Les données de la recherche et leur valorisation

au laboratoire LT2D (Lexiques, Textes, Discours et Dictionnaires)

Vade-mecum pour la création et l'utilisation de ressources langagières

Contributions : Thomas Gervais d'Aldin, Marine Delaborde

15 octobre 2025

Table des matières

Introduction

Alors que la recherche en sciences humaines repose de plus en plus sur le traitement de grandes quantités de données, la bonne gestion de celles-ci est un enjeu essentiel pour la recherche. En accord avec les initiatives d'état pour la science ouverte [1], les données de la recherche se doivent d'être diffusées et accessibles lorsque cela est possible. Il existe déjà de nombreuses ressources (guides, outils, plateformes, etc.) mais les informations qu'elles contiennent n'arrivent pas toujours jusqu'aux personnes qui pourraient en avoir besoin. C'est pour faire ce lien (en mettant l'accent sur les pratiques liées aux lexiques, textes, discours et dictionnaires) que l'idée de ce guide est née. Notre objectif n'est donc pas de produire de nouvelles recommandations mais plutôt de faire le point sur les pratiques et les recommandations existantes (gouvernement, consortiums, etc.).

Ce guide est produit dans le cadre de la Chaire de Professeure Junior « Ressources numériques en Sciences Humaines et Sociales » financée par l'Agence Nationale de la Recherche et portée par Marine Delaborde au laboratoire LT2D (Lexiques, Textes, Discours, Dictionnaires, UR 7518) de CY Cergy Paris Université. Il s'inscrit dans une démarche de valorisation des ressources du laboratoire et de création de nouveaux outils pour accompagner les membres du laboratoire dans leurs recherches. Ce guide sera partagé en accès libre, sous licence ouverte et pourra servir à des personnes issues d'autres laboratoires.

Pour établir ce guide, la collaboration des membres du LT2D a été précieuse. Nous avons aussi pu compter sur l'aide de membres de l'Institut Des Humanités Numériques (IDHN) de CY Cergy Paris Université ainsi que du laboratoire Lattice (UMR 8094). Nous remercions chaleureusement toutes les personnes qui ont contribué de près ou de loin à ce guide [2].

Pourquoi valoriser les données de la recherche ? Les enjeux sont grandissants, notamment concernant la sauvegarde, la visibilité ainsi que la réutilisation des données. La prise en compte des principes éthiques et légaux au début d'un projet permet d'envisager la valorisation dès la constitution des données. Pour la récolte, l'annotation et la mise en forme des données, il existe de nombreuses ressources et outils issus de différentes disciplines [3]. Une fois les données constituées, il existe différentes pratiques de partage et de valorisation (appliquer une licence, renseigner des métadonnées, identifier les données de manière pérenne, héberger les données sur un dépôt spécialisé ou non, écrire un *data paper*, faire de la vulgarisation, etc.).

Si vous avez des remarques ou des pistes d'amélioration à propos de ce guide, nous serons ravi·e·s de les prendre en compte pour le faire évoluer. Vous pouvez nous contacter à l'adresse suivante : marine.delaborde@cyu.fr.

1. https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte-pnso/
2. Cécilia Julien, Hélène Manuélian, Luc Massip, Patrick Haillet, Christophe Rey, Jana de Mattos Gibim, Pierre Chartier.
3. Nous référençons de préférence des catalogues d'outils plutôt que les outils directement.

Partie 1

Les enjeux de la valorisation des données de la recherche

D'après la définition officielle de l'OCDE (2007), les données de la recherche (DDR) [1] sont « l'ensemble des informations collectées, produites et utilisées dans le but d'un travail scientifique ». Cette définition englobe de nombreux domaines et il convient de l'adapter à chaque domaine. En linguistique, elles peuvent se définir par un « ensemble d'enregistrements écrits, ou oraux » (REBOUILLAT, 2019). Les volumes croissants de données nécessaires à l'analyse outillée induisent une gestion réfléchie et planifiée des données (MINEL, 2017).

Dans le cadre du Plan national pour la science ouverte, nous sommes incité·e·s à participer à un effort de collaboration pour rendre nos données pérennes, accessibles et réutilisables dans la mesure du possible. Les bénéfices se mesurent à différentes échelles : pour les individus, les équipes de recherche mais aussi pour les communautés scientifiques.

La valorisation des données de la recherche :

— permet la **pérennité** des données et des travaux de recherche (1.1) ;
— donne de la **visibilité** aux travaux de recherche (1.2) ;
— assure leur **réutilisation** dans le cadre d'autres travaux (1.3) ;
— sécurise le **financement** des travaux de recherche (1.4).

1.1 Sauvegarder les travaux

Jusqu'à récemment, la préservation des données de la recherche était loin d'être assurée de manière pérenne. VINES et al. (2014) montraient que la disponibilité des données de la recherche avait tendance à diminuer rapidement avec l'âge de la publication scientifique qui les mentionne. Cette difficulté à accéder aux données est notamment liée au fait qu'une grande partie des données était stockée sur des supports physiques qui devenaient obsolètes et sujets aux pertes et aux dégradations. C'est pour ces raisons que les recommandations gouvernementales intègrent des solutions plus durables.

Dans leur guide de bonnes pratiques sur la gestion des données de la recherche, HADROSSEK et al. (2023) distinguent et définissent différents degrés de pérennité dans les pratiques de conservation.

1. **Le stockage** : « C'est l'étape première qui consiste à déposer les données sur un support numérique pour les rendre accessibles. Cela peut être un ordinateur personnel, un disque partagé ou

1. Pour plus d'informations sur les données de la recherche : https://www.ouvrirlascience.fr/wp-content/uploads/2024/03/24-02-22-Donnees-FR-WEB.pdf

tout autre organe de dépôt. Le stockage permet d'assurer la continuité de l'exploitation sur du court terme. À ce stade, la donnée n'est ni sauvegardée et ni sécurisée. »

2. **La sauvegarde** : « La sauvegarde consiste à dupliquer les données sur un support numérique externe à celui où elles sont stockées. L'objectif est de pouvoir les retrouver en cas de perte ou de dégradation de l'organe de stockage. Il s'agit d'une sauvegarde octet par octet dans une perspective de court ou de moyen terme. La recherche de la préservation de l'intelligibilité des données n'est pas un élément pris en compte »
3. **L'archivage** : il « 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). »

Quelques recommandations

— **multiplier les supports de sauvegarde**, physiques ou dématérialisés : faire des sauvegardes sur une clé usb, un disque dur externe ou un cloud en prenant soin de vérifier les conditions d'utilisation de celui-ci ;

— **conserver un historique des sauvegardes**, afin de garder une trace des modifications apportées. On peut utiliser pour cela des logiciels de gestion de versions comme git. ;

— pour le dépôt des données sur un entrepôt de données de la recherche, **privilégier les entrepôts spécialisés** (cf. 3.1.1) [a].

a. Sources et plus :
https://www.cnil.fr/fr/securite-sauvegarder
https://www.science-ouverte.cnrs.fr/service/partager-et-gerer-mes-donnees/https ://www.ouvrirlascience.fr/science-ouverte-donnees-de-la-recherche/
https://bu.univ-lille.fr/chercheurs-doctorants/science-ouverte/donnees-de-recherche/

1.2 Donner de la visibilité à ses travaux

La visibilité des travaux de recherche est aussi un enjeu essentiel. Selon COLAVIZZA et al. (2020), les articles scientifiques basés sur des données ouvertes sont 25% plus cités que ceux basés sur des données partiellement accessibles ou inaccessibles. Pour cela, le passeport pour la science ouverte recommande de privilégier le mode de diffusion en accès ouvert : « la mise à disposition immédiate, gratuite et permanente sur Internet des publications scientifiques ».

La cohérence de l'identité numérique personnelle et de l'identité numérique des travaux (données, publications, etc.) permet de les rendre davantage accessibles aux humains (sites de références, réseaux de recherche) mais aussi aux machines, notamment à travers les métadonnées (moissonneurs, moteurs de recherche). La visibilité apporte donc une meilleure garantie de transparence et une plus

grande probabilité de reprise des travaux. C'est aussi l'occasion de mettre en valeur le travail de l'équipe de recherche, l'effort de gestion, de description et de partage.

Quelques recommandations

- **Privilégier les revues en accès ouvert**.
- **Assurer la cohérence de son identité numérique**[a] avec un PID (ex : identifiants ORCID, idHAL, idRef) pour permettre la désambiguïsation des homonymes par exemple.
- Lorsque cela est possible, **permettre la reproductibilité de ses travaux** en déposant ses données avec un identifiant pérenne et en les documentant à l'aide de métadonnées.
- **Publier un *data paper*** à propos des données produites.
- D'autres lectures :
  - Bonnes pratiques pour augmenter la visibilité de ses travaux de l'Université de Sherbrooke ;
  - *Augmenter la visibilité et l'impact d'une publication scientifique en maîtrisant le droit d'auteur*, de Janelise Favre1 et Tania Germond (2018, mise à jour en 2020) ;
  - Ouvrir la science :
    - « Je publie, quels sont mes droits ? »
    - « Mettre en œuvre la stratégie de non-cession des droits pour les publications scientifiques ».

a. https://scienceouverte.univ-rennes.fr/vos-identifiants-chercheurs-orcid-idhal

1.3 Assurer la réutilisation

Les coûts financiers et humains liés à la production de nouvelles données sont de plus en plus conséquents et une grande partie des données de la recherche est invisible et/ou peu réutilisée. Au delà de permettre la validation d'une hypothèse, la réutilisation des données d'un projet peut aussi permettre une pleine exploitation du potentiel scientifique du jeu de données par une analyse inédite ou une augmentation des données initiales.

Sur le plan économique, la réutilisation des données dans le cadre d'appels à projets peut être fortement encouragée voire requise pour le financement du projet. Cette pratique présente l'avantage d'économiser les moyens investis à la vérification de la conformité éthique et légale des données, car elle a déjà été réalisée à la constitution du jeu de données.

Pour une bonne réutilisation des données, il est important de leur attribuer une licence adéquate lors du partage, en respectant le principe « aussi ouvert que possible, aussi fermé que nécessaire ».

