés)
   (b) Faire référence aux autres données en relations, citées ou non dans le travail dans les métadonnées
3. Accès aux données
   (a) Sur internet en libre consultation
   (b) Sur des plateformes et dépots, aux membres en accès restreint
4. Réutilisation des données
   (a) Attribution de licences de réutilisation
   (b) Provenance des données
   (c) Structure conformes aux standards de la communauté pour faciliter leur ré-emploi et leur analyse

Liens https://www.ccsd.cnrs.fr/principes-fair/
https://www.ouvrirlascience.fr/fair-principles/
http://...

Mots clé données de la recherche, FAIR, Science Ouverte, métadonnées,licence de diffusion

2.1.2 Cycle de vie des données

Définition Le cycle de vie des données présente le processus de production, d'utilisation et de conservation ou destruction des données dans une organisation. Il liste les différentes étapes et les acteurs intervenants. Le cycle de vie des données s'applique à l'ensemble des données des organisations. Il permet de repérer la manière d'utiliser les données en fonction de leurs caractéristiques et de préciser les différents usages des données en fonction de leur spécificité. Il présente les différentes interventions nécessaires tout au long de la vie des données dans et hors de l'organisation.

1. Étapes

Planification

— Définir le projet de recherche et anticiper les prochaines étapes du cycle de vie des données
— Anticiper la façon dont les données seront obtenues et stockées pour faciliter la traçabilité en amont afin de permettre la réutilisation des données

Collecte

— Les données utilisées peuvent avoir plusieurs origines : elles peuvent être créées, modifiées, réutilisées

Organisation et analyse

— Organiser ses données pendant le projet est une étape importante car elle facilitera la gestion du cycle de vie
— Garantir l'identification, la localisation, la protection et l'accès à ces données

Conservation

— Mise en sécurité et sureté des données traitées
— Multiplication des supports de sauvegarde (physiques, serveurs, cloud)

Partage

— Une fois que les données d'un projet sont nettoyées et stabilisées, il est important de penser à les publier
— Les données de la recherche peuvent être publiées via un dépôt disciplinaire, institutionnel ou plus généraliste tel que l'entrepôt national Recherche Data Gouv

— Recommandé de publier ces jeux de données dans un entrepôt sécurisé générant automatiquement un DOI
(Digital Object Identifier)

Réutilisation

— Elles peuvent servir à d'autres travaux scientifiques permettant de faire avancer ou tester de nouvelles hypothèses

2. En savoir plus sur les bonnes pratiques

(a) Faire un inventaire des outils et sites informatifs.

Liens https://opendatafrance.gitbook.io/kit-de-ressources-odf/fiches-pratiques/comprendre/comprendre-le-
https://www.universite-paris-saclay.fr/recherche/science-ouverte/le-cycle-de-vie-des-donnees
http://...

Mots clé données de la recherche, PGD, RGPD, DOI

2.1.3 RGPD

1. Contexte

(a) Multiplication récente des masses de données numériques

(b) Des données à caractère personnel et intime sont générées par tous nos services (médical, bancaire, professionnel ...)

(c) Il fallait que le système législatif s'adapte à cette transformation

(d) En 2012, l'Europe lance un chantier de grande rénovation du cadre législatif qui donnera suite à l'adoption en 2016 puis la mise en application en 2018 du RGPD

(e) Son respect est un fondement éthique de la recherche en SHS

(f) Le non respect du RGPD expose à des sanctions juridiques et financières

2. Définition
Le Règlement Général sur la Protection des données a été conçu pour protéger les citoyens dans le contexte des GAFAM qui utilisent en masse nos données personnelles. Il encadre le traitement de données à caractère personnel c'est à dire " la collecte, l'enregistrement, l'organisation, la structuration, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, la diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, la limitation, l'effacement ou la destruction " (art. 4 du Règlement européen du 27 avril 2016)

3. Degrés de sensibilité des données

Moins protégées	Non personnelles	Les données non personnelles sont des données qui n'ont pas besoin de protection particulière. (Ex : mail d'accueil d'une entreprise, adresse d'une entreprise ...)
↓	Personnelles	Une donnée à caractère personnel désigne toute information se rapportant à une personne identifiée ou identifiable. Une personne est identifiable quand elle peut être identifiée directement ou indirectement. Directement : avec un nom, une photo, vidéo Indirectement : par recoupement de plusieurs données, par exemple, grâce à une date de naissance **et** une adresse postale
Plus protégées	Sensibles	Les données dites sensibles sont de nature confidentielles. Leur traitement est très strictement encadré par le Règlement Européen, et nécessite le consentement explicite de la personne concernée (Exemple : santé physique ou mentale, appartenance à un syndicat, opinions politiques ou religieuses, origine ethnique, données génétiques, biométriques...)

4. Stratégies de désidentification des données Souigner l'importance de choisir une stratégie adaptée à son projet si nécessaire.

Anonymisation

- Consiste à rendre **impossible** toute identification de la personne
- Deux principales méthodes d'anonymisation

La **randomisation** : Modifier les attributs de telle sorte qu'ils soient moins précis, tout en conservant la répartition globale. Par exemple, si l'on permute les dates de naissances des individus, on empêche le recoupement avec cette donnée, mais l'on peut conserver la répartition des âges dans l'échantillon. Il convient de mentionner l'action dans la description des données

La **généralisation** : Modifier l'échelle des attributs ou leur ordre de grandeur, afin de s'assurer qu'ils soient commun à un ensemble de personne. Par exemple, en agrégeant des adresses postales à l'échelle d'une ECPI ou d'une ville.

Pseudonymisation

- C'est un traitement de données réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire (comme un identifiant, un alias ou un numéro séquentiel).

5. Pour quoi faire ?
   (a) S'assurer qu'on reste dans le cadre de la loi
   (b) Bon respect des pratiques éthiques, donc plus de visibilité, plus de chances que les travaux soient cités ou repris.

Liens https://oeilpouroeilcreations.fr/formations/gdpr
https://www.ofis-france.fr/espaces-thematiques/integrite-scientifique-ethique-de-la-recherche-deon
https://oxfamilibrary.openrepository.com/bitstream/handle/10546/621092/gd-research-ethics-practical
pdf;jsessionid=C534B7C3F2572B2BA687EF10C277C8E7?sequence=17
https://u-paris.fr/societes-humanites/deontologie-ethique-de-la-recherche-et-integrite-scientifiqu
ethique-de-la-recherche/
(Delmotte 2016)
(Bouchet Moneret 2021)

Mots clé anonymisation,pseudonymisation,RGPD

2.2 Planification et spécificités liées aux pratiques

2.2.1 Élaborer un DMP (Plan de Gestion de Données)

1. Définition Le Data Management Plan (DMP) ou Plan de Gestion de Données (PGD) est un document synthétique qui aide à organiser et anticiper toutes les étapes du cycle de vie de la donnée. Il explique pour chaque jeu de données comment seront gérées les données d'un projet, depuis leur création ou collecte jusqu'à leur partage et leur archivage.
2. DMP Opidor
   (a) Outil mis à disposition par le CNRS
   (b) Accès à des modèles de DMP
   (c) Facilite la rédaction
   (d) Guides et exemples personnalisés

Liens https://dmp.opidor.fr/
https://doranum.fr/plan-gestion-donnees-dmp/plan-de-gestion-des-donnees-fiche-synthetique_10_13143_cgv4-0k53/
http://...

