



# Données de recherche: La transparence et les meilleures pratiques

Helene N. Andreassen, PhD  
UiT Université arctique de Norvège

*Ecole d'été « Phonologie de corpus »  
Lausanne, 22-26 juillet 2019*



Research data @ UiT

# Définir des données de recherche

« ... tous les enregistrements, notes et rapports qui sont produits ou générés au cours de la recherche, et qui sont considérés ayant un intérêt et/ou un potentiel scientifique. Le format de celles-ci inclut, mais n'est pas limité à, des chiffres, du texte, du code, des photos, des films et du son. » (traduit de UiT Arctic University of Norway, 2017)



# Données dans votre projet doctoral

Avez-vous déjà collecté des données de recherche dans votre projet doctoral, ou est-ce que vous pensez le faire? Si oui, quel type de données? Sinon, pourquoi pas?

**Passez 5 minutes à discuter ces questions avec votre voisin.** Est-ce que vous voyez des défis similaires ou différents en ce qui concerne la gestion de vos données?



# Objectifs d'apprentissage de cet après-midi

- Mieux comprendre le raisonnement derrière la bonne gestion des données de recherche (*research data management*, RDM).
- Mieux comprendre pourquoi l'archivage en libre accès devrait être l'option préférée, et pourquoi dans de nombreux cas ceci n'est pas possible.
- Mieux comprendre les exigences et les attentes concernant la gestion des données (et vous préparer à l'avenir).

\*\*\*

- Connaître les phases principales du cycle de vie des gestion des données.
- Connaître les principes de citation de données.
- Être au courant de la discussion RDM en linguistique au niveau international.

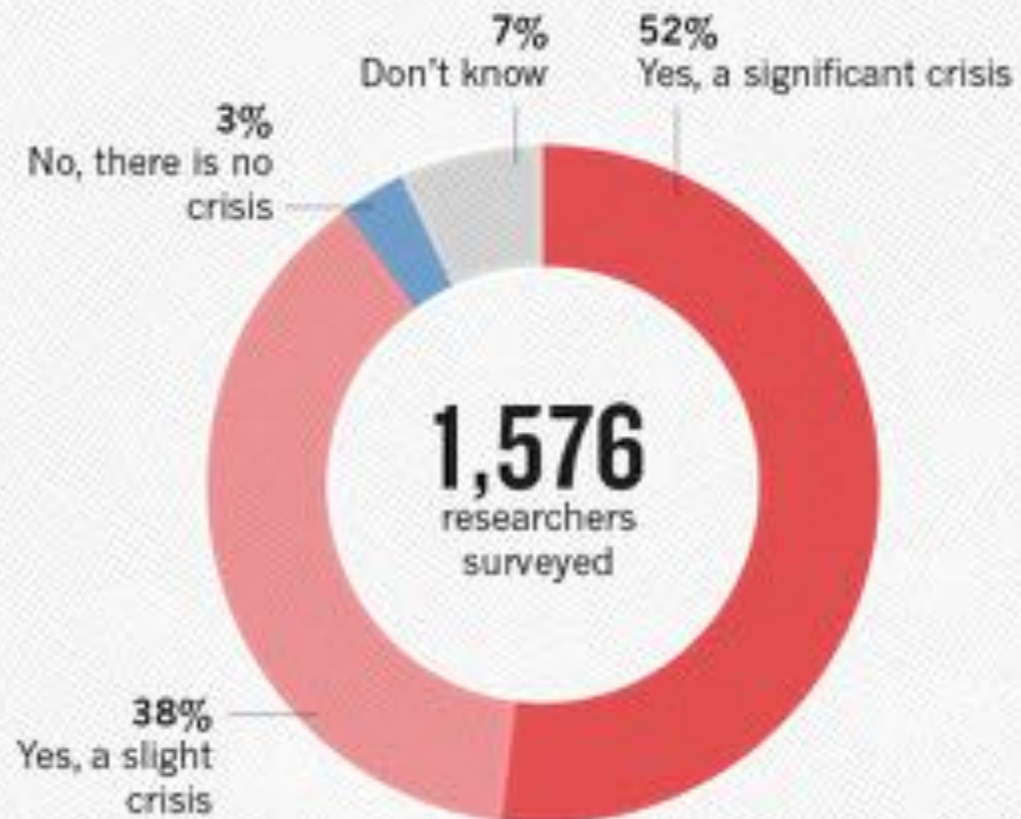


# Plan

- Introduction à la gestion de données
  - Pourquoi on devrait se préoccuper de la gestion de données
  - Exigences et attentes
  - Services de soutien & lieux de discussion
- Les meilleures pratiques
  - Le cycle de vie de gestion des données
  - La citation de données



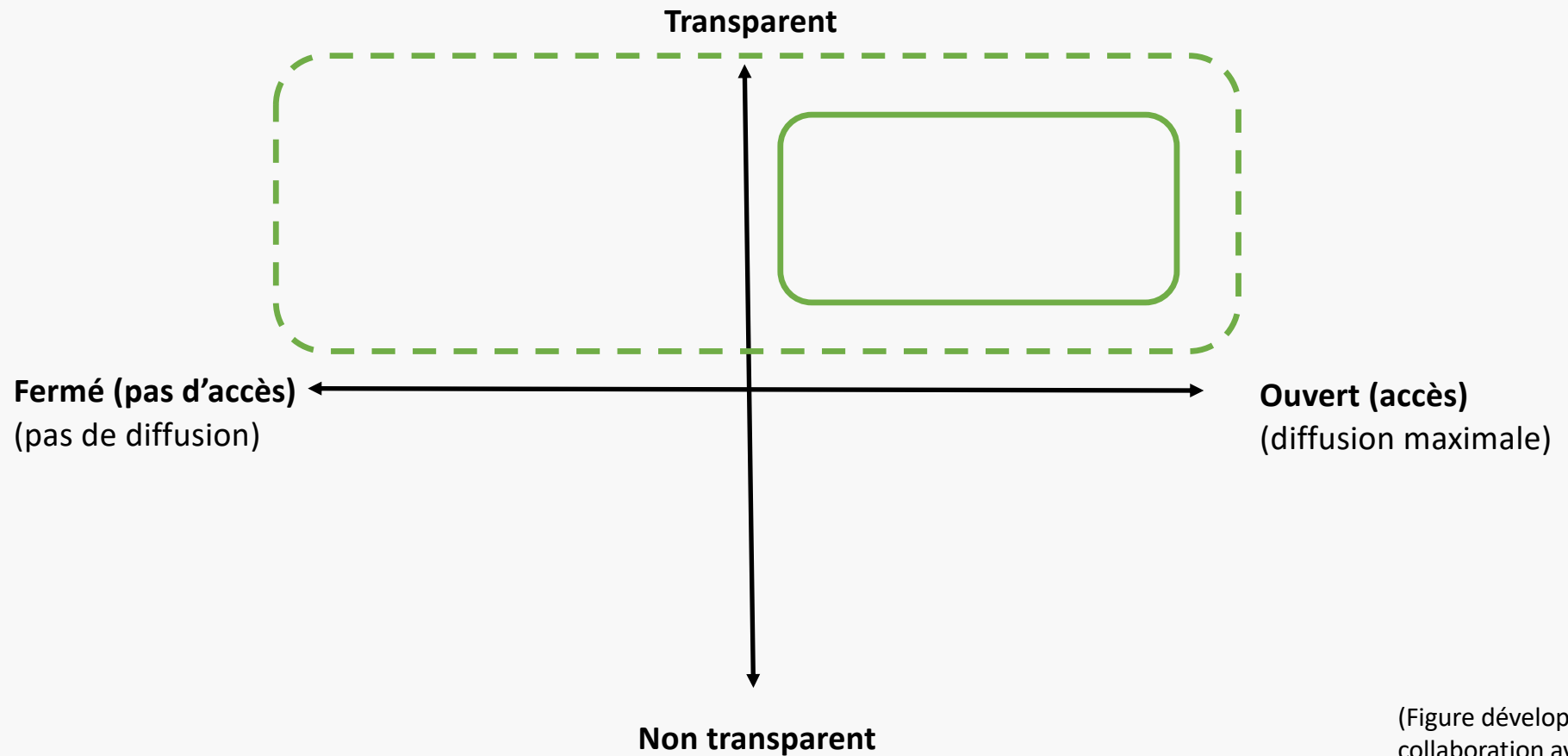
## IS THERE A REPRODUCIBILITY CRISIS?