Quelques recommandations

- S'assurer de la **conformité éthique et légale des données** (cf. 2.1).
- Privilégier **une licence de diffusion ouverte** pour ses jeux de données (cf. 3.2.1).
- Mettre à disposition ses jeux de données sur **des entrepôts maintenus et accessibles** (cf. 3.1.1).

1.4 Sécuriser le financement

Comme évoqué précédemment, le respect des principes de la science ouverte est un critère majeur d'éligibilité au financement de la recherche. Dans le cadre d'appels à projets financés sur fonds publics, les crédits sont alloués sur la base d'appels à projets compétitifs.

Les fondements des processus de sélection de l'ANR [2] :

- intégrité ;
- déontologie ;
- confidentialité des informations ;
- transparence des processus.

Les exigences communes pour le financement sont :

- l'accès immédiat (sans embargo) aux publications scientifiques et aux données de la recherche ;
- l'application d'une licence de diffusion ouverte (cf. 3.2.1) ;
- la rédaction d'un plan de gestion de données (PGD) (cf. 2.1.3) mettant en œuvre les principes FAIR (cf. 2.1.1).

Quelques recommandations

- Planifier le cycle de vie des données (cf. 2.1.2).
- Rédiger un Plan de Gestion des Données (PGD).

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

Partie 2

Préparer la constitution d'un jeu de données

La préparation du jeu de données est l'étape pendant laquelle il est utile de se poser les questions éthiques et légales mais aussi les questions liées aux formats de données.

2.1 Respecter les principes éthiques et légaux

La transition vers la science ouverte est accompagnée de considérations légales (DELMOTTE, 2016) mais aussi éthiques. L'éthique de la recherche occupe d'ailleurs une place de plus en plus importante dans la recherche française, notamment avec la création de nombreux comités d'éthique permettant d'accompagner le montage de projets de recherche.

2.1.1 Les principes FAIR

Les principes FAIR [1] font référence à une gestion responsable, altruiste, et transparente des données de la recherche afin qu'elles soient :

- **Faciles à trouver**
  - Une identification unique pérenne (PID) permet l'accès à la ressource sans ambiguïté.
  - Le dépôt des données sur un entrepôt spécialisé facilite la recherche.
  - Si les données ne peuvent pas être publiées, on peut publier seulement les métadonnées pour donner de la visibilité aux données (une version minimale des métadonnées permet aussi la désambiguïsation entre différents jeux de données).
- **Accessibles**
  - Sur internet, en libre consultation ou de manière restreinte si nécessaire, selon les autorisations obtenues.
  - Sur des plateformes et entrepôts connus de la communauté scientifique.
- **Interopérables** (utilisables dans différents environnements informatiques utilisés par les humains et les machines)
  - Réfléchir au format des données permet d'anticiper la possibilité d'utiliser facilement certains outils de TAL ensuite.
  - Si les données et les métadonnées peuvent être traitées par une machine, leur référencement et leur réutilisation sera plus facile.

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

    - L'utilisation d'un vocabulaire contrôlé (glossaire, lexique, liste de mots clés) adapté au domaine facilite l'interopérabilité.

- **Réutilisables**
    - L'attribution d'une licence de diffusion encourage la réutilisation tout en spécifiant les éventuelles restrictions (cf. 3.2.1).
    - Une documentation suffisante (métadonnées, *data paper*, etc.) permet la compréhension des données sans grand effort.
    - Une structure conforme aux standards de la communauté facilite leur ré-emploi et leur analyse.
    - Choisir un format pivot pour lequel il existe déjà des moyens de conversion favorise la réutilisation des données.
    - Préciser la provenance des données permet la citation correcte de la source au moment de la réutilisation.

2.1.2 Le cycle de vie des données

Dans un projet, les données sont souvent amenées à être modifiées, stockées et traitées par différentes personnes :

> 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 [2].

Un cycle est propre à une utilisation :

1. **La planification** : c'est la définition du projet de recherche et l'anticipation des prochaines étapes du cycle de vie des données. Cette étape permet d'anticiper la façon dont les données seront obtenues et stockées pour faciliter la traçabilité en amont afin de permettre la réutilisation des données.
2. **La collecte** : les données utilisées peuvent avoir plusieurs origines (elles peuvent être créées, modifiées, réutilisées).
3. **L'organisation et l'analyse** : organiser ses données pendant le projet est une étape importante car elle facilitera la gestion du cycle de vie. Cette étape permet de garantir l'identification, la

2. Source : https://opendatafrance.gitbook.io/kit-de-ressources-odf/fiches-pratiques/comprendre/comprendre-le-cycle-de-vie-des-donnees
Autres liens utiles : https://www.universite-paris-saclay.fr/recherche/science-ouverte/le-cycle-de-vie-des-donnees
https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_7.html

localisation, la protection et l'accès à ces données.

4. **La conservation** : la mise en sécurité des données traitées peut passer par la multiplication des supports de sauvegarde (clé USB, disque dur externe, cloud).
5. **Le partage** : une fois que les données d'un projet sont nettoyées et stabilisées, il est important de penser à les publier. Si la publication est possible, les données de la recherche peuvent être publiées via un entrepôt disciplinaire, institutionnel ou plus généraliste (cf. 3.1.1).
6. **La réutilisation** : les données de recherche peuvent servir à d'autres travaux scientifiques permettant de tester de nouvelles hypothèses.

2.1.3 Élaborer un Plan de Gestion des Données (PGD)

Le Plan de Gestion des Données (PGD) ou DMP (*Data Management Plan*) 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 [3].

Ce document peut être évolutif : une première version du PGD est idéale en amont du projet, définissant les différentes modalités de gestion et leur compatibilité avec les principes FAIR. En fonction des évolutions du projet, il est amené à être mis à jour et à rendre compte des spécificités liées à la gestion de certaines données, révélée par les avancées du projet. Il existe des outils permettant de faciliter ce travail (cf. 2.2.2).

De nombreuses ressources sur le PGD sont aussi disponibles sur le site internet de la BU de CY Cergy Paris Université (ressources, webinaires, accompagnement).

2.1.4 Le Règlement Général sur la Protection des données (RGPD)

Dans le contexte évoqué de la multiplication des données numériques, de nombreuses données à caractère personnel sont générées par tous nos services (médical, bancaire, professionnel, etc.), le système législatif a dû s'adapter à cette transformation. En 2012, l'Europe met à jour le cadre législatif qui donnera suite à l'adoption du RGPD en 2016, puis à son application en 2018. Le RGPD est aujourd'hui un fondement éthique et légal de la recherche en sciences humaines. Le non respect du RGPD expose à des sanctions juridiques et financières (BOUCHET MONERET, 2021). 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

3. Source : https://doranum.fr/wp-content/uploads/FicheSynthDMP.pdf

ou l'interconnexion, la limitation, l'effacement ou la destruction (art. 4 du Règlement européen du 27 avril 2016, Ministère de l'économie et des finances).

On peut distinguer différents degrés de sensibilité des données personnelles :

— Moins protégées :

1) **Non personnelles** : les données non personnelles sont des données qui n'ont pas besoin de protection particulière (mail d'accueil d'une entreprise, adresse d'une entreprise, etc.).
2) **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 (avec un nom, une photo, vidéo) ou indirectement (par recoupement de plusieurs données, par exemple, grâce à une date de naissance **et** une adresse postale).

— Plus protégées :

3) **Sensibles** : selon la CNIL, « Les données sensibles forment une catégorie particulière des données personnelles.
   Ce sont des informations qui révèlent la prétendue origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique.
   Le règlement européen interdit de recueillir ou d'utiliser ces données, sauf, notamment, dans les cas suivants :
   — si la personne concernée a donné son consentement exprès (démarche active, explicite et de préférence écrite, qui doit être libre, spécifique, et informée) ;
   — si les informations sont manifestement rendues publiques par la personne concernée ;
   — si elles sont nécessaires à la sauvegarde de la vie humaine ;
   — si leur utilisation est justifiée par l'intérêt public et autorisé par la CNIL ;
   — si elles concernent les membres ou adhérents d'une association ou d'une organisation politique, religieuse, philosophique, politique ou syndicale. »

En accord avec le RGPD, le recueil du consentement est nécessaire pour la récolte de données personnelles de manière légale. Selon la CNIL [4], le consentement doit être obtenu selon 4 critères cumulatifs :

— **libre** : sans contrainte ni influence ;
— **spécifique** : précision du traitement et de la finalité ;
— **éclairé** : les informations suivantes sont à fournir :
   — l'identité du responsable du traitement ;
   — la finalité du traitement ;

4. Source : https://www.cnil.fr/fr/les-bases-legales/consentement

    - les catégories de données récoltées ;
    - la possibilité du droit de retrait du consentement ;
    - selon les cas : transfert éventuel des données hors UE ou utilisation dans le cadre de décisions individuelles automatisées ;
- **univoque** : sans ambiguïté.

Le comité d'éthique de CY Cergy Paris Université propose des modèles de documents pour la récolte du consentement à adapter aux projets (cf. partie 2.2.1).

Désidentification des données

En fonction du projet de recherche, et si cela est nécessaire, différentes stratégies de désidentification des données peuvent être adoptées. Ces stratégies sont intéressantes car elles permettent l'étude de données sensibles tout en évitant de compromettre leur confidentialité et ainsi de rester dans le cadre de la loi. Selon la CNIL [5] :

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. »

La pseudonymisation est « un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire. En pratique, la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.). La pseudonymisation permet ainsi de traiter les données d'individus sans pouvoir identifier ceux-ci de façon directe. En pratique, il est toutefois bien souvent possible de retrouver l'identité de ceux-ci grâce à des données tierces : les données concernées conservent donc un caractère personnel. L'opération de pseudonymisation est également réversible, contrairement à l'anonymisation. La pseudonymisation constitue une des mesures recommandées par le RGPD pour limiter les risques liés au traitement de données personnelles. »

2.2 Les ressources : des instances, des outils et des formations

Avant de constituer des données à partir de zéro, il est intéressant de regarder ce qui peut être réutilisé. Il n'est pas toujours facile d'identifier les données disponibles, ni de savoir à quoi correspond leur format ou les outils qui y sont liés. Des ressources existent et sont de plus en plus accessibles : des données, des instances, des outils et des formations peuvent servir de soutien au travail de recherche.