Mots clé PGD,

2.3 Pratiques propres aux différents formats

Corpus écrits

1. Les pratiques
   (a) Dictionnaires

Mise en place du standard

— Au début de la numérisation, chaque maison d'édition à son propre langage
— Premier standard > le SGML, mais trop permissif > XML > création et adoption du format TEI

Avantages de la TEI

— Permet d'encoder des textes et documents numériques, particulièrement répandu dans les SH
— Shéma de balisage riche et flexible, adapté aux besoin des lexicographes
— Oblige la présence d'un header (md), ce qui permet une bonne tracabilité/normalisation
— Possibilité de hierarchie des documents
— La TEI est active et maintenue.

à introduire

— Omeka

Lexiques

—
— Formats LMF : Standard ISO pour les lexiques du TAL.
— Formats de tables classiques
— Format CONLL : décrit des données textuelles sous forme de colonne selon un nombre d'attributs catégorie d'entité nommée, nature grammaticale.
— à approfondir quand tu vas explorer les lexiques ORTOLANG

Thesaurus

— Format skos :

Bitextes

— format TMX, ou btxt.
—
— Pour ML, il existe des solutions open source pour entrainer ses propres modèles (openNMT), plutot facile d'accès, mais demande de grosses quantités de données avec d'obtenir des résultats satisfaisants

Corpus de copies d'élèves

— Corpus E-CALM écriture scolaire (ORTOLANG)

notes

— Dans quelle catégorie pour le format CONNL (commun à quelles pratiques ?)

Liens (Mangeot et Enguehard 2013)
https://omeka.org/
http://www.lexicalmarkupframework.org/

Mots clé TEI,teiHeader, langages de balisage, XML, TMX, métadonnées, LMF

Corpus oraux

1. Pré requis et méthodologie de constitution
   (a) Pour constituer un corpus oral, le linguiste doit définir une méthodologie :
      — **Objectifs de recherche visés** : si l'utilisateur veut construire un corpus pour étudier le vocabulaire des jeunes, il ne choisira pas les situations d'enregistrement de la même façon que s'il veut travailler sur les interactions de service.
      — **Type de corpus** : si le but est de constituer un corpus de référence, plusieurs critères interviennent en parallèle pour obtenir une meilleure représentativité possible.
      — **Modalités d'enregistrement des données** : si le corpus est construit pour travailler sur les caractéristiques acoustiques d'un son, il est important que les enregistrements soient faits dans des situations expérimentales optimales
   (b) Formulaires présentés aux volontaires pour enregistrement
   (c) Protocoles (TFC ...)
2. Autres éléments à intégrer
   (a) Présentation du Corpus de référence du français parlé (Delic, Teston-Bonnard et Véronis 2004)
   (b) Les grands corpus du français moderne (Wissner 2012)
   (c) Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France (Jacobson et Baude 2011)
3. Peut-être aborder ces sujet (attention listing)

Enregistrements
Discours
Vidéo

1. Présentation du Corpus de référence du français parlé (Delic, Teston-Bonnard et Véronis 2004)
2. Les grands corpus du français moderne (Wissner 2012)
3. Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France (Jacobson et Baude 2011)

Enregistrements
Discours
Vidéo
Liens http://...
http://...
http://...

Mots clé

Ressources et outils

3.1 Se former

3.1.1 Se former en ligne

Se former à la science ouverte

1. DoRANum
   (a) DoRANum est une plateforme de formation en ligne sur la gestion et le partage des données de la recherche selon les principes FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable), réalisée par l'Inist-CNRS et le GIS « Réseau Urfist » depuis 2015.
   (b) 130 ressources pédagogiques numériques réparties dans plusieurs thématiques générales et disciplinaires, qui permettent aux chercheurs et doctorants de se former selon leurs besoins et selon leurs niveaux de connaissance.
   (c) La plupart des ressources pédagogiques sont librement réutilisables et adaptables.
   (d) https://doranum.fr/
2. OPIDoR (Optimiser le Partage et l'Interopérabilité des Données de la Recherche)
   (a) Portail mis à la disposition de la communauté
   (b) Petit-b
   (c) Petit-c
   (d) https://opidor.fr/
3. FUN Mooc (France Université Numérique)
   (a) Plateforme proposant une grande variété de MOOC gratuits
   (b) Propose des modules pour se former à la science ouverte
   (c) https://www.fun-mooc.fr/fr/
4. CoopIST
   (a) Site sectoriel de la Délégation à l'information scientifique et à la science ouverte du CIRAD (Organisme français de recherche agronomique)
   (b) Propose un ensemble de fiches synthétiques sur la gestion de données et sur différents aspect de la recherche.
   (c) https://coop-ist.cirad.fr/lettre-coopist

Liens http://...
http://...
http://...

Mots clé Science Ouverte,données de la recherche

Se former au numérique

1. FUN Mooc (France Université Numérique)
   (a) Plateforme proposant une grande variété de MOOC gratuits
   (b) Nombreux mooc gratuits sur l'initiation à l'informatique (python, shell bash ...) et sur les techniques de traitement de données textuelles (manipulation, machine learning ...)
   (c) https://www.fun-mooc.fr/fr/

2. Mate-SHS
   (a) Réseau de professionnels de la recherche traitement des données appliquées au SHS
   (b) Propose les Tuto@Mate, séminaires de méthodes librement visionnables sur Youtube qui présentent différents outils numériques appliqués aux SHS.

3.1.2 Se former en ateliers

1. Formations bibliothèque CYU
   (a) Formations à la recherche documentatire, services d'appui à la recherche
   (b) Possibilité de prendre rdv pour ces formations sur demande par mail à bu-formation@ml.u-cergy.fr
2. Séminaires de l'école doctorale
   (a) Demander le programme aux doctorants

Liens https://bibliotheque.cyu.fr/version-francaise/se-former/se-former
http://...
http://...