nature



# Transparence et libre accès dans la diffusion de la recherche



(Figure développée en collaboration avec Aysa Ekanger, UiT)



**REDUCTION DE  
FABRICATION ET  
FALSIFICATION DE DONNÉES**



# Reproductibilité de la recherche

1. Transparence de la méthode appliquée lors de la collecte et l'analyse des données.
2. Transparence du statut des données, y compris le type de données, et si/comment le lecteur peut y avoir accès.

« I have come to think of the **accuracy of data as a serious problem** for all linguistics journals, and for the field at large. I do not see any way to deal with it unless we constantly remind ourselves and our students of the importance of working to maintain the quality of the data that we use.

[...]

Besides actual errors in cited data, the attribution of data and even well-known analyses to the last linguist who discussed the material in print, rather than to the original published source of the data and/or the analysis, is **so common that it has become a cliché.** »

(Thomason, 1994, p. 410, bolding ours)



# Reproductibilité et validation de la recherche

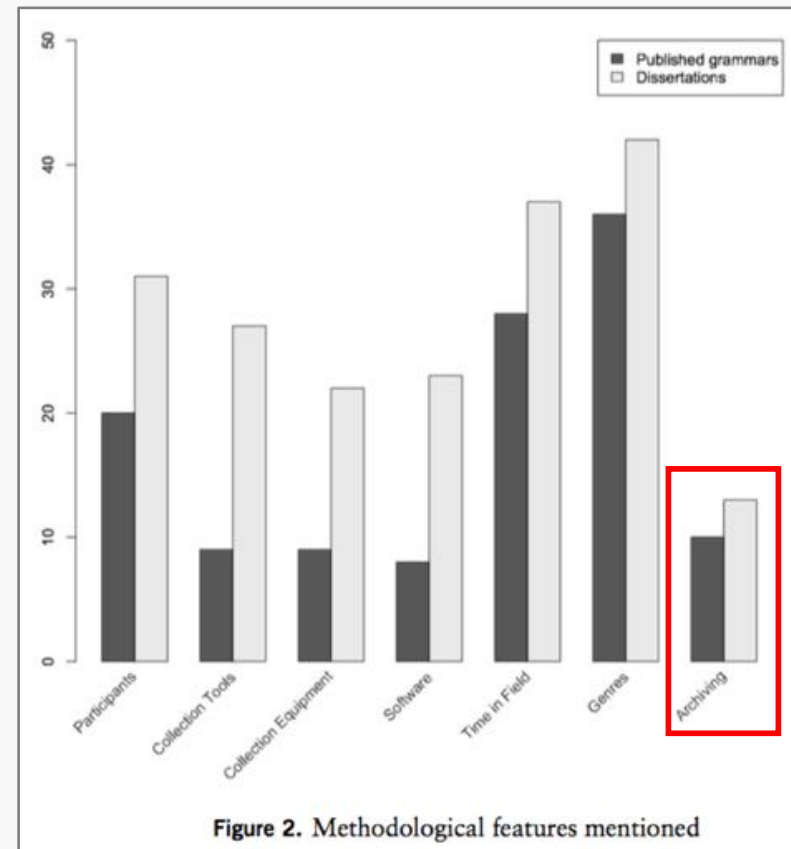
Analyse de 50 grammaires et 50 thèses  
(2003-2012)

## 1. Méthode

+ Période de collecte, type de données,  
participants

÷ Outils, équipement, logiciels, archivage

(Gawne et al., 2017, p. 172)



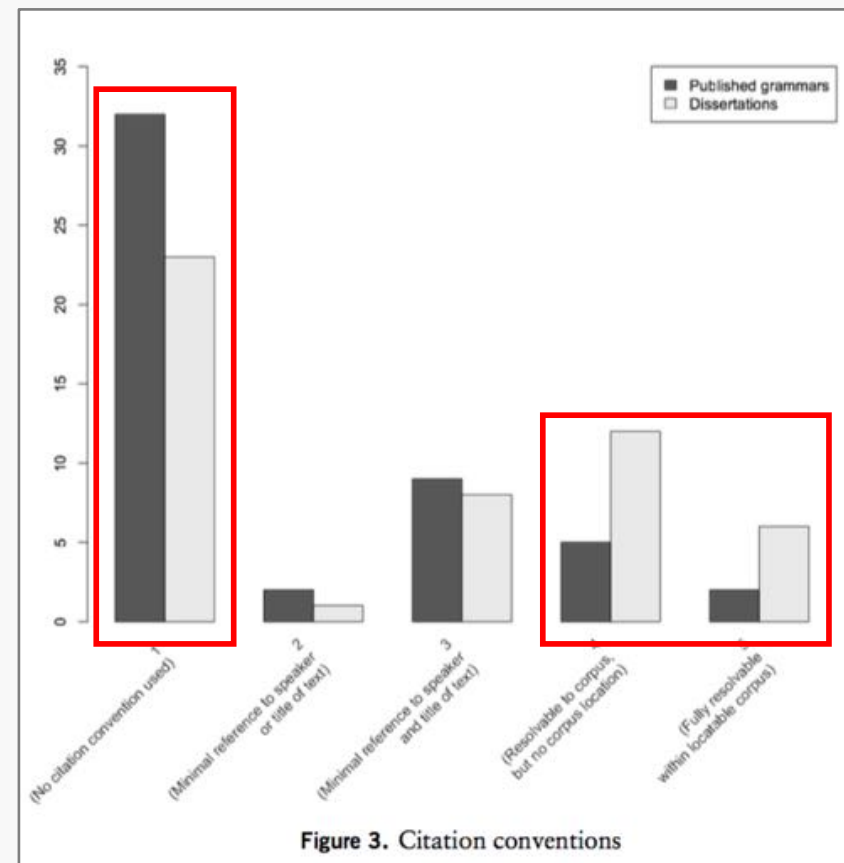
# Reproductibilité et validation de la recherche

Analyse de 50 grammaires et 50 thèses  
(2003-2012)

## 2. Conventions de citation

- La majorité ne donne pas de citation à côté des exemples.
- Un certain nombre de thèses réfère au corpus, archivé ou pas.