5. https://www.cnil.fr/fr/technologies/lanonymisation-de-donnees-personnelles (cette page présente aussi différentes stratégies de désidentification)

2.2.1 À CY Cergy Paris Université

CY Cergy Paris Université propose de nombreux services en appui au travail de recherche allant de la formation à l'aide au montage de projet en passant par des recommandations éthiques personnalisées.

Bibliothèque universitaire et école doctorale

La bibliothèque universitaire de CY Cergy Paris Université propose des formations à la recherche documentaire. Des tutoriels sont aussi disponibles sur le site. Ils concernent la recherche documentaire, l'outil Zotero et les bases de données (Europresse, Factiva, etc.). Il est aussi possible de prendre rendez-vous avec un·e bibliothécaire pour les questions liées à la recherche documentaire. De nombreuses ressources sur les données de la recherche sont accessibles sur cette page.

La bibliothèque universitaire propose aussi un service d'appui à la recherche permettant de se former à la science ouverte. Des guides sont disponibles sur la page web de ce service et des formations sont accessibles aux étudiant·e·s de Master et aux doctorant·e·s sur le site des études doctorales.

Via les identifiants CYU, l'accès à des bases de données payantes est possible gratuitement en passant par le portail de la bibliothèque.

Le service *données* de la bibliothèque universitaire propose un accompagnement dans la gestion des données de la recherche, notamment la relecture du PGD [6].

Comité d'éthique de la recherche

Depuis le 18 octobre 2022, CY Cergy Paris Université dispose d'un comité d'éthique de la recherche (CER). Il a pour objectif « de formuler des avis et recommandations afin de mieux intégrer les questionnements éthiques aux projets et protocoles de recherche non-interventionnelle impliquant la personne humaine (hors du cadre de la loi Jardé). » [7]. Des ressources comme des modèles de documents (formulaire d'information et de consentement, formulaire RGPD, etc.) sont par exemple disponibles sur la page web du CER. L'avis du CER ne peut pas être sollicité pour un projet qui a déjà commencé.

Service Ingénierie de Projets

L'obtention de financements de projets est un levier permettant de valoriser la recherche. Le service Ingénierie de Projets propose un accompagnement au montage de projets. L'espace Clic 4 Project permet de solliciter ce service en précisant le projet souhaité. L'accompagnement peut se faire depuis

6. Mail de contact : donnees-bu@ml.u-cergy.fr
7. https://www.cyu.fr/recherche-et-valorisation/appui-aux-chercheurs/ethique-et-integrite-scientifique/comite-ethique-recherche-cer-cy

la clarification des objectifs et le positionnement du projet pour la réponse à un appel à projet (AAP) et se poursuivre avec l'aide au chiffrage et à la rédaction des parties non scientifiques du projet.

Diffusion des savoirs

CY Cergy Paris Université s'engage pour la diffusion de la culture scientifique au plus grand nombre :

— « Le portail HAL de CY Cergy Paris Université est une archive ouverte qui collecte, conserve, valorise et diffuse la production scientifique des chercheurs de l'université. »
— Université Ouverte est un cycle de conférences gratuit et ouvert à toute personne intéressée à « réfléchir, débattre avec des spécialistes autour d'un sujet ».
— L'université participe au concours Ma Thèse en 180 secondes et propose des ateliers de préparation.
— « Autour d'évènements (Open access week), de conférences, de mise à disposition d'outils ou d'actions de sensibilisation, l'université s'engage dans une politique de déploiement à large échelle de l'Open science. »
— Chaque année, l'université organise des activités dans le cadre de la Fête de la science.

2.2.2 Se former à la science ouverte et au numérique

Des outils et des ressources existent pour se former à la science ouverte et faciliter la mise en œuvre de la politique nationale de science ouverte :

— Sur le site Ouvrir la science, des **guides** ont été réalisés par le Comité pour la science ouverte.
— DoRANum : **plateforme de formation** en ligne sur la gestion et le partage des données de la recherche selon les principes FAIR réalisée par l'Inist-CNRS et le GIS « Réseau Urfist » depuis 2015.
— Le service OPIDOR est un **outil** mis à disposition par le CNRS. Il met à disposition un large panel d'outils, de fiches techniques et de conseils pour accompagner le chercheur dans l'élaboration d'un PGD. Le site internet propose des recommandations spécifiques (pour CY Cergy Paris Université par exemple [8]) ainsi qu'un service d'aide.

Différentes organisations proposent des formations aux outils numériques :

— La plateforme FUN Mooc (France Université Numérique) propose de nombreux MOOC gratuits sur l'initiation à l'informatique et à la programmation (Python, Shell bash, etc.), sur les techniques de traitement de données textuelles (manipulation, machine learning, etc.) mais aussi sur la science ouverte.
— Le consortium CORLI propose un ensemble de formations consacrées à la création, la gestion et à l'analyse de corpus.

8. https://dmp.opidor.fr/public_guidance_groups

  - CORLI a publié un outil pour accompagner les linguistes dans leurs démarches liées au cadre éthique et juridique concernant les corpus
- Le réseau Mate-SHS est un réseau de professionnels de la recherche en traitement des données appliquées aux SHS. Il propose les Tuto@Mate, des séminaires de méthodes librement visionnables sur Youtube qui présentent différents outils numériques appliqués aux SHS.
- Le réseau URFIST (Unité Régionale de Formation à l'Information Scientifique et Technique) propose des formations aux outils numériques pour les universitaires.
- Les ateliers du numérique de l'Université Paris 8 : des supports écrits et des vidéos de formation aux outils de la recherche sont accessibles en ligne.
- En général, des formations aux outils numériques sont aussi organisées par les écoles doctorales [9] et les MSH.
- Digit_Hum propose aussi une boîte à outils avec des liens vers des supports de formation.
- Un langage de programmation dispose presque toujours d'une documentation complète disponible en ligne, mais le site W3Schools centralise les éléments de cette documentation en proposant une description structurée des notions et fonctions de base (avec des exemples et des petits exercices) pour la plupart des langages populaires actuels.

2.2.3 Différents formats de données langagières

Avant de constituer ou d'utiliser un jeu de données langagières, il est utile de se poser la question du format des données en fonction de l'usage souhaité.

Les ressources lexicales

Les ressources lexicales sont largement utilisées en traitement automatique des langues (TAL) comme ressources langagières. Avec cette approche, la question du format est incontournable.

> « **Les dictionnaires** traditionnels à caractère encyclopédique (Le Petit Robert, Larousse) ne permettent pas facilement le partage des informations et sont disponibles en formats "propriétaires" et, par conséquent, peu réutilisables en dehors du contexte pour lequel ils ont été développés.
> Par ailleurs, la communauté scientifique, en particulier dans le domaine du TAL, privilégie la réutilisation des ressources lexicales existantes ou des informations contenues dans certaines de ces ressources. Afin de rendre ces ressources partageables entre plusieurs applications informatiques, il est nécessaire que les dictionnaires respectent les standards et les normes disponibles, telles que Text Encoding Initiative (TEI) ou Lexical Markup Framework (LMF) aussi bien pour la structure des données que pour les normes de représentation des informations morphosyntaxiques ». (TODIRASCU, 2018)

9. Par exemple, à CY.

Les lexiques : en TAL, un lexique se concentre sur certaines informations présentes dans des dictionnaires encyclopédiques et peut avoir différentes formes : une collection de formes fléchies avec des informations comme le lemme, la partie du discours, des informations morphosyntaxiques (genre, nombre, mode, temps), la fréquence dans un corpus de grande taille, etc. Il peut aussi être simplement une liste de lemmes mais les collocations sont en général absentes (TODIRASCU, 2018).

Quels formats ?

— **Text Encoding Initiative (TEI)** : standard qui consiste en un ensemble de directives permettant l'encodage d'un texte, compatible avec XML (langage de balisage). La TEI propose un schéma de balisage riche et flexible (hiérarchisation possible), adapté aux besoin des lexicographes. Elle impose la présence d'un en-tête (*TeiHeader*) décrivant des métadonnées. Ce standard est soutenu par une communauté riche et active.

— **Lexical Markup Framework (LMF)** : méta-modèle (norme ISO) qui permet de séparer les parties lexicales, grammaticales et sémantiques mais les noms des éléments ne sont pas imposés. Il est possible d'utiliser l'en-tête TEI pour les métadonnées (MANGEOT & ENGUEHARD, 2013).

— **CONLL** : format tabulé [10] qui permet de décrire des données textuelles sous forme de colonnes selon un nombre d'attributs (lemmes, parties du discours, etc.).

— **Format TBX** : norme ISO pour la représentation des données terminologiques structurées axées sur les concepts.

Les thesaurus contiennent des informations supplémentaires, comme les relations verticales (ex : hyperonymie) ou horizontales (ex : synonymie).

Quels formats ?

— **SKOS** (Simple Knowledge Organization System) : langage de représentation de schémas de concepts qui est une recommandation du W3C pour les vocabulaires contrôlés (thésaurus, taxonomies) et qui utilise le langage RDF ;

— **Zthes** : format XML pour les représentations hiérarchiques et associatives ;

— **ISO 25964** : norme internationale, utilisée pour les thésaurus multilingues (modèle XML).

— **OWL** (Web Ontology Language) : basé sur RDF et pensé pour des ontologies (dans lesquelles il y a souvent des relations plus complexes entre les entrées).

Les corpus textuels

Pour représenter un corpus textuel, la première contrainte est souvent liée aux objectifs d'exploitation définis au moment de la création du corpus. Le corpus peut comporter une structure utile pour l'analyse : une partition en dates, en textes ou selon tout critère d'analyse pertinent. Cette structure

10. Texte brut qui peut être ouvert avec un tableur pour une lecture humaine mais qui est facile à analyser par un parseur.

peut être représentée à travers des fichiers et des répertoires mais aussi dans des langages structurés comme XML. De nombreux outils de textométrie ou de TAL possèdent leur propre format mais permettent quand même d'importer des corpus dans des formats universels.