Mots clé

3.2 Bases de corpus

Introduction

Liens https://books.openedition.org/septentrion/119418?lang=fr
https://wiki.frantext.fr/bin/view/Main/Manuel%20d%27utilisation/Corpus/
https://www.sketchengine.eu/comment-creer-un-corpus-a-partir-dinternet/
https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop/

Mots clé

ORTOLANG/NAKALA et les corpus dispo dessus > Exemple via des corpus, aussi parler d'autres gros corpus plus 'industriels'

3.2.1 Corpus de presse

1. EUROPRESSE
   (a) Europresse est une base de presse et actualités comportant plus de 8000 sources d'information reconnues : presse régionale, nationale et internationale, ressources généraliste et spécialisée, sites Web, télévision et radio, biographies, etc.
   (b) Elle vous permet d'interroger et consulter en texte intégral des articles de publications couvrant diverses thématiques, et parfois même directement les versions PDF des journaux et revues
   (c) Cette base est accessible gratuitement et intégralement à l'ensemble des étudiants de CYU et de nombreux autres établissements
   (d) Accès par CYU https://cyu.libguides.com/az.php
2. FACTIVA
   (a) Factiva est un outil d'information professionnelle de la société Dow Jones & Company. Factiva agrège des contenus provenant à la fois de sources sous licence et gratuites, et apporte aux entreprises des fonctionnalités de recherche, d'alerte, de diffusion et de gestion de l'information.
   (b) Accès par CYU https://cyu.libguides.com/az.php

Droits d'usage

- Articles de presse = soumis aux Licence Creative Commons (Attribution - Pas d'Utilisation Commerciale - Pas de modifications) pour la grande majorité.
- Cette licence permet le droit à la citation courte4.2.2
- Chaque revue de presse dispose de sections fournies qui détaillent les conditions d'utilisation de leur contenu
- Elles proposent également des formulaires de contact des auteurs, si un transfert de droit est nécessaire au travail en question (traduction par exemple)
- Pour plus d'informations sur les licences4.2.1

3.2.2 Corpus d'oeuvres littéraires

1. Projet Gutenberg
   (a) Le projet Gutenberg est une bibliothèque de versions électroniques libres
   (b) Les textes fournis sont essentiellement du domaine public soit parce qu'ils n'ont jamais été sujets à des droits d'auteur, soit parce que ces derniers sont expirés. Il contient toutefois quelques textes toujours sous droit d'auteur, qui sont rendus disponibles pour le projet avec la permission de l'auteur.
   (c) Accès : https://gutenberg.org/

Droits d'usage

- Le principe même du projet est de réunir un corpus libre de droit, et tout usage (même commercial) des corpus est autorisé.
- Pas de restriction quant à la modification des données (plutôt encouragé même pour normaliser des données déstinées à des traitements.)

— Lorsque des projets de recherche sur le corpus du projet Gutenberg, il est recommandé de le citer, mais ce n'est même pas obligatoire.
— Pour plus d'informations sur les licences4.2.1

2. Frantext
   (a) La base Frantext est conçue pour permettre des recherches de mots, lemmes et expressions régulières dans un corpus donné.
   (b) Frantext est une base de données comportant 5658 références, soit 270 millions de mots en Décembre 2023. Développée à l'ATILF (Analyse et Traitement Informatique de la Langue Française), elle est disponible en ligne depuis 1998.
   (c) Elle permet de faire des recherches simples et complexes sur des formes, des lemmes ou des catégories grammaticales et d'afficher les résultats dans un contexte de 700 signes
   (d) Les versions numériques des textes libres de droits sont téléchargeables
   (e) Accès par CYU https://cyu.libguides.com/az.php

   Droits d'usage
   — La base FRANTEXT départage clairement ses oeuvres libres de droits, ses oeuvres soumises à des droits.
   — Pour les oeuvres libres de droit
   — Pour les oeuvres soumises à des droits d'auteurs, le droit à la citation courte est toujours légitime4.2.2.
   — Pour plus d'informations sur les licences4.2.1

3. Google Books et Ngram viewer
   (a) Google Books = Banque d'oeuvres, certaines libres (domaine public) et entièrement consultables, d'autres soumis à des droits d'auteurs.
   (b) Quand droit d'auteurs » Extraits disponibles, avec l'accord à Google de l'auteur ou éditeur
   (c) Ngram Viewer est une application linguistique proposée par Google, permettant d'observer l'évolution de la fréquence d'un ou de plusieurs mots ou groupes de mots à travers le temps dans les sources imprimées. L'outil est entré en service en 2010. La dernière mise à jour de ce moteur de recherche web date de février 2021.
   (d) Le terme « ngram » désigne dans ce contexte une suite de « n » mots, ce qui est un cas particulier de la notion de n-gramme.

   Droits d'usage
   — Google Books : Pour les oeuvres libres » idem que pour Gutenberg
   — Pour les oeuvres soumis à des droits d'auteurs » droit citation courte 4.2.2
   — Pour créer des corpus »> oeuvres libres seulement.
   — Utilisable librement par navigateur, possibilité de reprendre les graphes en citant.

3.2.3 Corpus moissonnés sur le web

1. Common Corpus
   (a) Le plus gros corpus de textes de textes libres de droits de 500M de mots
   (b) Crée par la start-up Pleias, soutenue par les acteurs de la science ouverte
   (c) But : pouvoir entrainer des modèles de langages sur des données libres et transparentes.
   (d) Librement réutilisable pour la recherche, pour constituer des corpus.
2. FRWaC
   (a) Corpus du français de 1.6M de tokens
   (b) Méthodologie étique, transparente, reproductible, mais des lacunes (soulignées par les auteurs) pour statuer son echantillonage et pour retracer précisément l'origine de tout.
3. Paracrawl
   (a) Enorme corpus de segments tirées d'internet, monotexte et bitextes (EN>[41 langues])
   (b) initiative européenne, soucieuse du respect de l'éthique et de la loi
   (c) Licence : Creative Commons CC0 license ("no rights reserved")2.

Droits d'usage

- Pour certains corpus, respect des droits d'auteurs flou
- Utilisés surtout pour le ML, utiles si vous voulez entrainer vos propre modèles "de zéro" (de nombreux modèles pré-entrainés sont disponible sur le web)
- N'utiliser que si les droits sont clairement explicités (comme Paracrawl, Common Corpus)
- Le droit à la citation courte est toujours de rigueur 4.2.2
- Pour plus d'informations sur les licences4.2.1

Liens https://www.sketchengine.eu/frwac-french-corpus/#toggle-id-1
https://commoncrawl.org/research-papers
https://repository.ortolang.fr/api/content/cefc-orfeo/4/documentation/site-orfeo/corpus-source/index.html
https://paracrawl.eu/
https://commoncrawl.org/
Base textuelle FRANTEXT, ATILF - CNRS& Université de Lorraine. Site internet : http://www.frantext.fr. Version décembre 2016.
(Wissner 2012)

Mots clé corpus,licence de diffusion,données de la recherche

3.2.4 Banques grammaticales (arbres)

1. UD
   (a) Universal Dependencies (UD) est un projet qui consiste en une banques d'arbres de dépendances cohérentes d'un point de vue interlinguistique pour de nombreuses langues, dans le but de faciliter le développement d'analyseurs multilingues, l'apprentissage interlinguistique et la recherche sur l'analyse syntaxique du point de vue de la typologie des langues. Le schéma d'annotation est basé sur une évolution des dépendances Stanford (universelles), des étiquettes universelles de partie de discours de Google, et de l'interlingua Interset pour les étiquettes morphosyntaxiques. La philosophie générale est de fournir un inventaire universel de catégories et de lignes directrices pour faciliter l'annotation cohérente de constructions similaires à travers les langues, tout en permettant des extensions spécifiques à la langue si nécessaire.
   (b) Droits d'usage : Libre (open source)

3.3 Gestionnaires de corpus

1. Sketch Engine
   (a) Sketch Engine est un gestionnaire de corpus et un outil d'analyse textuelle développé par Lexical Computing Limited. Son objectif est de permettre aux personnes qui étudient les langues (lexicographes, chercheurs en linguistique de corpus, traducteurs ou apprenants en langues) ou quiconque qui souhaite trouver des exemples authentiques de rechercher de grandes collections de textes selon des requêtes complexes.
   (b) Droits d'usage : Contient de nombreux corpus libres de droit / Logiciel propriétaire (non open source)
   (c) Source : https://bu.univ-lyon3.fr/sketch-engine-un-outil-pour-ceux-qui-etudient-les-langues
2. CQPweb
   (a) Petit-a
   (b) Petit-b
   (c) Petit-c
3. Grand-3
   (a) Petit-a
   (b) Petit-b
   (c) Petit-c

Liens http://...
http://...
http://...