(Gawne et al., 2017, p. 175)



# Préservation des informations de valeur

Examen de 516 articles de la période 1991-2011

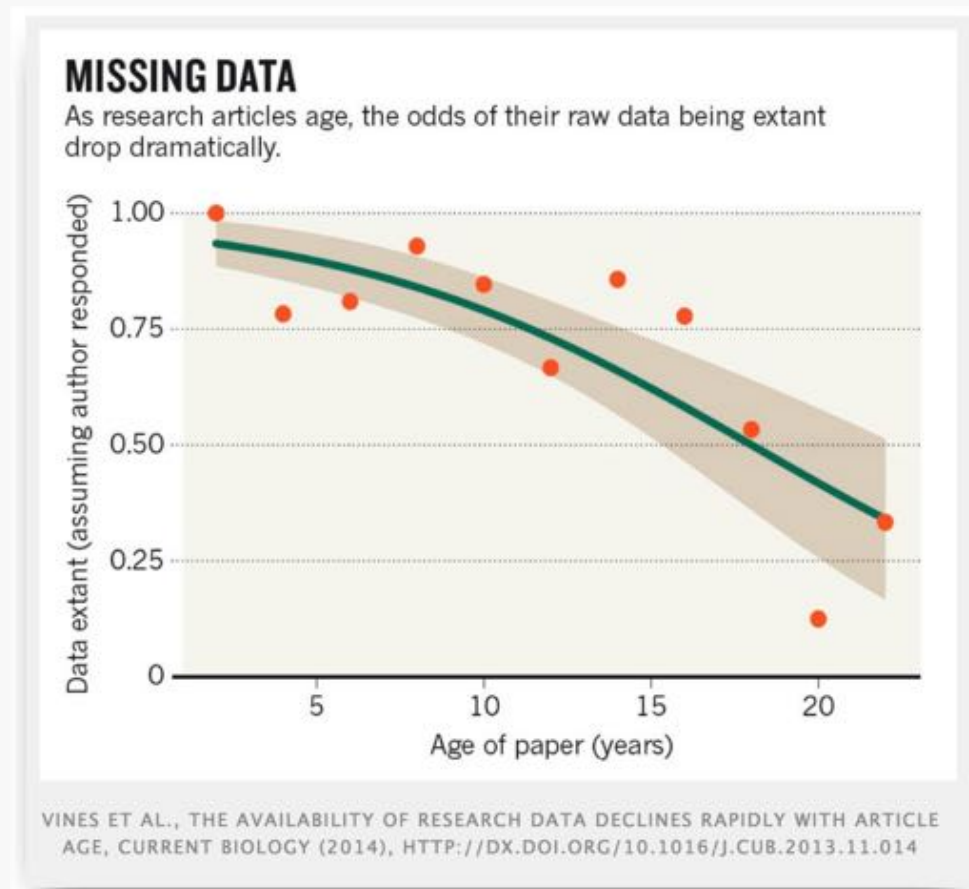
L'accès aux données affecté de manière négative par l'âge de l'article:

Existence des données: 17% réduction par année (disparues ou stockées sur des unités impossibles à ouvrir).

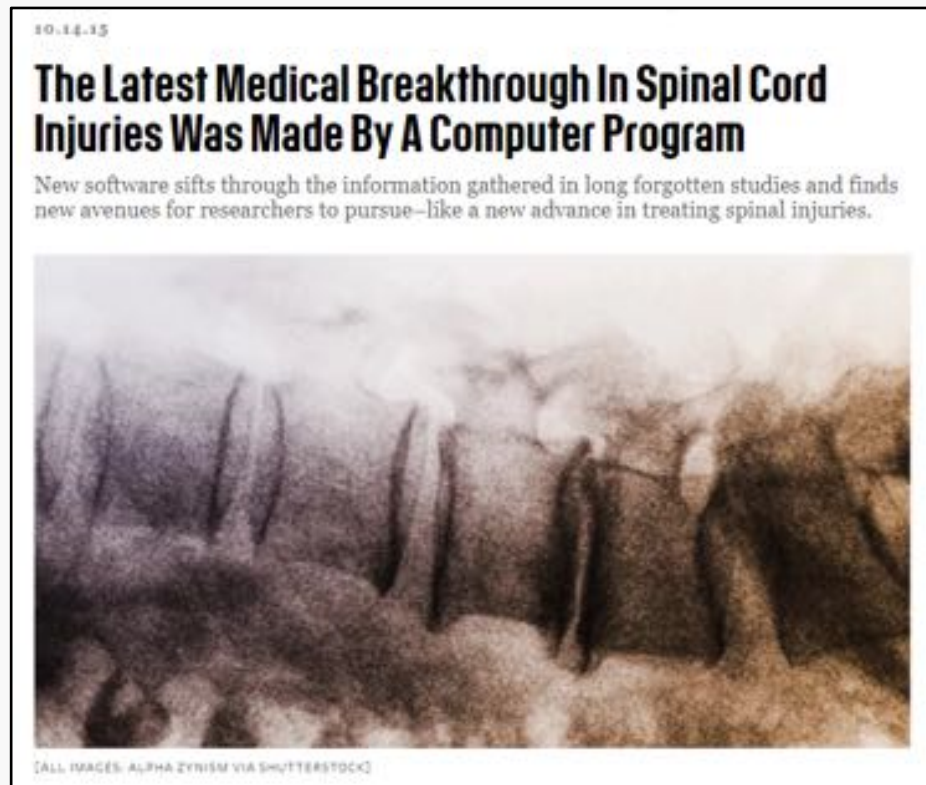
Validité de l'information de contact: 7% réduction par année.

**“Our results reinforce the notion that, in the long term, research data cannot be reliably preserved by individual researchers, and further demonstrate the urgent need for policies mandating data sharing via public archives.”**

(Vines et al., 2014, underlining ours)



# Contribution à des découvertes futures



- Méta-analyse de données, y comprise une large quantité de données non publiées, auparavant considérées inutilisables.
- Découverte d'une relation entre le rétablissement de personnes ayant eu une lésion de la moelle épinière et la tension artérielle durant les premières interventions chirurgicales.
- **“the finding raises several interesting questions [...] whether scientists should publish their raw data for posterity and whether their time and funding would be better spent poring through old experiments than conducting new ones.”**

(Lindsay, 2015, underlining ours)

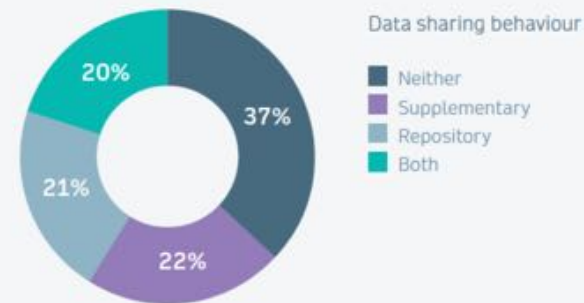


# Pratiques réelles

Plus de 30% des chercheurs ne publient pas leurs données.

Moins de 30% des chercheurs publient les données dans une banque de données.