Tout texte ne peut pas faire partie d'un corpus qui sera partagé en accès libre. Il convient de se renseigner sur les droits d'auteurs ou les autorisations (cf. partie 2.2.1) à obtenir pour les utiliser, comme pour un corpus de copies d'élèves par exemple.

Pour obtenir une version numérique (et exploitable automatiquement avec des outils) de certains textes (textes anciens, manuscrits, PDF, etc.), il est nécessaire d'utiliser des outils de reconnaissance optique de caractères (OCR).

Parfois, un corpus contient une partie alignée sur une autre, comme pour les traductions. Certains formats peuvent être adaptés à ces corpus comme TMX. Largement majoritaire, il permet l'alignement de segments de texte.

Audio et vidéo

Selon la CNIL [11], la voix peut être considérée comme une donnée personnelle dont la récolte nécessite le recueil du consentement (cf. partie 2.1.4).

La récolte de données langagières orales peut se faire à travers des enregistrements audios ou vidéos. Le choix de la vidéo peut présenter l'avantage de permettre des analyses multimodales. En revanche, les données vidéos seront plus volumineuses (variable à prendre en compte pour le stockage et le traitement) et le consentement sera probablement plus difficile à obtenir de la part des personnes participant à l'étude.

Après l'enregistrement, il est recommandé de mettre en application les bonnes pratiques de conservation des données, en faisant des copies sur différents supports sécurisés et distants. Les entrepôts institutionnels, comme ORTOLANG ou COCOON (cf. partie 3.1.1), prévoient des espaces de stockage sécurisés pour les projets en cours.

La transcription

La tâche de transcription permet de passer de l'oral à l'écrit, mais elle ne se limite pas à une simple tâche technique de reproduction car elle prend en compte des enjeux théoriques et interprétatifs (Baude et al., 2006 ; Ochs et al., 1979). Pour transcrire les caractéristiques de l'oral à l'écrit, il existe des conventions de transcription dans lesquelles des phénomènes [12] sont identifiés, hiérarchisés et exemplifiés dans le meilleur des cas, puis associés à une forme de surface pour les représenter. Le choix d'une convention de transcription permet d'assurer l'homogénéité du travail de transcription. Il existe différentes conventions, plus ou moins largement adoptées selon les phénomènes pris en

11. https://www.cnil.fr/fr/definition/donnee-personnelle

12. Par exemple : hésitations, chevauchements, élisions, allongements, montées et chutes intonatives, volume sonore, bruit, etc.

compte.

Quelques exemples :

— ICOR : réalisée par le groupe ICOR, au laboratoire ICAR (UMR 5191), elle concerne « la notation des phénomènes verbaux et vocaux ». Une autre convention vient la compléter pour les aspects multimodaux (gestes, mimiques, etc.).

— VALIBEL : réalisée par le centre de recherche VALIBEL pour la création de sa banque de données orales.

— CIEL : inspirée des conventions des équipes d'ICOR, Freiburg et VALIBEL.

— DELIC : utilisé pour la transcription du *Corpus de référence du français parlé* (DELIC et al., 2004).

Pour effectuer une transcription manuelle, il existe des outils permettant l'alignement de la transcription au signal sonore (cf. partie 2.2.5). Des outils de reconnaissance automatique de la parole peuvent aussi être utilisés comme première étape avant de passer aux éventuelles étapes de correction et d'application d'une convention de transcription [13].

2.2.4 Des bases de données en ligne : des catalogues de ressources

Il existe des bases de données en ligne permettant d'effectuer des recherches dans un corpus ou de trouver des documents.

Articles de presse

Les bases de corpus de presse Europresse et Factiva donnent accès à des articles de journaux grâce à un abonnement de l'université.

— Europresse est une base de presse et actualités comportant plus de 8 000 sources d'information reconnues : presse régionale, nationale et internationale, ressources généraliste et spécialisée, sites web, télévision et radio, biographies et autres. Il est possible de télécharger les articles au format PDF sur le site mais l'outil Press Corpus Scraper de Florent Moncomble permet de récupérer du texte brut.

— 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. Il est possible de télécharger les articles aux formats PDF et RTF.

Droits d'usage : ces bases mentionnent les informations de droits d'auteurs, les articles sont sous copyright pour la majorité. Il existe une exception à ce droit pour des fins pédagogiques et pour les citations courtes. En revanche, si l'article est sous copyright, il n'est pas possible de le faire figurer dans un corpus ouvert et partagé.

13. Certaines études ne nécessitent pas de choisir une convention de transcription.

Bibliothèques d'œuvres littéraires (mais pas seulement)

— Gallica est la bibliothèque numérique de la Bibliothèque nationale de France et de ses partenaires depuis 1997. Les textes sont téléchargeables aux formats PDF, JPEG, EPUB et TXT.

— Le projet Gutenberg est une bibliothèque internationale de versions électroniques d'œuvres littéraires. Les textes sont téléchargeables aux formats PDF, EPUB, MOBI, HTML et TXT.

— Wikisource est une bibliothèque en ligne d'œuvres libres de droits à télécharger aux formats EPUB, MOBI, PDF, HTML, RTF, TXT.

— Ebooksgratuits est une base de données d'e-books d'œuvres libres de droits aux formats DOC, HTML, PDF et EPUB.

— Digit_Hum a listé les sources potentielles pour « trouver et gérer des ressources ».

Droits d'usage : le droit patrimonial dure 70 ans après le décès du dernier auteur. Avant cela, l'œuvre est couverte par le droit d'auteur avec l'exception pédagogique et de citation courte (cf. partie 3.2.2). Le copyright est normalement indiqué sur ces plateformes. Si un corpus est réalisé à partir d'une bibliothèque numérique, il est recommandé de citer les sources de récolte du corpus.

Corpus interrogeables en ligne

— Frantext [14] est une base conçue pour permettre des recherches de formes, lemmes et expressions CQL (avec ou sans expressions régulières) dans un corpus donné. Développée au laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française), elle est disponible en ligne depuis 1998. Les résultats sont affichés dans un contexte de 700 signes. Certains textes sont libres de droits et d'autres non. Un manuel est disponible sur le site. Depuis le CNRTL, on accède sous forme de concordances aux textes libres de droits.

— Sketch Engine est un gestionnaire de corpus et un outil d'analyse textuelle développé par Lexical Computing Limited qui permet de faire des recherches complexes dans de grandes collections de textes. Différents types de corpus sont disponibles dans plusieurs langues. CY Cergy Paris Université ne possède pas d'abonnement à cet outil.

— Google books est la plus large banque d'œuvres du web. Elle est interrogeable *via* l'outil Ngram Viewer avec des requêtes simples à complexes [15] portant sur des suites de *n* tokens. Il existe des corpus pour différentes langues mais il est impossible d'en connaître la composition exacte. Le résultat des requêtes est affiché sous forme de graphique, il n'est pas possible d'avoir accès aux extraits textuels. Il est autorisé de partager les graphiques en citant la source.

14. Ce lien est celui de la BU de CY Cergy Paris Université, qui possède un abonnement payant à Frantext donnant accès à la base intégrale.

15. Un manuel est disponible sur le site : https://books.google.com/ngrams/info

Corpus moissonnés sur le web

Ils peuvent être massifs et hétérogènes. Les corpus moissonnés sur le web peuvent avoir deux principaux types d'utilisation : pour l'étude de phénomènes linguistiques contemporains à grande échelle sur de grandes quantités de données ; pour l'entraînement de systèmes d'intelligence artificielle variés. Selon les objectifs de recherche, il est possible de récolter des données à l'aide de parseurs spécifiques (cf. partie 2.2.4) ou génériques mais dans ce cas il convient de faire attention à la nature des données récoltées selon l'usage souhaité (enjeux éthiques et légaux). Il est aussi possible d'utiliser des corpus massifs déjà récoltés et libres de droits, comme les corpus suivants :

— Common Corpus est le plus gros corpus de textes libres de droits disponible (plus de 500 milliards de tokens). Créé par la start-up Pleias, cette initiative est soutenue par les acteurs de la science ouverte, notamment par le Ministère de la Culture et la Direction interministérielle du numérique (DINUM). L'objectif de ce projet est de créer une alternative libre de droits et transparente aux « mega-corpus » utilisés dans le cadre d'entraînement de LLM à l'international. Ce corpus est librement utilisable à des fins de recherche.

— FRWaC (BARONI et al., 2009) est un corpus du français constitué de 1,3 milliard de tokens récoltés à partir du domaine *.fr*. FRWaC-2010 est disponible dans Sketch Engine (cf. partie 2.2.4).

— Paracrawl est issu d'un projet co-financé par l'Union Européenne. Il s'agit d'un ensemble de corpus de langues de l'UE (ainsi qu'un bonus en langues peu dotées). On peut y trouver des corpus monolingues, mais également des bitextes alignés spécialement prévus pour la traduction automatique. Le corpus est soumis à la licence Creative Commons CC0 *no rights reserved* (cf. partie 3.2.1 pour les licences).

Corpus arborés (*treebanks*)

Les treebanks contiennent des informations d'ordre syntaxique. Ils sont principalement utilisés en TAL pour l'entraînement de modèles d'analyse syntaxique, pour l'amélioration de systèmes de traduction automatique et en linguistique pour effectuer des analyses.

Pour son approche unifiée et évolutive, il est intéressant de mentionner Universal Dependancies (UD). C'est une collection multilingue (plus de 100 langues), collaborative et ouverte de phrases annotées selon des principes d'annotation universels en dépendances syntaxiques. On peut aussi citer le Penn Treebank pour l'anglais et le French Treebank pour le français, mais le site de l'European Language Resources Association (ELRA) en répertorie un grand nombre (parfois payants).

Lexiques

Un lexique est un ensemble structuré de lexèmes accompagnés d'informations linguistiques (morphologie, catégorie grammaticale, sémantique, etc.). Pour le français, on peut citer par exemple [16] :

— Le Lefff (Lexique de Formes Fléchies du Français) est un lexique complet à large couverture du français particulièrement utile pour les tâches de TAL. Il permet entre autres de faire de l'étiquetage morphosyntaxique, de l'analyse morphologique et peut être intégré dans des outils de linguistique de corpus.