Mots clé

3.4 Lexiques

1. Morphalou
   (a) Morphalou3 est un lexique à large couverture. Les lexies sont accessibles par leurs formes lemmatiques (forme canonique non fléchie). À chacun de ces lemmes sont associées toutes ses formes fléchies (déclinaisons et conjugaisons du lemme).
   (b) Droits d'usage : licence LGPL-LR (Lesser General Public License For Linguistic Resources)
2. Leff
   (a)

3.5 Textométrie

Définition de la textometrie : La textométrie est l'application de calculs sur des données textuelles : statistique lexicale, analyses factorielles, classifications.https://www.encyclopedie.fr/definition/Textom%C3%A9trie

1. Iramuteq
   (a) Iramuteq est une Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires, son fonctionnement consiste à préparer les données et écrire des scripts qui sont ensuite analysés dans le logiciel statistique R. Les résultats sont finalement affichés par l'interface.
   (b) Demande une installation de R et de python (facile à faire, possible à l'interface graphique)
   (c) Droits d'usage : Licence GNU GPL (licence libre)
2. Le Trameur et iTrameur
   (a) Le Trameur est un programme de génération puis de gestion de la Trame et du Cadre d'un texte (i.e découpage en unité et partitionnement du texte : le métier textométrique) pour construire des opérations lexicométriques / textométriques (ventilation des unités, carte des sections, cooccurrence, spécificité, AFC...).
   (b) Le Trameur est compatible avec l'outil treetagger : système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation. Il permet aussi de génèner et de gérer des annotations multiples sur les unités du texte (et de traiter les niveaux d'annotations visés)
   (c) iTrameur – Outils d'analyse textométrique de données est un ensemble d'outils en ligne comportant plusieurs fonctionnalités de l'analyse automatique de textes en vue de leur profilage sémantique, thématique et de leur interprétation. C'est une version en ligne du Trameur. iTrameur est à l'origine un outil de textométrie. Il dispose aussi des fonctionnalités particulières qui permettent d'annoter dynamiquement des corpus ou d'explorer des ressources annotées (treebanks monolingues/multilingues) ou des alignements.
   (d) Sur le site de l'outil, seules les versions allégées sont directement téléchargeables. Ces version offrent exactement les mêmes fonctionnalités que les version complètes, à la différence qu'elles n'intègrent pas le module *treetagger* (outil d'étiquetage morpho-syntaxique). *treetagger* est un outil libre et gratuit, mais ne peut pour des raisons de droits être directement intégrées à l'outil. Pour installer *treetagger* sur l'interface, veuillez vous référer à la documentation de l'auteur fournie avec un tutoriel d'installation (accessible et réalisable à l'interface graphique).
   (e) Droit d'usage : Ouvert à la communauté scientifique
3. TXM
   (a) TXM est un logiciel de textométrie open-source et gratuit utilisé dans le traitement automatique du langage naturel, l'analyse de données textuelles, l'analyse du discours, l'analyse de contenu, la logométrie, la littérométrie, ou autres fouilles de textes effectuées en linguistique, mais aussi et de plus en plus, en sciences humaines et sociales (par exemple en sociologie1 et en géographie) et dans les autres disciplines connexes que regroupe le champ des humanités numériques.
   (b) Besoin de procéder à l'installation de treetagger manuellement
   (c) Droits d'usage : Libre (open source)
4. 1 ou 2 outils de textométrie en plus

Liens https://bu.univ-lyon3.fr/sketch-engine-un-outil-pour-ceux-qui-etudient-les-langues
https://universaldependencies.org/introduction.html
https://repository.ortolang.fr/api/content/morphalou/2/LISEZ_MOI.html#idp37270384
http://www.tal.univ-paris3.fr/trameur/iTrameur/ (Pincemin 2020)

Mots clé

3.6 HN et Linguistique outillée (?)

Quelques taches et outils de TAL pour le linguiste

1. Segmentation (Tokenization)
   (a) Processus de segmentation du textes en unités appellées "tokens" (élément de base d'annotation)
   (b) Ces unités peuvent prendre la forme de "mots", de "sous-mots", de phrases ...
   (c) Cette opération est la première étape de tout traitement de données textuelles en TAL
2. Reconnaissance d'entités nommées (NER)
   (a) Tache d' identification et de classification automatique d'entités nommées (noms propres, dates, toponymes...) sur un ensemble de "token"
3. Étiquetage en parties du discours (POS Tagging)
   (a) Étiquetage en partie du discours des "token" qui composent une phrase, pour identifier les catégories gramaticales et les relations entre les "token" d'une phrase.
4. Catalogues d'outils

PostLab catalogue de logiciels et applications académiques d'intelligence artificielle »> https://www.postlab.fr/

TAPoR 3.0 (Text Analysis Portal for Research) : catalogue d'outils d'analyse et de manipulation de données textuelles pour la recherche. Les utilisateurs proposent des combinaisons d'outils efficaces à certaines approches. Référence toutes les licences attribuées à ces outils. »> https://tapor.ca/home

taire des outils de CORLI Inventaire réunissant une grande variété d'outils de traitement de corpus de langage (écrit/oraux) https://corli.huma-num.fr/inventaire-des-outils/

s du consortium ARIANE : Chaînes éditoriales : Gestion de format, génération, métadonnées, édition critique, script de validation de format https://axe-1-gt3-outils-et-pratiques-editoriales.gitpages.huma-num.fr/scripts/recensement

5. Récupérer des données textuelles issues du web

Webscrapping* et *webcrawling

— Webcrawling (moissonage) : Consiste à récupérer en masse des données depuis des pages internet, par un système de sauts de liens en liens
— Webscraping (grattage) : outil spécifiquement conçu pour l'extraction des données d'un site web ciblé

Méthodes

— 2 approches possible : approche "code" (bs, selenium, wget ...) et approche interface
— Crawl par extensions de navigateur : Instant Data Scraper, web scraper.
— Avantages de l'approche code : permet un examen plus fin et sur un grand nombre de pages, mais plus difficile d'accès
— Avantages de l'approche plug in : très facile à utiliser, mais demande un examen "page par page", moins pratique pour récupérer des grandes quantités de texte.