Figure 2: Q - Generally, when submitting a manuscript to a journal what do you do with the data files generated by your research? (n=7,697)



Stuart et al, 2018 (N = 7700)

Figure 1. Dissemination of research data (% n=1162)



Wouters & Haak, 2017 (N = 1162)

# Activité 1:

## Inquiétudes concernant l'archivage ouvert de données

1. J'aimerais publier mon travail avant que d'autres personnes voient mes données.
2. Je n'ai ni le temps, ni l'argent pour pouvoir préparer mes données pour partage.
3. D'autres chercheurs ne comprendraient pas mes données.
4. Je fais des recherches très sensibles et ne peux pas permettre à d'autres personnes d'accéder mes données.
5. J'ai promis de détruire mes données dès que le projet est terminé.
6. Mes données comprennent entre autres des données que j'ai achetées, et je ne peux pas les publier.

Avec votre voisin, passez 10 minutes à discuter ces inquiétudes. Sont-elles valables? Comment promouvoir de la transparence et le libre accès et au même temps réduire l'effet de ces inquiétudes? Conclusion en plénum après.

(Corti et al. 2014, pp. 10-12)



# Exigences et attentes

## UiT Université arctique de Norvège

« The researcher shall make the research data openly available for future use by all relevant users, providing this is not prevented by any legal, ethical, security, or commercial reasons. »

## Conseil de Recherche de Norvège

« Open as standard. »

## EU, Horizon 2020

« As open as possible, as closed as necessary. »





Exigences et attentes:

# *Natural Language & Linguistic Theory*

## DATA MANAGEMENT

Natural language data is central to the mission of NLLT. Authors must ensure that their data is reliable, and must provide sufficient information on data collection in their manuscript for the editors and reviewers to evaluate this. Depending on the source of data, this may include details regarding the consultants (number, dialect/region, age, etc), the experimental design, the search procedures, or the other data collection methods used. If the data is archived, provide the archival information. If secondary sources are used, sourcing information must be provided for every example, including page numbers if the source has them. Use of tertiary sources is discouraged.

When submitting a manuscript, use the "additional information" to highlight for the editors and reviewers where in the manuscript the information on data collection can be found.

Manuscripts that do not include this information will be returned to the authors for the information to be provided before review.



Exigences et attentes:

# *Journal of French Language Studies*

## Data and Supporting Evidence

We support transparency and openness around data, code, and other materials associated with research. We expect authors to maintain accurate records of supporting evidence necessary to allow others to understand, verify, and replicate new findings, and to supply or provide access to this supporting evidence, on reasonable request. Where appropriate and where allowed by employer, funding body or others who might have an interest, we encourage authors to deposit data in a suitable repository or storage location, for sharing and further use by others. Cambridge University Press aims to provide authors with the ability to connect supporting data with their manuscripts, either on our own platform or through third party services.

Cambridge University Press is also a signatory of [Transparency and Openness Promotion \(TOP\) Guidelines](#). Please contact individual journals for information on their transparency requirements.



Research data @ UiT

<https://www.cambridge.org/core/services/aop-file-manager/file/5b44807ace5b3fca0954531e/CUP-Research-Publishing-Ethics-Guidelines-2019.pdf>

Exigences et attentes:

## *Science*

### **Data and Materials Availability after Publication**

After publication, all data and materials necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of a *Science Journal*. After publication, all reasonable requests for data, code, or materials must be fulfilled. Any restrictions on the availability of data, code, or materials, including fees and restrictions on original data obtained from other sources must be disclosed to the editors as must any Material Transfer Agreements (MTAs) pertaining to data or materials used or produced in this research, that place constraints on providing these data, code, or materials. Patents (whether applications or awards to the authors or home institutions) related to the work should also be declared.

Fossils or other rare specimens must be deposited in a public museum or repository and available for research.

Unreasonable restrictions on data, code, or material availability may preclude publication.

Problems in obtaining access to published data are taken seriously by the *Science Journals* and can be reported at [science\\_data@aaas.org](mailto:science_data@aaas.org).



# Partager les données: Quelques avantages

## **Augmenter l'impact et la visibilité de votre recherche**

Être cité

Autoriser le réusage des données pour la recherche et l'innovation  
Trouver de nouvelles collaborations

**Encourager la validation et  
l'amélioration de votre recherche (et  
des méthodes de recherche en général)**

**Assurer l'accès future à vos  
données (stockage et description)**

**Répondre aux exigences (bailleurs de  
fonds, revues) concernant le partage  
de données**

# Avant de classifier nos données ...

## **Données brutes (primaires)**

Fichiers audio/vidéo  
Transcriptions

## **Données structurées**

Données annotées  
Listes d'occurrences  
Extraits sonores

## **Métadonnées**

Informations sur le locuteur  
et la situation de collection

# Quelles données archiver en libre accès?

Archiver avec

- libre accès aux données brutes, données structurées et métadonnées.

Archiver avec

- accès fermé/restricté aux données brutes
- libre accès aux données structurées et métadonnées.

Archiver avec

- accès fermé/restricté aux données brutes et structurées
- libre accès aux métadonnées

Archiver avec

- accès fermé/restricté aux données brutes, données structurées et métadonnées

**Où se place votre propre projet? Trouvez la réponse en discutant 3 minutes avec votre voisin.**



# Ressources utiles sur la gestion de données

## Matériel de formation

FOSTER training material: <https://www.fosteropenscience.eu/>

TU Delft Open science MOOC: <https://online-learning.tudelft.nl/courses/open-science-sharing-your-research-with-the-world/>

Jisc Research Data Management toolkit: <https://rdmtoolkit.jisc.ac.uk/>

DCC How-to-guides & checklists: <http://www.dcc.ac.uk/resources/how-guides>

CESSDA Training: <https://www.cessda.eu/Training>

## Littérature

Alter, G. & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, 73(2), 146-156. <http://dx.doi.org/10.1037/amp0000258>

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., . . . Yarkoni, T. (2016). How open science helps researchers succeed. *eLife*, 5, e16800. <https://doi.org/10.7554/eLife.16800>

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021. <https://doi.org/10.1038/s41562-016-0021>

Van den Eynden, V., Corti, L., Woollard, M., Bishop, L. & Horton L. (2011). *Managing and sharing data: A best practice guide for researchers*. 3 ed. Colchester: UK Data Archive. Retrieved from <http://repository.essex.ac.uk/2156/>



# Littérature sur la gestion de données linguistiques

- **Reproducible research in linguistics:  
A position statement on data citation and attribution in our field.**

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., . . . Woodbury, A. C. (2018). *Linguistics*, 56(1), 1-18.

- **The Open Handbook of Linguistic Data Management**

Open Handbooks in Linguistics series, MIT Press Open, 2020

Editeurs: Andrea L. Berez-Kroeker, Bradley McDonnell, & Eve Koller





# Linguistics Data Interest Group (LDIG)

- Réseau international d'environ 100 membres au sein du Research Data Alliance.
- Groupe noyau d'environ 8 personnes, codirigé par Helene N. Andreassen (UiT), Andrea Berez-Kroeker (Hawai'i at Manoa) et Lauren Gawne (La Trobe).