— 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). Cette ressource est utile pour faire de l'analyse morphosyntaxique (POS tagging, désambiguisation lexicale). Morphalou3 est une augmentation du Lefff. Cette ressource est sous licence LGPL-LR (Lesser General Public License For Linguistic Resources) (cf. partie 3.2.1 pour les licences).

Bases de données pour le TAL

— Hugging Face : modèles computationnels et jeux de données (*datasets*).

— OPUS : corpus de traductions issues du web.

2.2.5 Des outils de traitement du langage

Les tâches de TAL comme la tokenisation, la reconnaissance d'entités nommées ou l'étiquetage morpho-syntaxique peuvent être utiles pour l'analyse de la langue. C'est aussi le cas des mesures de textométrie qui permettent de faire des statistiques textuelles. Ces tâches sont parfois implémentées dans des modèles à lancer depuis un terminal avec des paramètres particuliers selon les résultats souhaités et cela nécessite plus ou moins de connaissances en programmation selon les modèles. Il existe aussi des outils à utiliser en ligne ou à installer sous forme de logiciel.

Catalogues d'outils de traitement du langage

Parfois développés dans le cadre d'un projet ou d'un laboratoire, certains outils traversent les disciplines et d'autres sont encore méconnus. Afin de trouver les outils utiles et utilisables, des bases de données d'outils existent.

— The Social Sciences Humanities Open Market Place référence des ressources pour les SHS (outils, services, matériel de formation, ensembles de données, publications et *workflows* (flux de travaux)).

16. D'autres ressources lexicales sont disponibles sur la plateforme Ortolang par exemple.

— LINDAT est une initiative soutenue par DARIAH-EU et CLARIN (cf. partie 3.4.2) pour créer un répertoire d'outils pour le TAL.
— Le consortium de recherche CORLI a réalisé et maintient un inventaire réunissant une grande variété d'outils de traitement de corpus de langage (écrit/oraux).
— PostLab est une plateforme dédiée à la démocratisation des solutions d'intelligence artificielle au monde de la recherche académique avec un moteur de recherche.
— TAPoR 3.0 (Text Analysis Portal for Research) est un catalogue d'outils pour l'analyse et la manipulation de données textuelles avec un moteur de recherche.
— Digit_Hum a référencé différents outils pour « analyser un corpus de données ».
— Ortolang répertorie aussi les outils et modèles de traitement de la langue, sans moteur de recherche.
— Le consortium Ariane réalise actuellement un travail de recensement collaboratif de scripts utiles pour les opérations sur les chaînes éditoriales (gestion et conversion de formats).
— Des bibliothèques Python pour le TAL sont intéressantes quand on a quelques bases de programmation pour : les expressions régulières, la lemmatisation, la racinisation, l'étiquetage morpho-syntaxique, la reconnaissance d'entités nommées, la traduction, la classification, l'analyse de sentiments, etc.
— *Introduction to Data Analysis* (GOMIDE & SCHÖCH, 2023) contient une liste d'outils pour l'analyse de données.
— Le projet ATRIUM est un catalogue de ressources pour le traitement des données numériques en art et en sciences humaines.

Récolte automatique de données issues du web

Il existe deux types de récoltes sur le web : le *webcrawling* (moissonnage), qui consiste à récupérer en masse des données depuis des pages internet, par un système de suivi d'hyperliens ; et le *webscraping* (grattage), dont l'objectif est de récolter des données depuis un site ciblé avec un outil spécifiquement conçu pour cela.

Deux approches sont possibles pour la récolte :

— l'approche « code » : réalisation/utilisation d'un programme (bibliothèques Python `BeautifulSoup`, commandes `curl` ou `wget`, etc.) ;
— l'approche « interface » : utilisation d'une interface dédiée, *via* un outil comme gromoteur ou des extensions de navigateur comme Web Scraper ou les outils de Florent Moncomble.

Les sites internet contrôlent de plus en plus la récolte automatique de leurs données et les interdisent parfois. Le *scraping* est légal si les données récoltées sont publiques. Il est encadré par le RGPD, il faut donc faire attention aux données personnelles. Il faut aussi se référer aux conditions générales d'utilisation (CGU) 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. Certains sites proposent

des interfaces API permettant de collecter des données en toute conformité à leurs CGU et en toute légalité. Il est tout à fait possible de demander l'autorisation aux auteur·ice·s d'utiliser leurs données dans le cadre de travaux de recherche.

Des outils pour le traitement de l'oral

Pour l'annotation et la transcription manuelles, il existe Transcriber, ELAN, CLAN, Praat. Pour un inventaire mis à jour, vous pouvez consulter les catalogues d'outils en ligne (cf. partie 2.2.5).

Textométrie

La textométrie est l'application de calculs sur des données textuelles afin de produire des études qualitatives et quantitatives.

— TXM est un logiciel permettant de charger et d'analyser un corpus. Il est possible d'installer différentes extensions, comme Treetagger pour l'étiquetage morpho-syntaxique. Ce logiciel est compatible avec les systèmes d'exploitation Windows 64 bits, MacOS X et Linux 64 bits.

— Iramuteq est une « Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires ». L'utilisation du logiciel demande une installation de R et de Python3 sur sa machine.

— Le Trameur « 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…). ». C'est une application de bureau compatible Windows (seulement). Il est compatible avec `TreeTagger` [17], un système d'étiquetage automatique des catégories grammaticales et de lemmatisation, libre d'usage à des fins de recherche.

— iTrameur est une version en ligne du Trameur, accessible depuis un navigateur web (pas d'installation nécessaire). Pour l'annotation, des scripts proposés par Serge Fleury permettent d'adapter la chaîne UDpipe (pour l'étiquetage morpho-syntaxique, la lemmatisation et les dépendances syntaxiques).

— Tropes est un logiciel d'analyse sémantique permettant d'éditer des ontologies, de faire de l'extraction terminologique ou encore de faire une analyse chronologique de corpus.

— LancsBox est une boîte à outils textométrique développée par la Lancaster University. Ce logiciel est compatible Windows/Mac/Linux. Il réunit un ensemble d'outils comme KWIC (Key Words In Context), un concordancier, GraphColl, un outils d'analyse de collocations, mais aussi des fonctions de lemmatisation et d'étiquetage morphosyntaxique. Droits d'usage : Licence BY-NC-ND Creative Commons

17. Avec une installation supplémentaire de ce module particulier.

Méthodologie de veille scientifique

La veille scientifique est un aspect essentiel du travail de recherche, notamment dans les domaines liés aux humanités numériques, où la recherche est prolifique et les avancées sont rapides avec la course à l'IA.

Les revues et conférences de TAL, linguistique de corpus outillée et humanités numériques sont une source de nouveautés d'un point de vue scientifique mais aussi concernant les méthodes et les outils.

Pour le **TAL**, on peut se référer à la liste de revues et conférences de Wikiversity, ainsi qu'à la revue TAL et à la conférence TALN, sous l'égide de l'Association pour le Traitement Automatique des Langues (ATALA) pour la France.

Pour les **Humanités numériques**, on peut se référer à la liste de revues de Digit_Hum et à la liste de D+ART+H, comprenant aussi des conférences et des séries de livres. Sans oublier la conférence annuelle DARIAH et le congrès Humanistica.

Pour la **linguistique de corpus**, il existe des revues spécialisées que l'on peut retrouver par exemple dans la liste des 458 revues de linguistique de Denis Apothéloz. Pour les conférences, il y a en France les Journées Internationales de Linguistique de Corpus (JLC).

Les articles publiés dans ces revues et actes de conférences ne sont pas les seuls éléments intéressants pour la veille :

— les ***shared tasks*** sont des tâches communes pour lesquelles un groupe de travail impliquant différentes équipes travaillent pour répondre à une problématique commune. Par exemple en TAL, les *shared tasks*, comme DEFT, portent sur des traitements à effectuer sur un même jeu de données langagières.

— les ***workshops*** sont des ateliers collaboratifs ayant pour objectif d'échanger sur une thématique précise du domaine. Par exemple, les conférences annuelles DARIAH organisent différents *workshop* pour réunir autour de projets l'expertise de professionnels de branches variées des Humanités Numériques.

— les ***demos*** sont des présentations de solutions à une problématique définie, il s'agit souvent d'outils en phase finale de développement.

Il existe différents outils permettant l'automatisation de la veille, comme :

— la recherche de conférences par mot-clé pour les conférences dans Calenda et l'abonnement au fil RSS ;

— l'enregistrement de recherches HAL à relancer à la main ;

— les alertes automatiques par mail depuis Google Scholar ;

— les alertes ou abonnements aux fils RSS des revues ;

— l'inscription sur les listes de diffusion spécialisées (cf. partie 3.4.1).

Partie 3

Le partage et la valorisation des données de la recherche

Une fois les données récoltées, le partage de ces dernières est une forme de valorisation permettant, entre autres [1], leur accès et leur réutilisation par d'autres personnes. Cependant, cette valorisation doit respecter certaines conditions selon la nature des données. En effet, le dépôt sur un entrepôt de données, le choix de la licence, le renseignement des métadonnées, l'attribution d'un identifiant unique pérenne et la communication à propos des données peuvent impliquer certaines contraintes à respecter selon les projets de recherche.

3.1 Donner accès aux données

3.1.1 Les entrepôts de données de la recherche

Un entrepôt de données de la recherche permet le dépôt, la description, l'accès et le partage des données de la recherche pour une réutilisation future, conformément aux principes FAIR. Il est recommandé de privilégier le dépôt des données sur des entrepôts spécialisés ou sur un entrepôt institutionnel si aucun autre entrepôt ne semble adapté.

On distingue plusieurs types d'entrepôts :

— Les entrepôts institutionnels sont rattachés à Recherche Data Gouv. CYU dispose par exemple de son propre entrepôt institutionnel.

— Les entrepôts pluridisciplinaires :

  — Nakala : **humanités numériques**. Nakala dépend du TGIR Humanum (cf. partie 3.4.2) et propose deux niveaux de préservation des données (stockage et stockage avancé avec la collaboration avec le CINES). NakalaPress permet de créer un site internet minimal pour présenter ses données. Nakala ne supporte plus la mise en ligne de site web réalisés avec Omeka. Nakaka Quarto View est un outil de la MRSH de Poitier qui permet de créer un site modulaire à partir de ressources Nakala et Quarto avec des scripts Python.