Sur le plan légal

— Le scraping est légal si les données récoltées sont publiques
— Il est encadré par le RGPD, attention aux données personnelles
— Il faut se référer aux Conditions générales d'utilisation du site que l'on veut examiner pour savoir dans un premier temps si il est possible de le faire, et dans un second temps dans quelles conditions.
— Les sites se prémunissent de plus en plus du scraping en l'empéchant ou en l'encadrant
— Certains sites proposent des interfaces API permettant de collecter des données en toute conformité à leurs CGU et en toute légalité. (ex :Twitter ...)
— Il est tout à fait possible de demander l'autorisation à/aux auteur(s) d'utiliser leurs données dans le cadre de travaux de recherche.

Liens https://www.geeksforgeeks.org/nlp-libraries-in-python/https://www.postlab.fr/
(Webster et Kit 1992)
(Straka et Straková 2017)
(Sun et al. 2018)

Mots clé TAL,Tokenisation,POS Tagging,NER,token

Méthodologie de veille en TAL

1. Un domaine en ébullition
   (a) Linguistique outillée = de plus en plus nécessaire car plus de données
   (b) Omniprésence des modèles de langage dans la presse/réseaux/revues
   (c) Nombreuses application possibles qui peuvent faciliter la vie du chercheur et lui donner de nouveaux axes de recherche sur ses trauvaux.
   (d) Les avancées sont rapides > intérêt certain à la veille sur le domaine
2. Sources d'informations

Revues

— **La revue TAL** : Publiée par l'Association pour le Traitement Automatique des Langues (ATALA), cette revue paraît 3 fois par an sous format éléctronique réunit papiers, thèses et états de l'art sur divers problématiques de l'industrie de la langue en jetant une passerelle entre la linguistique et l'informatique.

— Les contenus qui y sont publiés sont soumis à la licence Creative Commons Attribution 4.0 International License.

3. Conférences

Quelles activités au conférences ?

— Shared task : groupe de travail qui se réunit pour réaliser un projet dans un intervalle déterminé, autour d'un sujet défini en amont.

— Workshop/ Ateliers : un atelier collaboratif ayant pour objectif d'échanger sur une thématique précise. À la différence d'une réunion, tous les participants d'un workshop interagissent pour construire une réflexion, trouver une idée, partager un savoir particulier ou débattre sur la problématique définie

— Demo : présentation d'une solution à un problème par un outil présenté par une entreprise (état fini ou non)

La conférence JEP-TALN RECITAL

— évènement annuel organisée par l'ATALA réunissant les Journées d'études sur la parole (JEP), axées sur la recherche en phonétique, phonétique et phonologie, la conférence Traitement Automatique du Langage Naturel (TALN), consacrée aux avancées en traitement automatique des langues, et les Rencontres des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) dédiée aux jeunes chercheurs pour présenter leurs travaux de recherche à la communauté.

— Elle est l'occasion de réunir chercheurs et industriels, juniors et expérimentés autour des problématiques actuelles du TAL.

— Liens vers les actes publiés chaque année : https://www.atala.org/index.php/-Conference-TALN-RECITAL

La conférence LREC (International Conference on Language Resources and Evaluation)

— Conférence bianuelle organisée par l'ELRA sur des thématiques transverses informatique/linguistique.

— organisé par l'ELRA

— Evenement majeur dans le champs de la linguistique computationelle

La conférence ACL

— Edition américaine + européenne chaque année
— axée linguistique computationelle

4. Veille automatique
   (a) Google Scholar
   (b) Collection HAL de l'ATALA : https://hal.science/TALN-RECITAL/browse/period

Liens http://...
http://...
http://...

Mots clé

Partage et valorisation des données

4.1 Entrepôts

4.2 Licences et droits

4.2.1 Licences

1. La licence CC-by 4.0 (Creative Commons Attribution)
   (a) Permet de partager, copier, distribuer et communiquer les données par tous moyens et sous tous formats, de les réutiliser pour créer de nouveaux jeux de données. Toutes les utilisations, y compris commerciales, sont possibles, sous réserve de créditer les données à leurs créateurs (obligation d'attribution).
   (b) Cette licence est préconisée par un certain nombre d'entrepôts de données.
   (c) Petit-c
2. La licence CC0 (Creative Commons Public Domain Dedication)
   (a) permet aux producteurs de données de les placer dans le domaine public, sans aucune restriction de réutilisation. La citation du producteur du jeu de données n'est pas obligatoire, même si, d'un point de vue éthique et scientifique, il est conseillé aux utilisateurs de citer les créateurs originels des données lors de la réutilisation.
   (b) Petit-b
   (c) Petit-c
3. La Licence ouverte (LO)
   (a) Petit-a
   (b) Petit-b
   (c) Petit-c

4.2.2 Droits

Droits d'auteurs

— Le droit d'auteur est l'ensemble des droits dont dispose un auteur ou ses ayants droit (héritiers, sociétés de production), sur ses œuvres originales définissant notamment l'utilisation et à la réutilisation de ses œuvres sous certaines conditions.

Le droit à la citation courte

— Le droit à la citation courte est une exception au Code de la propriété intellectuelle, permettant de citer des extraits de textes soumis à des droits d'auteur, "sous réserve que soient indiqués clairement le nom de l'auteur et la source " (Article L122-5 du Code de la propriété intellectuelle)

— Le Code de la propriété intellectuelle admet *"les analyses et courtes citations justifiées par le caractère critique, polémique, pédagogique, scientifique ou d'information de l'oeuvre à laquelle elles sont incorporées"* (Article L122-5 du Code de la propriété intellectuelle)

— La citation ne doit pas dénaturer le propos de l'auteur, et doit être dans le texte clairement identifiable par l'usage de guillemets ou d'une police différente du corps de texte.

— La loi ne prévoit pas de limite quantifiable en mots ou en caractères pour les, et il est de la responsabilité de l'auteur d'en faire un usage raisonné.

— C'est la proportion de citations dans l'oeuvre qui est étudiée en cas de suspicion de plagiat (30 citations courtes peuvent être jugées excessives dans un court article, mais tout à fait acceptable dans une thèse).

Liens https://coop-ist.cirad.fr/gerer-des-donnees/rendre-publics-ses-jeux-de-donnees/6-les-principales-l
http://...
http://...