## Objectifs

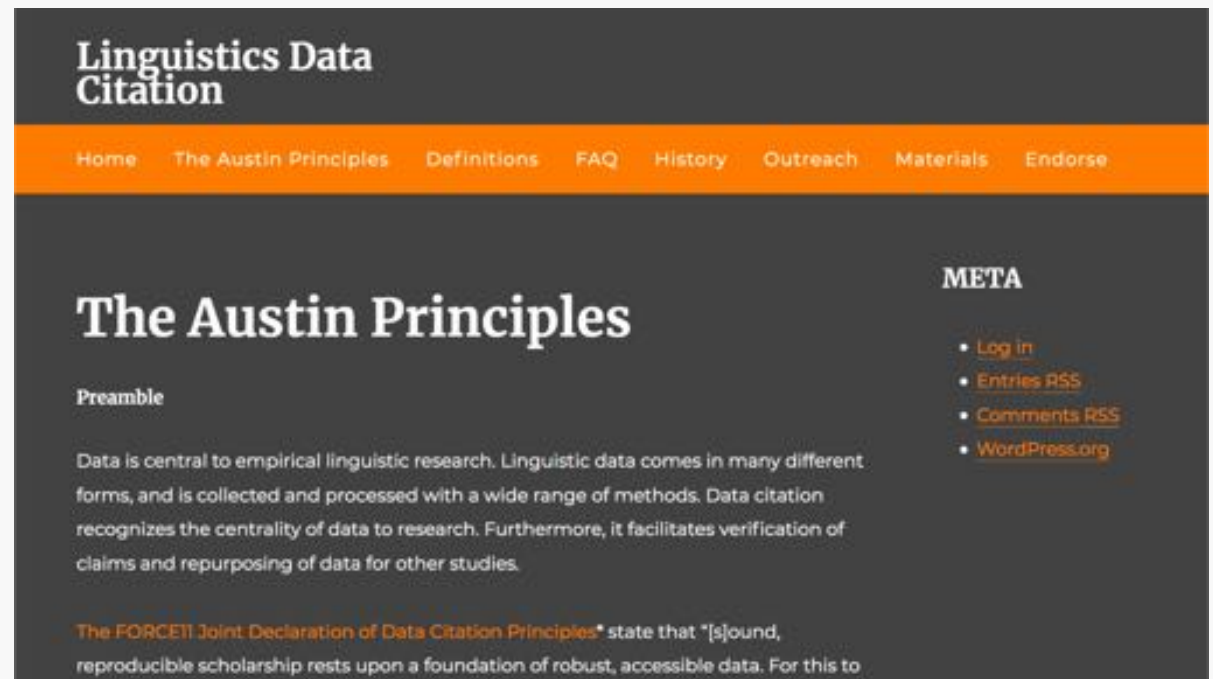
1. Développement et adoption de principes et guidelines pour la citation et l'attribution des données.
2. Education et dissémination, changement de culture.
3. Valorisation du traitement des données dans la profession

<https://www.rd-alliance.org/groups/linguistics-data-ig>

# The Austin Principles of Data Citation in Linguistics

- Encourager et améliorer la visibilité et la récupérabilité des données
- Guidelines pour le formatage de citations de données
- Applicable à toute sous-discipline ainsi que tout type de données
- Basé sur le FORCE11 Joint Declaration of Data Citation Principles)

*2020: Recommendations for citation of research data in linguistics (working title)*



The screenshot shows the website for Linguistics Data Citation. The header includes the site name and a navigation menu with links for Home, The Austin Principles, Definitions, FAQ, History, Outreach, Materials, and Endorse. The main content area features the title 'The Austin Principles' and a 'Preamble' section. The preamble text states: 'Data is central to empirical linguistic research. Linguistic data comes in many different forms, and is collected and processed with a wide range of methods. Data citation recognizes the centrality of data to research. Furthermore, it facilitates verification of claims and repurposing of data for other studies.' Below this, it mentions 'The FORCE11 Joint Declaration of Data Citation Principles\* state that "[s]ound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to'. On the right side, there is a 'META' section with links for 'Log In', 'Entries RSS', 'Comments RSS', and 'WordPress.org'.

linguisticsdatacitation.org



Research data @ UiT

# Devoir pour après ce cours

## A déterminer:

- Est-ce que vous avez des données et/ou du code à partager, en libre accès ou accès restreint?
- Est-ce que votre institution a une politique sur la gestion des données de recherche?
- Est-ce que votre institution offre une banque de données?
- Est-ce que vous connaissez les personnes s'occupant des services liés à la gestion de données dans votre institution? Suggestion: Allez dans la bibliothèque et demandez!