  — Cocoon : **ressources orales** (anthropologie, ethnomusicologie, histoire, linguistique). Cocoon dépend aussi du TGIR Humanum et collabore également avec le CINES.

— Les entrepôts disciplinaires :

  — Ortolang : **linguistique**. Ortolang valorise des outils et des ressources pour le traitement de la langue française (corpus, lexiques, terminologies, outils).

1. cf. partie 1

Il est recommandé de déposer les données sur un seul entrepôt afin d'obtenir un identifiant unique pérenne, un PID (cf. partie 3.1.2). La multiplication des dépôts pourrait nuire au référencement de la ressource. Le choix de l'entrepôt [2] est donc très important et la décision est à mettre en perspective avec les usages du ou des domaines liés au projet.

3.1.2 Attribuer un identifiant unique pérenne

L'attribution d'un identifiant unique pérenne (PID, pour *Persistent IDentifier*) permet d'identifier de manière précise des objets ou des personnes et des institutions.

Pour l'identification d'objets numériques, le DOI (*Digital Object Identifier*) est un PID largement utilisé par les entrepôts de données. Il permet un accès stable aux ressources, même après la suppression ou la modification de l'URL de la ressource en question. Cette méthode permet également une délimitation précise des données concernées par une publication. C'est une norme internationale, hautement interopérable et échangeable. La multiplication des DOI pour une seule et même ressource compromet l'intérêt même du DOI. Certains entrepôts de données fournissent des identifiants uniques propres à leur plateforme, ce ne sont pas pour autant des DOI. Il existe aussi des outils de gestion de DOI comme DataCite ou Crossref. Selon Doranum :

> Les PID contributeur (pour les auteur·ice·s et institutions) visent aussi à désambiguïser les noms et résoudre les problèmes d'homonymie, translittération, etc. Ils augmentent ainsi la visibilité académique [3].

3.1.3 Rendre ses recherches accessibles

Selon le contrat signé avec un éditeur, il n'est pas toujours possible de publier immédiatement et gratuitement le contenu d'un article en accès libre [4] (*open access*).

Des frais de publication [5] peuvent être demandés (ou non) pour la publication, menant à différentes voies de publication en accès libre :

— Voie verte : pas de frais de publication mais souvent un embargo (publication de la version acceptée pour publication : *pre-print*) ;
— Voie dorée : publication en accès ouvert dans des revues nativement ouvertes mais qui impliquentin :inbox des frais de publication ;
— Voie diamant : publication en accès libre et sans frais de publication.

Une chercheuse ou chercheur est titulaire des droits de propriété intellectuelle sur ses publications mais les droits d'auteurs sont parfois cédés aux revues. La stratégie de non-cession des droits en-

2. La liste des entrepôts de confiance de Recherche Data Gouv : https://recherche.data.gouv.fr/fr/entrepots
3. Citation et exemples : https://doranum.fr/wp-content/uploads/Fiche_synthetique_PID.pdf
4. Libguide de la bibliothèque universitaire de CY Cergy Paris Université
5. APC : *Article Processing Charges*

courage les chercheuses et chercheurs à ne plus céder de manière exclusive leurs droits d'auteur aux éditeurs de revues scientifiques. Cela permet de maîtriser la diffusion des manuscrits sans entraîner de frais supplémentaires. Cette stratégie est portée par la cOAlition S [6] (dont l'Agence nationale de la recherche et la Commission européenne). Pour mettre en œuvre cette stratégie il suffit d'avertir l'éditeur de l'application d'une licence libre (cf. 3.2.1) au manuscrit soumis et à ses versions successives [7].

3.2 Les conditions d'utilisation des données

3.2.1 La licence : un instrument juridique

La licence est un instrument juridique permettant de clarifier les droits et les conditions d'utilisation de données. Certains entrepôts imposent une licence particulière et il est nécessaire de comprendre ce qu'elle implique [8]. Les licences ouvertes et permissives accordent des droits d'utilisation et/ou de modification des données selon certaines contraintes (ou non) :

- Etalab est une licence ouverte dédiée aux données publiques françaises ;
- Creative Commons (CC) propose des licences modulables selon différents degrés de permissivité :
  - **CC BY** : attribution
  - **CC BY-SA** : attribution - partage dans les mêmes conditions
  - **CC BY-NC** : attribution - usage non-commercial
  - **CC BY-NC-SA** : attribution - usage non-commercial - partage dans les mêmes conditions
  - **CC BY-ND** : attribution - pas de dérivé/adaptation
  - **CC BY-NC-ND** : attribution - usage non-commercial - pas de dérivé/adaptation
  - **CC0** (Public Domain Dedication) : domaine public = abandon du copyright [9]

Pour le code, Etalab et Creative Commons sont parfois utilisées mais il est plutôt recommandé d'utiliser des licences ouvertes spécifiques aux logiciels comme MIT, Apache 2.0. Des licences copyleft sont aussi utilisées pour le code, comme : GPL, LGPL ou encore AGPL.

Une licence a été spécialement réalisée pour les ressources linguistiques : la LGPL-LR.

6. https://www.coalition-s.org/plan-s-principes-et-mise-en-oeuvre/
7. Source et précisions : Ouvrir la science et CNRS
8. Pour plus d'informations : https://openscience.pasteur.fr/2023/04/12/comment-choisir-une-licence-de-diffusion-pour-ses-donnees-de-recherche/
9. Certains entrepôts et certaines revues imposent cette licence.

3.2.2 Les droits liés aux œuvres

Le droit d'auteur, ou le *copyright*, est l'ensemble des droits dont dispose un auteur sur ses œuvres originales. Il a donc le contrôle sur l'utilisation et la distribution de ces œuvres. L'objectif du droit d'auteur est de protéger les intérêts financiers et moraux de l'auteur dès qu'une œuvre est créée. L'utilisation d'une œuvre sous *copyright* nécessite dans la majorité des cas l'accord explicite de l'auteur.

Le droit à la citation courte est une exception au Code de la propriété intellectuelle. Elle permet 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'œuvre 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 clairement identifiable par l'utilisation de guillemets ou d'une police différente du corps de texte. La loi ne prévoit pas de limite quantifiable en nombre de mots ou de caractères. Elle doit se limiter au passage nécessaire à la compréhension. « Aucune longueur n'est explicitement définie mais elle correspond souvent à une logique de proportionnalité par rapport au texte intégral : en général 10% selon la jurisprudence. ». [10]

3.3 Documenter les données

3.3.1 Les métadonnées : des données à propos des données

Les métadonnées sont un ensemble de données fournissant de l'information sur des données. Elles permettent de situer les données en question dans leur contexte de production. La conception des métadonnées est une phase essentielle de la constitution d'un jeu de données, car les métadonnées sont indispensables à la bonne traçabilité des données. Elles contribuent par exemple à éviter les biais d'analyse dans le cas de l'étude du corpus par d'autres personnes (plus on a d'informations sur les données, plus il est facile de les analyser correctement). Si les données ne sont pas accessibles, l'accessibilité des métadonnées permet d'attester l'existence d'un jeu de données et de donner sa description.

Il existe des **standards de métadonnées**, propres à la nature des données et souvent à un domaine en particulier. La grande majorité des métadonnées sont décrites selon des schémas établis en XML, langage de balisage largement reconnu dans la classification et la hiérarchisation des informations.

L'utilisation d'un standard de métadonnées permet d'améliorer l'indexation des données et d'assurer leur interopérabilité à travers l'emploi de vocabulaires dédiés. Un standard de métadonnées peut

10. Source : http://thesesenligne.parisdescartes.fr/Diffuser/Droits-d-auteur-et-citation

être imposé par un entrepôt de données [11] ou induit par une discipline ou un format de données. Par exemple :

— Dublin Core est un standard de métadonnées générique largement utilisé pour décrire des ressources numériques et physiques. Il comprend 15 éléments de base mais des extensions spécifiques peuvent venir le compléter.
— OLAC (Open Language Archives Community) est un projet international visant à la création puis au maintien d'un dépôt de ressources numériques linguistiques. Le standard OLAC lié à ce projet est une extension au Dublin Core spécifiquement adapté aux ressources linguistiques (type de données, type de discours, langue, champ linguistique, rôles spécifiques : interviewer, annotateur, etc.) en proposant des vocabulaires contrôlés.
— CMDI : solution proposée par CLARIN pour modéliser des métadonnées de façon modulaire (composants de description, composants techniques, composants contextuels). Le CMDI peut intégrer aussi des éléments OLAC.
— teiHeader : la TEI (Text Encoding Initiative) est un standard utilisé notamment pour la structuration et l'annotation de textes littéraires [12]. Les spécifications sont précises et prévoient l'intégration de métadonnées (titre, auteur, traducteur, date de publication électronique, licence, identifiants, index géographique, index chronologique, etc.).

Pour le renseignement des métadonnées, des recommandations sont faites pour chaque standard concernant le format des informations à remplir.

Un outil d'aide au renseignement des métadonnées est en cours de finalisation au LT2D. En attendant, voici déjà des d'outils permettant d'aider à la réalisation de fichiers de métadonnées :

— TEIMETA est un outil CORLI [13] pour l'édition de métadonnées TEI et OLAC.
— Des scripts CLARIN sont disponibles sur le site pour la conversion de différents formats de métadonnées vers CMDI.

3.3.2 Les *data papers* : pour une meilleure description des données

Un *data paper* est un article court décrivant un jeu de données accessible. La rédaction de ce type d'article montre une volonté de mise à disposition du jeu de données à la communauté par le dépôt sur un entrepôt. Il informe par la même occasion de la disponibilité des données et de leurs conditions d'utilisation.