Mots clé

4.3 Métadonnées

1. **Définition**
   (a) C'est une donnée fournissant de l'information sur une autre donnée
   (b) Elle permettent de situer les données en question dans leur contexte (qui, quoi, ou, quand, comment)
   (c) En recherche, il existe des standards de métadonnées, propres à la nature des données et à la pratique de la discipline
   (d) **XML** : langage de balises, référence pour ce qui est des métadonnées.
2. **Standards de métadonnées**
   (a) Le standard a pour objectif de fournir un ensemble d'éléments caractéristiques qui permettent de décrire les productions scientifiques. Ainsi la recherche peut être facilitée en portant sur les critères définis. La description des éléments peut elle-même être précisée par l'emploi de vocabulaires dédiés. Le standard est choisi en fonction de la destination des données, dépôt, publication, archivage, etc. Il peut aussi être spécialisé par discipline, par type de données, etc., ainsi que son vocabulaire.
   (b) Quelques standards de métadonnées (toutes en XML)

Dublin Core
- — Utilisé par tous les gestionnaires de bibliothèques numériques et les plateformes généralistes de dépôt et de publication des données
- — Description générale des documents, précisée par de nouveau éléments apportés par le Dublin Core étendu (permet un usage plus adapté aux disciplines spécifiques)

teiHeader
- — Le but de la TEI est de " fournir des recommandations pour la création et la gestion sous forme numérique de tout type de données créées et utilisées par les chercheurs en Sciences humaines et sociales " (Lou Bunard, un des fondateur de la TEI)
- — Priorité au sens du texte plutôt qu'à son apparence
- — Indépendant de tout environnement logiciel
- — Conçu par la communauté scientifique qui est aussi en charge de son développement continu

MARC
- — MARC = MAchine-Readable Cataloging
- — Il présente de nombreuses variantes (UNIMARC, INTERMARC, USMARC ...)
- —

EAD
- — Format XML (DTD et schéma) pour l'encodage des descriptions de fonds d'archives
- — Également utilisé en bibliothèque pour la description de manuscrits.

DDI
- — Standard de documentation technique pour décrire et conserver les informations statistiques et plus globalement les informations et données d'enquêtes en sciences humaines et sociales

Liens https://doranum.fr/wp-content/uploads/Fiche-Synth%C3%A9tique-M%C3%A9tadonn%C3%A9es.pdf
https://axe-1-gt3-outils-et-pratiques-editoriales.gitpages.huma-num.fr/scripts/recensement
http://...

Mots clé métadonnées,langages de balisage,teiHeader

4.4 Identifiant Unique Pérenne

4.5 Data Papers

4.6 Canaux de communication

4.6.1 Listes de diffusion

parislinguists Une liste de diffusion pour les linguistes français substituant la liste de Yahoo, où les membres de la liste pourraient échanger des infos sur les conférences, présentations, colloques et postes dans leur domaine.
Inscription :https://listes.services.cnrs.fr/wws/info/parislinguists

DH Liste francophone de discussion autour des Digital Humanities (DH)
Inscription :https://groupes.renater.fr/sympa/info/dh

CORLI CORLI est un consortium d'Huma-Num Information détaillées disponibles ici : https ://corli.huma-num.fr/les-groupes-projets/gp5/ Le groupe de travail "annotation" a pour mission de réfléchir aux bonnes pratiques pour la gestion d'une campagne d'annotation, depuis la conception de la campagne jusqu'à la mesure de l'accord inter-annotateur, en passant par la prise en main d'outils d'annotation.
Inscription :https://groupes.renater.fr/sympa/info/annotation-corli

ln Liste de diffusion des annonces à destination des chercheurs en Traitement Automatique des Langues

quanti La liste de discussion "Quanti", créée après la journée d'études "Enseigner le quanti" qui a eu lieu à Paris le 5 juin 2015, a pour vocation d'accueillir les contributions et les échanges de toutes celles et tous ceux qui s'intéressent aux questions d'enseignement des méthodes quantitatives dans les sciences sociales.
Inscricription :https://groupes.renater.fr/sympa/info/quanti

linguistlist La linguist list dépend du département de linguistique de l' Indiana University, et vise à fournir un forum d'échange de problématiques et d'information sur la recherche en linguistique. (en anglais).
Inscription : https://linguistlist.org/

4.6.2 Consortiums HN

1. En quoi consistent-ils ?
   (a) Les Consortiums-HN réunissent plusieurs personnels d'unités et équipes de recherche françaises, aux métiers variés (chercheur.e.s, ingénieur.e.s, archivistes, documentalistes, etc.) autour de thématiques et/ou d'objets communs pour lesquels ils définissent des procédures et standards numériques partagés (méthodes, outils, partages d'expériences).
   (b) L'existence des consortium sont limitées dans le temps
   (c) L'affiliation à un consortium induit la participation active aux projets menés par le groupe de recherche
   (d) Il est recommandé cependant de suivre l'activité des consortiums pour se maintenir informé des avancées (à la manière d'une liste de diffusion
2. Consortium de l'IR HUMA-NUM axés linguistiques

Ariane (Analyses, Recherches, Intelligence Artificielle et Nouvelles Editions numériques)

- ARIANE réunit des spécialistes du texte (littéraires, linguistes, historiens...) et de l'informatique en vue de créer un espace de dialogue véritablement interdisciplinaire entre ces deux communautés.
- L'objectif du consortium ARIANE est de progresser dans la connaissance et le raffinement des méthodes informatiques appliquées aux objets et données des sciences humaines et plus particulièrement, des sciences du texte.
- Le groupe de travail sur les métadonnées : reflexions sur les définitions, identifier les besoins chercheurs.
-

CORLI 2 (Corpus, Langues et Interactions 2)

- CORLI est un réseau de laboratoires et de chercheurs travaillant sur les corpus de langage. Son but est d'offrir à tous des données, des outils, de la documentation et des formations autour de l'utilisation scientifique des corpus de langage en suivant les principes FAIR.

— Ressources utiles sur les outils, formats, bonnes pratiques et aspect juridiques

3. Renvoi vers le catalogue des consortium HUMA-NUM 4.6.2
4. Consortiums européens

CLARIN (Common Language Resources and Technology Infrastructure)

— Consortium européen pour le partage de ressources et d'outils du langage au niveau européen
— Propose des ressources pour la recherche en SH, principalement axées linguistique et traitement automatique du langage
— Différence centre C et centres K :
  — Centre B : Les centres techniques, ils sont fournisseurs de services, rattachés à des université ou des instituts académiques. Les centre C sont aussi des centre techniques, dont les métadonnées sont intégrées dans CLARIN, mais autogérées
  — Centre K :
— En france : CORLI= centre K, plus récemment, ORTOLANG= centre C.
—

DARIAH (Digital Research Infrastructure for the Arts and Humanities)

— Initiative européenne pour developpement et soutien de la recherche dans différentes disciplines des SH numériques (texte, son, image, vidéo).
— Différents centres réunis en catégories de lettres, les centres K (Knowledge Centres) se focalisent sur les ressources linguistiques dans toutes les formes qu'elle peuvent prendre dans les différents domaines de la recherche (
— Réunit 22 pays européens et 197 institutions partenaires sur des projets de recherches variés
— Elle s'inscrit par définition dans la logique science ouverte (libre accès des ressources, certification des entrepots de données, procédure d'archivage, mise à disposition en open source, formats standardisés, interopérabilité)

Liens https://www.huma-num.fr/les-consortiums-hn/
https://cst-ariane.huma-num.fr/
https://corli.huma-num.fr/
https://ptm.huma-num.fr/
https://www.dariah.eu/
https://clarin-fr.fr/