# Les catégories à se rappeler

## **Données brutes (primaires)**

Fichiers audio/vidéo  
Transcriptions

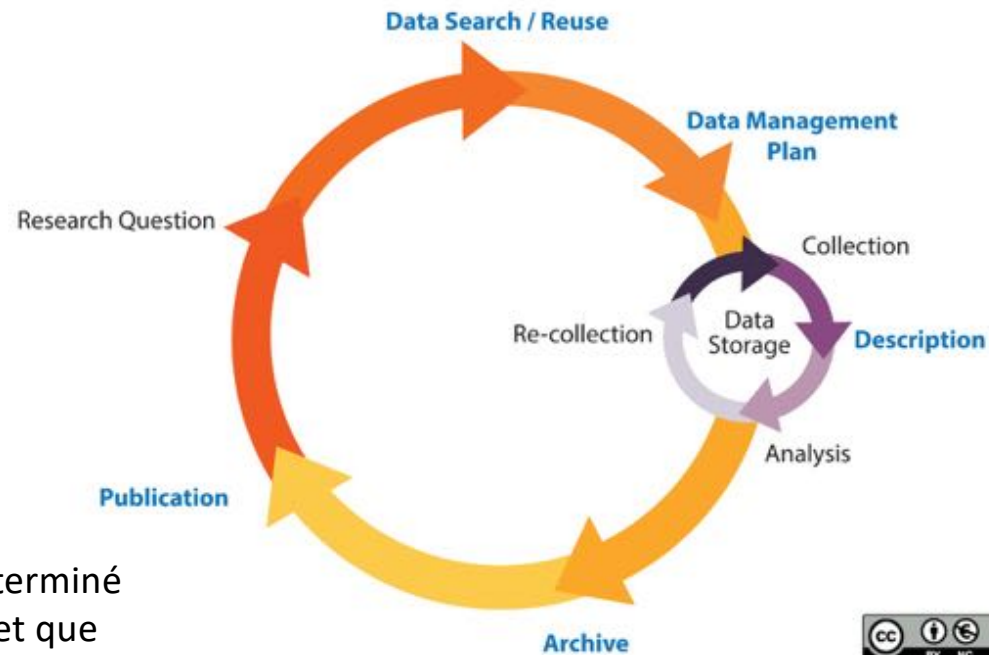
## **Données structurées**

Données annotées  
Listes d'occurrences  
Extraits sonores

## **Métadonnées**

Informations sur le locuteur  
et la situation de collection

## The Research Data Management Lifecycle

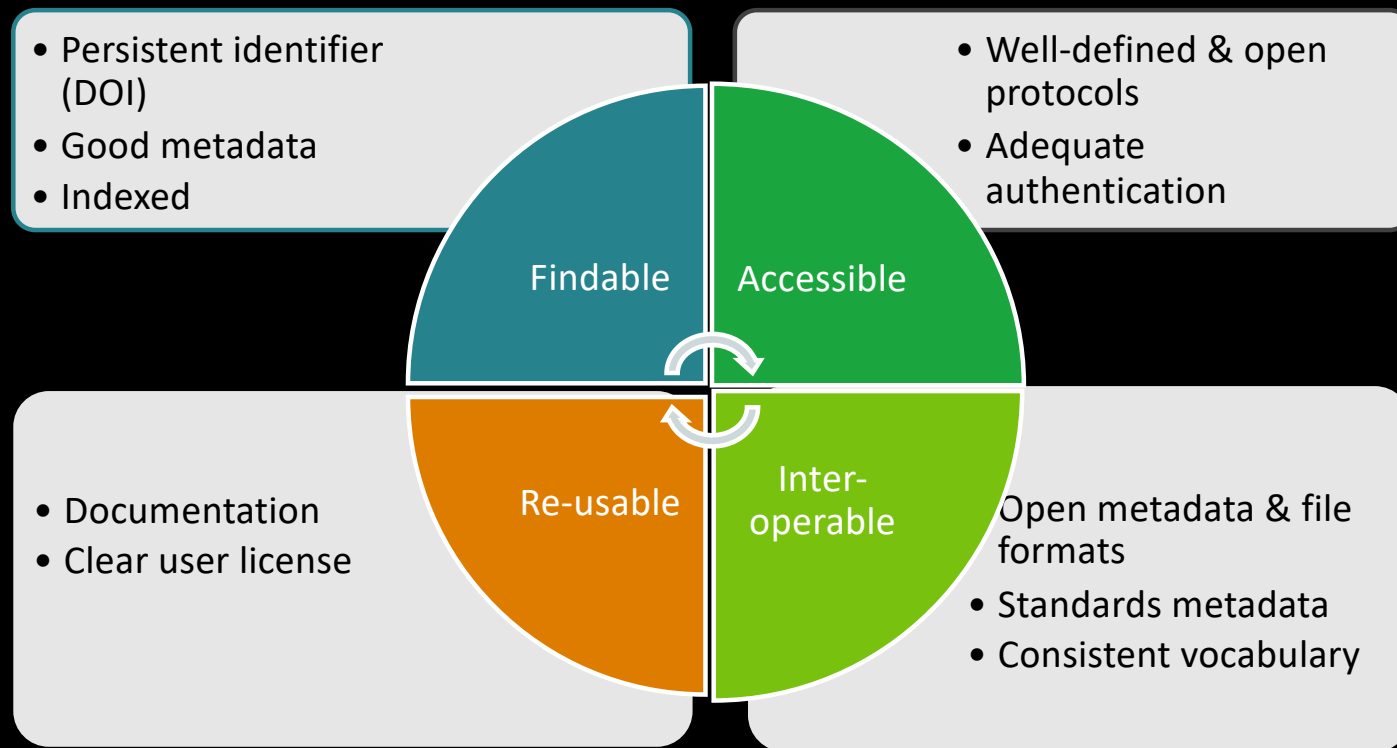


**1. La phase de planification**

**2. La phase active,** quand vous traitez et analysez vos données, et que vous travaillez sur votre papier.

**3. La phase d'archivage et de publication,** quand vous avez terminé le traitement de vos données et que vous êtes prêt(e) à soumettre le papier. Dépendant de la nature des données, vous les archivez en libre accès ou avec un accès restreint.

# Les meilleures pratiques et les principes FAIR



# La phase de planification

Etablir la question de recherche

Identifier de données existantes

Préparer la gestion de données, y comprise la collecte

Préparer les documents de consentement





# Rechercher des données collectées pour d'autres projets

Si la licence le permet, réutiliser les données pour votre propre projet.

Si collectées pour un projet similaire et qu'il y a déjà des réponses, modifier votre question de recherche?

L'examen des données permet d'identifier le besoin d'ajuster la méthode prévue pour la collecte de données: Qu'est-ce qui a marché/pas marché?

## MOTEUR DE RECHERCHE

DataCite  
BASE – Bielefeld Academic Search Engine  
Google Dataset Search

## BANQUES DE DONNEES

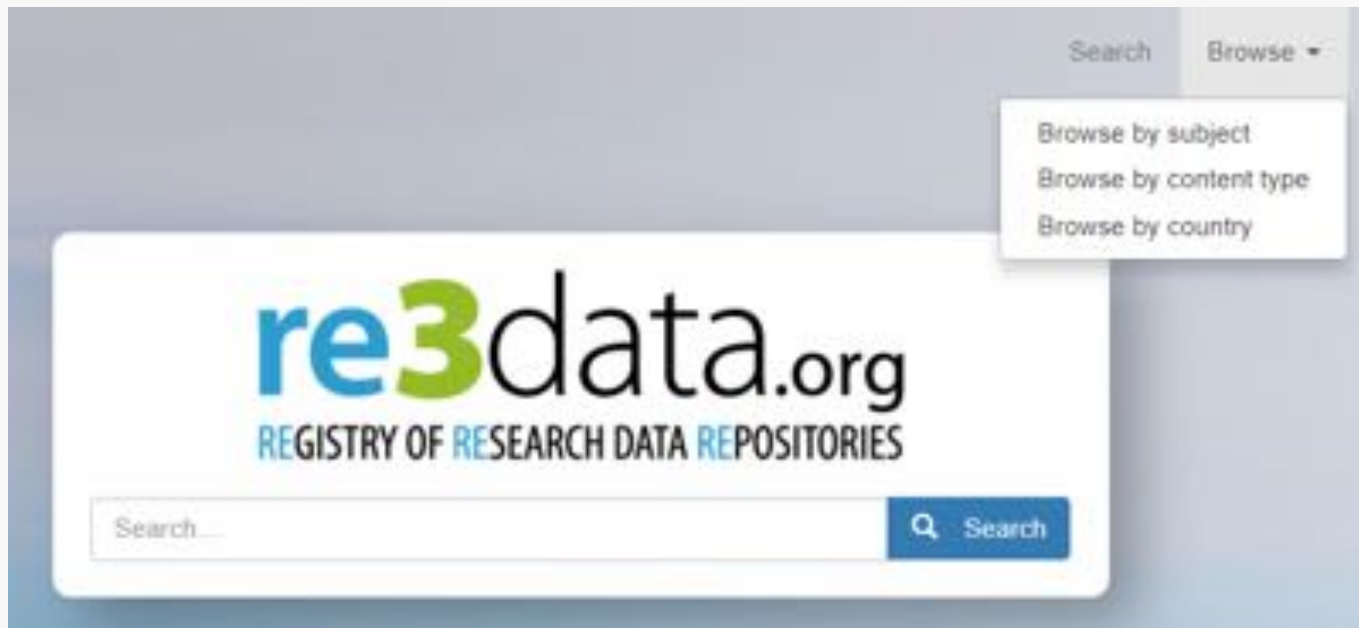
Figshare  
Dryad  
TROLLing  
Zenodo

## INDEX DE BANQUES DE DONNEES

Re3data



# L'index de banques de données



Found 93 result(s)

## Ortolang

Outils et Ressources pour un Traitement Optimisé de la LANGue

Subject(s) [Linguistics](#) [Humanities](#) [Humanities and Social Sciences](#)

Content type(s) [Standard office documents](#) [Structured text](#) [Audiovisual data](#) [Plain text](#)

Country [France](#) [European Union](#) [Search with content type 'Audiovisual data'](#)

ORTOLANG is an EQUIPEX project accepted in February 2012 in the framework of investissements d'avenir. Its aim is to construct a network infrastructure including a repository of language data (corpora, lexicons, dictionaries etc.) and readily available, well-documented tools for its processing. Expected outcomes comprise: promoting research on analysis, modelling and automatic processing of our language to their highest international levels thanks to effective resource pooling; facilitating the use and transfer of resources and tools set up within public laboratories to industrial partners, notably SMEs which often cannot develop such resources and tools for language processing given the cost of investment; promoting French language and the regional languages of France by sharing expertise acquired by public laboratories. ORTOLANG is a service for the language, which is complementary to the service offered by Huma-Num (très grande infrastructure de recherche). Ortolang gives access to SLDR for speech, and CNRTL for text resources.