Certaines revues proposent une section et des modèles de *data papers*. La publication de ce type d'article permet de donner de la visibilité aux données et aux travaux qui en découlent. Il consti-

11. Par exemple CMDI pour les centres CLARIN.
12. Le but de la TEI, selon Lou Burnard, l'un des fondateurs 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 en Sciences humaines et sociales ».
13. cf. partie 3.4.2 pour les consortiums

tue aussi une clé de lecture des données à la communauté, favorisant leur réutilisation grâce à une meilleure compréhension des données et de leur contexte de réalisation. Les objectifs sont donc :

— la description d'un jeu de données et des métadonnées associées ;
— la description de la méthode d'obtention des données ;
— la démonstration du potentiel de réutilisation du jeu de données ;
— seulement la description, aucun résultat ni analyse.

Des ressources existent pour aider à la rédaction d'un *data paper* :

— l'URFIST de Lyon propose un guide d'aide à la rédaction ;
— l'URFIST de Bordeaux propose des sessions à distance « Comment publier un *data paper* ? » ;
— à partir du DOI, Recherche Data Gouv propose un outil permettant la génération d'une ébauche de *data paper*.

3.4 Des canaux de communication : pour la centralisation de l'actualité scientifique

3.4.1 Des listes de diffusion

Les listes de diffusions permettent la centralisation et le partage de l'actualité scientifique d'un réseau lié à un domaine d'études (publications, appels à soumissions, évènements scientifiques, formations, offres d'emploi, etc.). Quelques listes de diffusion dans nos domaines :

— Linguist list : linguistique, anglophone, supportée par le département de linguistique de l'université de l'Indiana à Bloomington (USA) ;
— parislinguists : linguistique, France ;
— corpora : linguistique de corpus ;
— CORLI : activités du consortium CORLI (cf. partie 3.4.2) corpus, langues, interactions ;
— LN : traitement automatique des langues (TAL), liste supportée par l'Association pour le Traitement Automatique des Langues (ATALA) ;
— lingtyp : liste de The Association for Linguistic Typology ;
— histling-l : Historical Linguistics (Yale, USA) ;
— rfs : liste du Réseau francophone de sociolinguistique ;
— RISC : liste du Relais d'Information sur les Sciences de la Cognition ;
— legram : sciences de l'information et de la communication, analyse du discours ;
— theuth : épistémologie et histoire des sciences ;
— SAES : liste de discussion de la SAES, anglicistes français.

— ALAES : liste de discussion de l'Association des Linguistes Anglicistes de l'Enseignement Supérieur ;
— chercheurs SDL : sciences du langage ;
— histoire_eco : histoire économique ;
— hdls : histoire de la santé ;
— humanist : liste de discussion du séminaire Humanist, humanités numériques ;
— DH : humanités numériques, francophone, liste supportée par l'association Humanistica ;
— Ariane : activités du consortium Ariane (cf. partie 3.4.2), intelligence artificielle et éditions numériques ;
— Quanti : enseignement des méthodes quantitatives dans les sciences sociales.

3.4.2 Des consortiums

Dans la recherche, les consortiums sont des réseaux réunissant des personnels d'unités autour d'une thématique et d'objectifs communs. La durée d'un consortium est en général définie à l'avance pour une durée donnée. Les consortiums qui suivent s'inscrivent dans la logique de la science ouverte.

Les consortiums de L'IR* HumaNum

L'Infrastructure de Recherche « étoile » Huma-Num a pour vocation de construire une infrastructure numérique pour les SHS au niveau international. Pour cela, elle labellise et finance des consortiums et assure une présence dans les Maisons des Sciences de l'Homme (MSH). Actuellement, il existe 10 consortiums-HN et certains d'entre eux traitent de thématiques liées au langage :

— Ariane : Analyses, Recherches, Intelligence Artificielle et Nouvelles Editions numériques ;
— CANEVAS : Consortium pour l'annotation, l'analyse et l'archive de la vidéo appliquées aux activités scientifiques
— CORLI 2 : Corpus, Langues et Interactions ;
— DISTAM : DIgital STudies Africa Asia and the Middle east ;
— 3DHN : Consortium 3D pour les humanités numériques ;

Des Infrastructures européennes

Deux infrastructures sont complémentaires (et collaborent parfois) au niveau européen :

— CLARIN (Common Language Resources and Technology Infrastructure) propose des ressources pour la recherche en Sciences Humaines, principalement axées linguistique et TAL.
  — **CORLI** est un centre [14] CLARIN K (*Knowledge Centre*).

14. Les différents centres CLARIN : https://www.clarin.eu/content/overview-clarin-centres

  - **ORTOLANG** est un centre CLARIN B (*Service Providing Centre*) ;
  - **COCOON** est un centre CLARIN C (*Technical Centre*).
- DARIAH (Digital Research Infrastructure for the Arts and Humanities) a une mission plus large : elle réunit 22 pays européens et 197 institutions partenaires sur des projets de recherche variés en sciences humaines numériques.

3.5 La vulgarisation des données : pour le grand public

La vulgarisation est un pont entre le monde de la recherche et le grand public. C'est un moyen utile de promouvoir ses travaux et d'augmenter leur visibilité. Différents moyens existent pour cela, comme la rédaction de livres grand public ou d'articles pour des médias comme The conversation qui publie des articles de vulgarisation écrits par des universitaires. Le service de valorisation de CY Cergy Paris Université (cf. partie 2.2.1) peut représenter une aide précieuse pour accompagner la vulgarisation de la recherche.

3.5.1 Le référencement sur des répertoires d'expert·e·s

Les répertoires d'expert·e·s sont des listes sur lesquelles des spécialistes d'un sujet peuvent être répertoriés de manière à être sollicité·e·s par les médias sur un sujet en lien avec leur expertise. On peut espérer que plus ces plateformes seront complètes, plus les journalistes inviteront de véritables expert·e·s du domaine et des questions qui les intéressent [15].

- EXPERTalia est un répertoire belge d'experts et expertes « issus de la diversité d'origine » à destination des médias ;
- Expertes.com est un annuaire des femmes expertes francophones.

3.5.2 Des podcasts et des chroniques

Les podcasts permettent de vulgariser les avancées scientifiques, dans un format plus ou moins court et régulier. Quelques exemples en linguistique :

- les interventions de Jean Pruvost sur Radio France.
- Vox (EFL) est le podcast du Labex EFL, lancé en 2021 ;
- Parler comme jamais est soutenu par la DGLFLF et allie linguistique et sujets de société ;
- Avec la langue est un podcast France Inter de la linguiste Julie Neveux ;
- La chronique langue de Laélia Véron sur France Inter ;

15. Cf. distinction entre *opinion* et *expertise* par Julien Longhi : https://theconversation.com/linguistique-education-quand-les-medias-confondent-opinion-et-expertise-215615

— Complètement gaga ... des parlers d'ici est la chronique d'Olivier Glain sur France Bleu ;
— Dites-le en marseillais est la chronique de Médéric Gasquet-Cyrus sur France Bleu Provence.

3.5.3 Des ateliers et des conférences

— TED Talks : conférences courtes et percutantes ;
— La Nuit des idées organisée par l'Institut français ;
— La fête de la Science (cf. partie 2.2.1 pour l'organisation à CY Cergy Paris Université).

Il existe d'autres stratégies de vulgarisation et la limite reste celle de votre imagination : chaînes Youtube, jeux vidéos, jeux de société, applications mobiles, comptes sur les réseaux sociaux, etc.

Conclusion

Ce guide a été réalisé au LT2D en 2024 dans sa première version et a pour vocation d'être diffusé en accès libre à toute personne intéressée par la valorisation des ressources langagières. Notre objectif est de le rendre évolutif, c'est pourquoi une nouvelle version de ce guide sera publiée après chaque modification.

Comme précisé en introduction, si vous avez des remarques ou des pistes d'amélioration à propos de ce guide, nous serons ravi·e·s de les prendre en compte pour le faire évoluer. Vous pouvez nous contacter à l'adresse suivante : marine.delaborde@cyu.fr.

Glossaire

cloud Réseau de serveurs en ligne permettant le stockage dématérialisé des données. Pour les données de la recherche, il est recommandé de privilégier les serveurs institutionnels. 3, 8

copyleft « Le copyleft est une méthode générale pour rendre libre un programme (ou toute autre œuvre) et obliger toutes les versions modifiées ou étendues de ce programme à être libres également. »
Source : https://www.gnu.org/licenses/copyleft.fr.html. 25

data paper « Le *data paper* (data article, data descriptor) est une publication qui décrit un jeu de données scientifiques, notamment à l'aide d'informations structurées, appelées métadonnées. Le data paper fournit une voie formalisée au partage des données contrairement à l'article de recherche classique qui teste des hypothèses ou présente de nouvelles analyses. Le data paper et l'article de recherche classiques se complètent. »
Source : https://doranum.fr/data-paper-data-journal/. 4, 7

DOI « Le Digital Object Identifier (DOI, littéralement "identifiant d'objet numérique") est le cœur d'un mécanisme d'identification de ressources numériques, comme les revues, articles scientifiques, rapports, vidéos, etc. Il est parfois comparé aux ISSN ou ISBN pour le web, mais c'est aussi une alternative à l'instabilité des URL par l'association de la localisation du document et des métadonnées qui lui sont liées. ». Le DOI est un PID largement utilisé pour les objets.
Source : https://objs-fr.hypotheses.org/367. 24, 28

donnée personnelle Une donnée personnelle est « toute information se rapportant à une personne physique identifiée ou identifiable ». Une personne peut être identifiée directement (par son nom, son prénom) ou indirectement (téléphone, mail, données biométriques, voix, image, etc.).
Source : https://www.cnil.fr/fr/definition/donnee-personnelle. 20

données de la recherche (DDR) Selon l'OCDE : « 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. »
Source : https://recherche.data.gouv.fr/fr/page/quelles-donnees-de-recherche. 2

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. ».
Source : https://socle.univ-rennes2.fr/vos-besoins/diffuser-ses-donnees-dans-entrepot. 3, 23

étiquetage morpho-syntaxique « L'étiquetage morpho-syntaxique (plus connu sous le nom anglais Part-Of-Speech tagging ou POS-tag) est le processus d'attribution de sa catégorie morpho-syntaxique (catégorie grammaticale) à chaque mot-forme. En d'autres termes, il consiste à étiqueter chaque mot d'une phrase avec sa partie appropriée du discours »
Source : https://corli.huma-num.fr/wp-content/uploads/2022/05/Fiche-10-Etiquetage-morpho-syntaxique.pdf. 19, 20, 21