Mots clé

4.7 Vulgarisation

Référencement sur des répertoires d'experts

1. Youmanity : répertoire d'experts pour les médias https://www.youmanity.org/expertalia-repertoire-dexpert-e-
2. Expertes.com : annuaire de femmes expertes francophones https://expertes.fr/
3. Article de Julien Longhi à citer ? Celui là ? (Longhi 2022)

Podcast

(a) Vox (EFL)
   i. Podcast lancé en 2021 à l'initiative du Labex EFL
   ii. Divers podcast qui traitent de sujets de linguistique, et qui met en avant les chercheurs du labex et leurs travaux
   iii. Etablit un pont entre les chercheurs et le gand public
   iv. Moyen de valorisation des travaux de la recherche par excellence

(b) Parler comme jamais
   i. Podcast animé par Laélia Véron, mélant linguistiques et sujets de société
   ii. Petit-b
   iii. Petit-c

(c) Les podcast linguistiques Radio France
   i. Podcasts sur des sujets linguistique/humanités, linguistique/société

Liens https://www.radiofrance.fr/sujets/langues-linguistique
https://www.labex-efl.fr/post/vox-le-podcast-du-labex-efl-qui-vous-parle-de-linguistique
https://www.binge.audio/podcast/parler-comme-jamais

Mots clé archives

Glossaire

anonymisation L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible. . 16

archives Une organisation dont la mission est la conservation des informations afin d'assurer l'accès et l'utilisation par une communauté spécifique, ou un site où des données lisibles par machine sont stockées, conservées et éventuellement redistribuées aux personnes intéressées à utiliser lesdites données. Source : https://bu.univ-lille.fr/chercheurs-doctorants/science-ouverte/donnees-de-recherche/sauvegarde-et-stockage-des-donnees . 8, 39

bibliométrie La bibliométrie est l'application de méthodes statistiques et mathématiques pour mesurer et évaluer la production et la diffusion de publications. Elle génère des formules, parfois sophistiquées, visant à donner un indice de performance de la recherche pour un chercheur ou une chercheuse, un laboratoire, un établissement, un pays, etc. source:https://bu.univ-larochelle.fr/lappui-a-la-recherche/valorisation-de-la-recherche/bibliometrie-et-impact-de-la-recherche-2/ . 9

Conditions générales d'utilisation Les CGU sont un ensemble de clauses contractuelles qui régissent les relations entre un fournisseur de services en ligne (site web, application, plateforme, etc.) et ses utilisateurs. Elles ont pour objectif de définir les droits et obligations des parties, d'encadrer l'utilisation du service et de protéger les intérêts du fournisseur. . 29

corpus Ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique . 25

criticité La détermination et la hiérarchisation du degré d'importance et de la disponibilité d'un système d'information. . 8

DOI Le digital object identifier (DOI, littéralement « identifiant numérique d'objet») est un mécanisme d'identification de ressources stable, qui peuvent être des ressources numériques, comme un film, un rapport, des articles scientifiques, ainsi que des personnes ou tout autre type d'objet.
source : https://fr.wikipedia.org/wiki/Digital_Object_Identifier . 9, 15

données de la recherche Les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. (recherche.data.gouv) . 5, 7, 8, 9, 10, 14, 15, 21, 25

Entrepôt de données de la recherche Un entrepôt de données de recherche (Research Data Repository ou Data Repository) est une base de données destinée à accueillir, conserver, rendre visibles et accessibles des données de recherche. Son rôle est de permettre le dépôt ou la collecte de données, leur description, leur accès, et leur partage en vue de leur réutilisation. Chaque entrepôt dispose généralement d'une politique de dépôt, de description et de diffusion des données. . 8

FAIR Les principes FAIR (Findable, Accessible, Interoperable, Reusable) décrivent comment les données doivent être organisées pour être plus facilement accessibles, comprises, échangeables et réutilisables . 14

h-index Le h-index (ou facteur h), créé par le physicien Jorge Hirsch en 2005, est un indicateur d'impact des publications d'un chercheur. Il prend en compte le nombre de publications d'un chercheur et le nombre de leurs citations. Le h-index d'un auteur est égal au nombre h le plus élevé de ses publications qui ont reçu au moins h citations chacune. . 9

langages de balisage En informatique, les langages de balisage représentent une classe de langages spécialisés dans l'enrichissement d'information textuelle. Ils utilisent des balises, unités syntaxiques délimitant une séquence de caractères ou marquant une position précise à l'intérieur d'un flux de caractères. . 18, 35

licence de diffusion Une licence de diffusion est un instrument juridique, complémentaire au droit d'auteur. Elle permet au titulaire des droits sur une œuvre d'accorder à l'avance aux utilisateurs certains droits d'utilisation de cette œuvre. Elle préserve les droits moraux de l'auteur en imposant toujours l'obligation d'attribution (citation de la source). Source : https://coop-ist.cirad.fr/etre-auteur/utiliser-les-licences-creative-commons/2-qu-est-ce-qu-une-licence-de-diffusion . 10, 14, 25

LMF Lexical Markup Framework (LMF ou cadre de balisage lexical, en français) est le standard de l'Organisation internationale de normalisation (plus spécifiquement au sein de l'ISO/TC37) pour les lexiques du traitement automatique des langues (TAL). . 18

métadonnées Les métadonnées sont utiles pour exploiter un jeu de données produit par d'autres. Ce sont les nombreuses informations relatives au contexte de production, à la méthodologie, à la description du jeu, etc. Les entrepôts de données sont des espaces qui rendent accessibles les jeux de données et les métadonnées qui y sont associées. Les entrepôts en auto-dépôt sont des espaces de partage libre de données et sans vérification de la qualité des métadonnées. (CNRS) . 14, 18, 35

NER La reconnaissance d'entités nommées est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mots) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc. . 30

PGD Document dont la rédaction doit être initiée au commencement d'un projet de recherche, décrivant les données et comment elles seront partagées et conservées pendant et après le projet . 11, 15, 17

POS Tagging En linguistique, l'étiquetage morpho-syntaxique (aussi appelé étiquetage grammatical, POS tagging (part-of-speech tagging) en anglais) est le processus qui consiste à associer aux mots d'un texte les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre, etc. à l'aide d'un outil informatique . 30

pseudonymisation La pseudonymisation est un procédé de remplacement des données permettant d'identifier une personne physique par d'autres données. Cette technique permet de protéger les données personnelles de l'individu concerné. Contrairement à l'anonymisation, la pseudonymisation est un procédé réversible permettant de retrouver la trace des données remplacées. . 16

RGPD Le règlement général de protection des données (RGPD) est un texte réglementaire européen qui encadre le traitement des données de manière égalitaire sur tout le territoire de l'Union européenne (UE). Il est entré en application le 25 mai 2018. Le RGPD s'inscrit dans la continuité de la loi française « Informatique et Libertés » de 1978, modifiée par la loi du 20 juin 2018 relative à la protection des données personnelles, établissant des règles sur la collecte et l'utilisation des données sur le territoire français. . 15, 16

sauvegarde Une copie de tout ou une partie des fichiers sur un système séparé des données originelles, à des fins de récupération sur le court terme en cas de perte ou de dégradation des données. Il s'agit d'une image figée dans le temps des fichiers ; la fréquence des sauvegardes et le nombre de versions conservées simultanément dépendent des outils, services et besoins. Source : https://bu.univ-lille.fr/chercheurs-doctorants/science-ouverte/donnees-de-recherche/sauvegarde-et-stockag . 8