## Das Deutsche Referenzkorpus

DeReKo

Subject(s) [Linguistics](#) [Humanities](#) [Humanities and Social Sciences](#)

Content type(s) [Standard office documents](#) [Audiovisual data](#) [Scientific and statistical data formats](#) [Raw data](#) [Plain text](#)

Country [Germany](#) [European Union](#)

The project is set up in order to improve the infrastructure for text-based linguistic research and development by building a huge, automatic corresponding tools for corpus annotation and exploitation. DeReKo constitutes the largest linguistically motivated collection of contemporary newspaper texts, as well as several other text types, contains only licensed texts, is encoded with rich meta-textual information, is fully annotated and

Repository details

## Ortolang

[General](#) [Institutions](#) [Terms](#) [Standards](#)

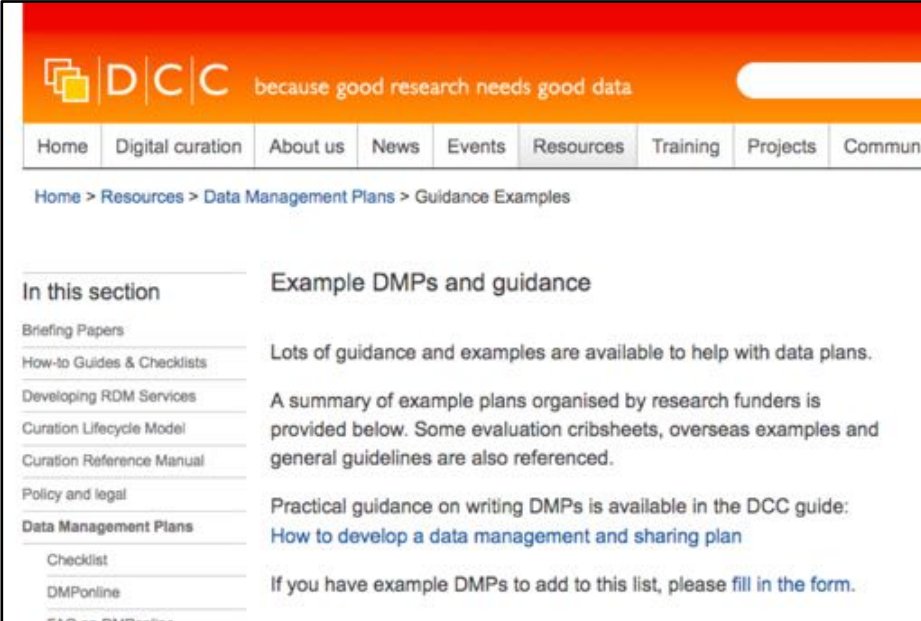
|                    |  |
|--------------------|--|
| Name of repository | Ortolang   |
| Additional name(s) | Outils et Ressources pour un Traitement Optimisé de la LANGue<br>Open Resources and TOols for LanGuage   |
| Repository URL     | <a href="https://www.ortolang.fr/">https://www.ortolang.fr/</a>  |
| Subject(s)         | <a href="#">Linguistics</a> <a href="#">Humanities</a> <a href="#">Humanities and Social Sciences</a>  |
| Description        | ORTOLANG is an EQUIPEX project accepted in February 2012 in the framework of investissements d'avenir. Its aim is to construct a network infrastructure including a repository of language data (corpora, lexicons, dictionaries etc.) and readily available, well-documented tools for its processing. Expected outcomes comprise: promoting research on analysis, modelling and automatic processing of our language to their highest international levels thanks to effective resource pooling; facilitating the use and transfer of resources and tools set up within public laboratories to industrial partners, notably SMEs which often cannot develop such resources and tools for language processing given the cost of investment; promoting French language and the regional languages of France by sharing expertise acquired by public laboratories. ORTOLANG is a service for the language, which is complementary to the service offered by Huma-Num (très grande infrastructure de recherche). Ortolang gives access to SLDR for speech, and CNRTL for text resources. |
| Contact            | <a href="mailto:contact@ortolang.fr">contact@ortolang.fr</a>   |
| Content type(s)    | <a href="#">Standard office documents</a> <a href="#">Structured text</a> <a href="#">Audiovisual data</a> <a href="#">Plain text</a>  |
| Keyword(s)         | <a href="#">corpora</a> <a href="#">lexicon</a> <a href="#">dictionary</a> <a href="#">morphology</a> <a href="#">etymology</a>  |



Research data @ UiT

# Le plan de gestion des données

- Le plan de gestion des données (*data management plan*, DMP) sert à documenter la gestion des données pendant et après le projet.
- Couvre toutes les phases du cycle de vie des données.
- Créé avant/au début du projet, normalement révisé au cours du projet.



The screenshot shows the DCC website interface. At the top, the logo 'D|C|C' is displayed with the tagline 'because good research needs good data'. Below the logo is a navigation menu with links for Home, Digital curation, About us, News, Events, Resources, Training, Projects, and Commun. The main content area is titled 'Home > Resources > Data Management Plans > Guidance Examples'. It features a sidebar on the left with a list of links under the heading 'In this section', including Briefing Papers, How-to Guides & Checklists, Developing RDM Services, Curation Lifecycle Model, Curation Reference Manual, Policy and legal, and Data Management Plans. The main content area is titled 'Example DMPs and guidance' and contains three paragraphs of text. The first paragraph states that lots of guidance and examples are available to help with data plans. The second paragraph mentions a summary of example plans organized by research funders, including evaluation cribsheets, overseas examples, and general guidelines. The third paragraph provides practical guidance on writing DMPs, referencing a guide titled 'How to develop a data management and sharing plan'. Below the text, there is a link to 'Checklist' and 'DMPonline', and a note that if users have example DMPs to add to the list, they should fill in a form.