FAIR Les principes FAIR (*Findable, Accessible, Interoperable, Reusable*) ont été formulés par un groupe de travail en 2014 pour caractériser un traitement des données qui leur permette d'être plus faciles à trouver, accessibles, interopérables et réutilisables.
Source : Document de travail *Draft FAIR*. 5, 6, 12, 23

langage 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. »
Source : https://www.arthurperret.fr/digithum-glossaire-hn.html. 14, 26

lexème « Unité minimale de signification appartenant au lexique. »
Source : https://www.cnrtl.fr. 19

licence de diffusion « Document accompagnant la publication en ligne d'un jeu de données qui détermine les conditions de réutilisation des données publiées. Ces conditions ne peuvent apporter de restrictions à la réutilisation que pour des motifs d'intérêt général et de façon proportionnée. Elles ne peuvent avoir pour objet ou pour effet de restreindre la concurrence. Le recours à une licence est obligatoire lorsque l'administration soumet les réutilisations au paiement d'une redevance. »
Source : https://www.cnil.fr/sites/cnil/files/atoms/files/guide-open-data.pdf. 5, 7

LLM « Modèle statistique de la distribution d'unité linguistiques (par exemple : lettres, phonèmes, mots) dans une langue naturelle. Un modèle de langage peut par exemple prédire le mot suivant dans une séquence de mots. On parle de modèles de langage de grande taille ou « Large Language Models » (LLM) en anglais pour les modèles possédant un grand nombre de paramètres (généralement de l'ordre du milliard de paramètres ou plus) comme GPT-3, BLOOM, Megatron NLG, Llama ou encore PaLM. »
Source : https://www.cnil.fr/fr/definition/modele-de-langage. 18

métadonnées Les métadonnées sont les données à propos des données. « Dans un cadre scientifique, les métadonnées contribuent à décrire les ressources, les données de recherche et les productions réalisées (article, dépôt, photo, mesure, logiciel, page Web, etc.). ». Elles peuvent documenter les ressources de manière interne ou externe.
Source : https://doranum.fr/wp-content/uploads/Fiche-Synthétique-Métadonnées.pdf. 3, 6, 14, 26

moissonnage du web *Web scrapping* en anglais, « technique d'extraction automatique de données à partir d'un ou de plusieurs sites Web dans le but d'utiliser celles-ci, après traitement, dans un autre contexte. »
Source : https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/26507119/moissonnage-du-web. 3

MOOC (*Massive Open Online Course*) cours en ligne ouvert et massif. 12

parseur Outil informatique qui parcourt le contenu d'un document pour l'analyser, vérifier sa syntaxe ou extraire des informations. 14, 18

PGD « Le Plan de Gestion de Données (PGD) ou Data Management Plan (DMP) 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. »
Source : https://doranum.fr/plan-gestion-donnees-dmp/plan-de-gestion-des-donnees-fiche-synthetique_10_13143_cgv4-0k53/. ii, 5, 8, 11, 12

PID « Un identifiant pérenne (Persistent identifier ou PID) est un identifiant qui est assigné à un objet de façon permanente. Il est disponible et gérable à long terme ; il ne changera pas si l'objet est renommé ou déplacé (changement de site, d'entrepôts de données...). »
Source : https://doranum.fr/wp-content/uploads/Fiche_synthetique_PID.pdf. 4, 6, 24, 35

reconnaissance d'entités nommées « En anglais « Named-entity recognition » (NER), sous-tâche d'extraction d'informations qui cherche à localiser et classifier les mentions d'entités nommées dans du texte non structuré en catégories prédéfinies, emplacements, codes médicaux, expressions de temps, quantités, valeurs monétaires, pourcentages, etc. »
Source : https://www.cnil.fr/fr/definition/reconnaissance-dentites-nommees . 19, 20

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. ».
Source : https://www.cnil.fr/fr/comprendre-le-rgpd. 8, 20

TAL « Le traitement automatique des langues (TAL) est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer des outils de traitement de textes et de la parole (incluant la parole signée) pour diverses applications. Le TAL combine les apports de la linguistique computationnelle — modèles du langage basées sur des règles —, et des méthodes à base statistique, d'apprentissage machine et d'apprentissage profond. Le traitement automatique des langues est l'un des domaines d'application majeur de l'Intelligence Artificielle. »

Source : https://www.inshs.cnrs.fr/fr/traitement-automatique-de-la-langue . 6, 13, 18, 19, 20, 22, 28, 29

TEI La *Text Encoding Initiative* (abrégé en TEI, en français « initiative pour l'encodage du texte »). « La TEI c'est principalement un ensemble de recommandations de mise en forme de l'information dans les textes électroniques.
Lancé au départ avec SGML puis ayant migré sur XML. Il s'agit de baliser les texte afin de faciliter le traitement de ces documents. Ce format est aussi voué à être interdisciplinaire afin de promouvoir le partage de la connaissance et de l'information. ».
Source : https://stph.scenari-community.org/contribs/doc/cdt/tei1/co/definition.html. 14

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. »
Source : https://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-teiHeader.html. 14

TMX « La norme TMX (Translation Memory eXchange) décrit un format de fichier basé sur le standard XML. TMX est tout particulièrement destiné à fournir des traductions de phrases dans différentes langues. »
Source : http://www.xmlfacile.com/guide_xml/fichier_de_traduction_tmx_1.php5. 15

token La plus petite unité d'informations détectée lors de la tokenisation. Selon l'étape de tokenisation, un token peut être un mot, un signe de ponctuation, un nombre ou encore une abréviation. 17, 18, 36

tokenisation Étape de segmentation lexicale, fondamentale en traitement automatique du langage naturel. Elle consiste à découper un texte en unités minimales, appelées tokens, qui peuvent ensuite être traitées par des modèles d'apprentissage automatique par exemple. 19

treebank Un *treebank* (corpus arboré) est une base de données qui contient des phrases annotées avec des informations syntaxiques. 18

XML *eXtensible Markup Language* ou « langage de balisage extensible » en français. Il s'agit d'un métalangage informatique de balisage générique qui est un sous-ensemble du *Standard Generalized Markup Language* (SGML). 14

Bibliographie

BARONI, M., BERNARDINI, S., FERRARESI, A., & ZANCHETTA, E. (2009). The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*, 209-226 (Cité page 18).

BAUDE, O., BLANCHE-BENVENISTE, C., CALAS, M.-F., CAPPEAU, P., CORDEREIX, P., GOURY, L., JACOBSON, M., DE LAMBERTERIE, I., MARCHELLO-NIZIA, C., & MONDADA, L. (2006). *Corpus oraux, guide des bonnes pratiques 2006*. Presses universitaires d'Orléans ; CNRS Éditions. (Cité page 15).

BOUCHET MONERET, F. (2021, mars). Les données personnelles de recherche et le RGPD [Guide sur les données personnelles de recherche dans le cadre du RGPD]. https://hal.univ-lorraine.fr/hal-03636697 (Cité page 8).

COLAVIZZA, G., HRYNASZKIEWICZ, I., STADEN, I., WHITAKER, K., & MCGILLIVRAY, B. (2020). The citation advantage of linking publications to research data. *PloS one*, *15*(4), e0230416 (Cité page 3).

DELIC, E., TESTON-BONNARD, S., & VÉRONIS, J. (2004). Présentation du Corpus de référence du français parlé [Equipe DELIC]. *Recherches sur le français parlé*, *18*, 11-42. https://shs.hal.science/halshs-01388193 (Cité page 16).

DELMOTTE, A. (2016, octobre). *Les aspects juridiques de la valorisation de la recherche*. mare & martin. https://hal.science/hal-01940124 (Cité page 6).

GOMIDE, A., & SCHÖCH, C. (2023). Introduction to data analysis. In C. SCHÖCH, J. DUDAR & E. FILEVA (Éd.), *Survey of Methods in Computational Literary Studies (= D 3.2 : Series of Five Short Survey Papers on Methodological Issues)*. CLS INFRA. https://doi.org/10.5281/zenodo.7892112 (Cité page 20).

HADROSSEK, C., JANIK, J., LIBES, M., LOUVET, V., QUIDOZ, M.-C., RIVET, A., & ROMIER, G. (2023, janvier). Guide de bonnes pratiques sur la gestion des données de la Recherche. https://hal.science/hal-03152732 (Cité page 2).

MANGEOT, M., & ENGUEHARD, C. (2013). Des dictionnaires éditoriaux aux représentations XML standardisées. In GALA, NURIA, ZOCK & MICHAEL (Éd.), *Ressources Lexicales : contenu, construction, utilisation, évaluation* (p. 24). John Benjamins. https://doi.org/10.1075/lis.30.08man (Cité page 14).

MINEL, J.-L. (2017). La linguistique face à la multiplication des données langagières numériques. Méthodes, risques et enjeux. *7° Séminaire International de Linguistique et 3° Symposium de Linguistique Textuelle*. https://shs.hal.science/halshs-01590750 (Cité page 2).

OCDE. (2007). *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*. https://doi.org/https://doi.org/10.1787/9789264034020-en-fr (Cité page 2).

OCHS, E., et al. (1979). Transcription as theory. *Developmental pragmatics*, *10*(1), 43-72 (Cité page 15).

REBOUILLAT, V. (2019, décembre). *Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs* [Theses]. Conservatoire national des arts et metiers - CNAM. https://theses.hal.science/tel-02447653 (Cité page 2).

TODIRASCU, A. (2018). Dictionnaires électroniques : normes de représentation. *Cahiers du plurilinguisme européen*, (10) (Cité pages 13, 14).

VINES, T. H., ALBERT, A. Y. K., ANDREW, R. L., DÉBARRE, F., BOCK, D. G., FRANKLIN, M. T., GILBERT, K. J., MOORE, J.-S., RENAUT, S., & RENNISON, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age [Publisher : Elsevier]. *Current Biology*, *24*(1), 94-97. https://doi.org/10.1016/j.cub.2013.11.014 (Cité page 2).