Science Ouverte La science ouverte est la diffusion sans entrave des publications et des données de la recherche. Elle s'appuie sur l'opportunité que représente la mutation numérique pour développer l'accès ouvert aux publications et – autant que possible – aux données de la recherche. . 5, 7, 11, 14, 21

TAL Le traitement automatique des langues, en anglais natural language processing ou NLP, est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer des outils de traitement du langage naturel pour diverses applications. . 30

TEI La Text Encoding Initiative (abrégé en TEI, en français « initiative pour l'encodage du texte ») est un format de balisage et une communauté académique internationale dans le champ des humanités numériques visant à définir des recommandations pour l'encodage de ressources numériques, et plus particulièrement de documents textuels. . 18

teiHeader L'en-tête TEI (teiHeader) fournit des informations descriptives et déclaratives qui constituent une page de titre électronique au début de tout texte conforme à la TEI. . 18, 35

TMX Translation Memory eXchange est une spécification XML pour l'échange de données de mémoire de traduction entre des outils de traduction et de localisation assistés par ordinateur avec peu ou pas de perte de données critiques. . 18

token Un token est une séquence de caractères dans un document particulier qui sont regroupés en tant qu'unités sémantique utile pour le traitement. . 30

Tokenisation La tokenisation est une étape fondamentale du traitement du langage naturel (NLP). Elle consiste à découper un texte en unités plus petites, appelées tokens, qui peuvent ensuite être traitées par des modèles d'apprentissage automatique. . 30

XML L'Extensible Markup Language, généralement appelé XML, « langage de balisage extensible » en français, est un métalangage informatique de balisage générique qui est un sous-ensemble du Standard Generalized Markup Language. . 18

Bibliographie

Bouchet Moneret, Florence (mars 2021). *Les données personnelles de recherche et le RGPD*. Guide sur les données personnelles de recherche dans le cadre du RGPD. URL : https://hal.univ-lorraine.fr/hal-03636697.

Delic, Equipe, Sandra Teston-Bonnard et Jean Véronis (2004). « Présentation du Corpus de référence du français parlé ». In : *Recherches sur le français parlé* 18. Equipe DELIC, p. 11-42. URL : https://shs.hal.science/halshs-01388193.

Delmotte, Alexandre (oct. 2016). *Les aspects juridiques de la valorisation de la recherche*. Bibliothèque des thèses. mare & martin, p. 958. URL : https://hal.science/hal-01940124.

Hadrossek, Christine et al. (jan. 2023). *Guide de bonnes pratiques sur la gestion des données de la Recherche*. Dans leurs différentes pratiques, les réseaux métiers du CNRS, regroupés au sein de la Mission pour les Initiatives Transverses Interdisciplinaires (MITI) ou soutenus par les Instituts sont en première ligne pour participer au mouvement d'ouverture et de partage des données.Les personnels des organismes de recherche qui les constituent, œuvrent pour mettre en place de bonnes pratiques de gestion et participent également au processus de production des données scientifiques aux côtés des équipes de recherche.Ce guide est la production du groupe de travail inter-réseaux " Atelier Données ". Il s'agit d'un groupe composé de plusieurs réseaux métiers de la MITI (Calcul, Devlog, QeR, rBDD, Renatis, Resinfo, RIS, Medici), du réseau SIST, (labellisé par l'INSU et regroupant les gestionnaires de données environnementales), de l'INIST-CNRS, du GDS EcoInfo et de la Direction des données ouvertes de la recherche (DDOR-CNRS). Dans ce document, nous avons donc voulu témoigner des travaux réalisés au sein de nos réseaux métiers qui rendent compte de la gestion des données de la recherche tout en guidant le lecteur vers des bonnes pratiques.Ce guide est donc un condensé des actions menées autour de la gestion des données de la recherche. Il est le fruit d'un travail collaboratif qui a consisté à collecter, sélectionner et mettre à disposition des ressources vers les actions phares des réseaux métiers, enrichis d'informations et de conseils. URL : https://hal.science/hal-03152732.

Jacobson, Michel et Olivier Baude (2011). « Corpus de la parole : collecte, catalogage, conservation et diffusion des ressources orales sur le français et les langues de France ». In : *Revue TAL : traitement automatique des langues*. Ressources libres 52, p. 47-69. URL : https://shs.hal.science/halshs-01165884.

Longhi, Julien (2022). « The Parascientific Communication around Didier Raoult's Expertise and the Debates in the Media and on Digital Social Networks during the COVID-19 Crisis in France ». In : *Publications* 10.1, p. 7. DOI : 10.3390/publications10010007. URL : https://hal.science/hal-03547921.

Mangeot, Mathieu et Chantal Enguehard (2013). « Des dictionnaires éditoriaux aux représentations XML standardisées ». In : *Ressources Lexicales : contenu, construction, utilisation, évaluation*. Sous la dir. de Gala et al. John Benjamins, p. 24. DOI : 10.1075/lis.30.08man. URL : https://hal.science/hal-00959229.

Minel, Jean-Luc (sept. 2017). « La linguistique face à la multiplication des données langagières numériques. Méthodes, risques et enjeux ». In : *7° Séminaire International de Linguiistique et 3° Symposium de Linguistique Textuelle*. Université Cruzeiros del Sud, Sao Paulo. Sao Paulo, Brazil. URL : https://shs.hal.science/halshs-01590750.

Pincemin, Bénédicte (mars 2020). « La textométrie en question ». In : *Le Français Moderne - Revue de linguistique Française*. Linguistique et traitements quantitatifs 88.1. numéro dirigé par Véronique Magri, p. 26-43. URL : https://shs.hal.science/halshs-02902088.

Rebouillat, Violaine (déc. 2019). « Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs ». Theses. Conservatoire national des arts et metiers - CNAM. URL : https://theses.hal.science/tel-02447653.

Straka, Milan et Jana Straková (août 2017). « Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe ». In : *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*. Sous la dir. de Jan Hajič et Dan Zeman. Vancouver, Canada : Association for Computational Linguistics, p. 88-99. doi : 10.18653/v1/K17-3009. url : https://aclanthology.org/K17-3009.

Sun, Peng et al. (2018). « An overview of named entity recognition ». In : *2018 International Conference on Asian Language Processing (IALP)*. IEEE, p. 273-278.

Webster, Jonathan J et Chunyu Kit (1992). « Tokenization as the initial phase in NLP ». In : *COLING 1992 volume 4 : The 14th international conference on computational linguistics*.

Wissner, Inka (2012). « Les grands corpus du français moderne ». In : *SKY Journal of Linguistics* 25, p. 233-272. url : https://hal.science/hal-03604977.