# L'objectif du plan de gestion des données

- Aide à garder le contrôle des données pendant le projet.
- Aide à gagner du temps et à éviter du double travail.
- De plus en plus demandé par des institutions et des bailleurs de fonds.
- P.ex. demandé par le Fonds National Suisse à partir d'octobre 2017:

*Les données et résultats de la recherche encouragée par des fonds publics doivent être accessibles à tous. Le Fonds national suisse (FNS) soutient le principe du libre accès aux données de la recherche (Open Research Data ou ORD). Il introduit à ce titre une première mesure: **à partir d'octobre 2017, des plans de gestion des données (DMP) feront partie intégrante des requêtes en encouragement de projets.***

***Les requérant-e-s soumettent un DMP dont la pertinence est vérifiable et qui correspond aux besoins de leur projet et à ceux de leur communauté de recherche.** A ce stade, le DMP est considéré comme une ébauche et n'est pas pris en compte dans l'évaluation de la requête. Si la requête est acceptée par le FNS, le DMP peut être adapté, modifié ou élargi durant toute la durée du subside. Un DMP définitif doit être fourni au plus tard lors de l'achèvement du projet de recherche.*




# Le contenu du plan de gestion des données

- Information générale sur le projet
  - Responsabilités et droits
1. Collecte et traitement des données
  2. Documentation et métadonnées
  3. Stockage et backups pendant le projet
  4. Archivage et partage



# Le contenu du plan de gestion des données

- Ethique et protection des données
  - Formulaire de consentement
  - Evaluation de la probabilité de collecte de données avec de l'information personnelles
  - Identification de la nécessité de faire une demande auprès d'une commission éthique.



Avoir toutes les phases futures à l'esprit en faisant ce travail.

*[T]he ideal stance [...] would be to involve participants in the process of data sharing. Given the extended interactions qualitative researchers typically have with their interviewees, they are able to explain the nature, purpose and benefits of data sharing and also get a sense of the types of risks their participants might be worried about (or not) and thus calibrate an initial hypothetical assessment. (Kirilova and Karcher 2017, p. 3, underlining ours)*

# Règlement général sur la protection des données

## Ce que c'est le RGPD (ang. GDPR)

- « The GDPR applies to personal data, meaning any information relating to an identifiable person who can be directly or indirectly identified »
- Entré en vigueur en mai 2018 et s'applique à tous les Etats Membres de l'UE
- Principes clés
  - le consentement
  - le droit des personnes (y compris minimisation de données)
  - la transparence
  - la responsabilité

## Comment y répondre

- Discutez des défis potentiels avec votre directeur/-trice de thèse
- Identifiez les personnes administratives dans votre institution pouvant vous donner des conseils, si besoin.

<https://eugdpr.org/>





# Traitement de données avec des informations personnelles

- Si possible, anonymiser pendant la transcription. Si vous le mettez pour plus tard, vous risquez de ne pas avoir le temps de le faire, et au pire des cas, vous devez détruire les données brutes.
- Stocker la clé d'embrouillage, qui lie les données et la personne, séparément. La détruire quand l'anonymisation est complétée.
- Conseils pour l'anonymisation: Corti et al., 2014, p. 118-124.
- Explications de terminologie:  
<http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html>



# Des données personnelles: Points importants

- Avez-vous vraiment besoin de collecter des données personnelles?
  - Si oui, les stocker sur une entité différente dans la phase active.
- Informez bien vos locuteurs comment vous allez traiter les données personnelles.
  - Stockage, usage, destruction
- Une donnée provenant d'un locuteur ne contient pas nécessairement d'information personnelle, p.ex. si anonyme/anonymisée.
  - A déterminer: des identifiants directs, p.ex. nom, adresse, numéro de téléphone, photos, et des identifiants indirects, p.ex. lieu de travail, occupation, âge (peut-être problématiques si combinés).



# La phase active

Collecter de nouvelles données

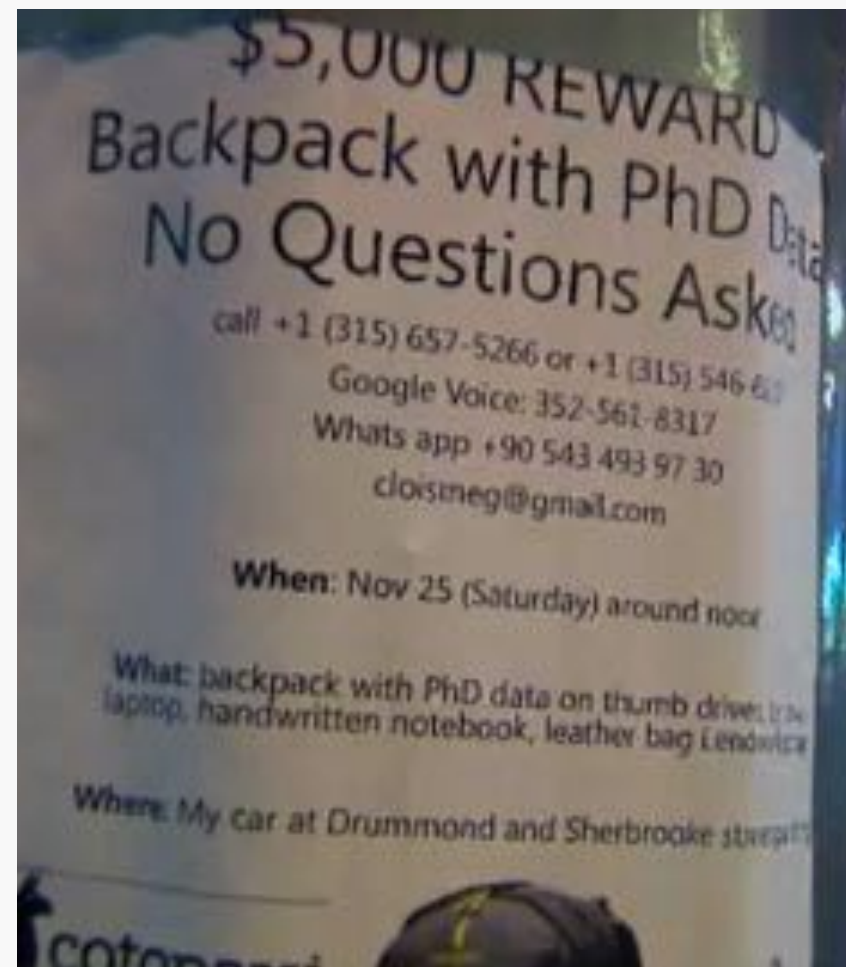
Collecter et créer des métadonnées

Stocker, nettoyer, transcrire, anonymiser et valider les données

Décrire et documenter les données

Analyser et interpréter les données, préparer la publication





<https://montreal.ctvnews.ca/phd-student-offering-5-000-reward-after-car-thief-steals-all-his-research-1.3700484>



Research data @ UiT

# Stockage des données

## Collecte

- Utiliser un laptop.
- Transfer: Stocker des enregistrements sur un dispositif USB encrypté.

## Stockage/traitement

- L'endroit de stockage dépend de la sensibilité des données.
- Demander les possibilités de stockage à votre département d'informatique.

| Catégorie              | Impact   |
|------------------------|--|
| Niveau de protection 2 | Elevé: Information extrêmement sensible, qui peut identifier l'individu.                   |
| Niveau de protection 1 | Modéré: Information moyennement sensible, qui peut identifier l'individu.                  |
| Niveau de protection 0 | Faible: Information non-sensible, qui peut identifier l'individu, ou information publique. |

<https://security.berkeley.edu/education-awareness/cybersecurity-best-practices/how-classify-research-data>



# Structure et documentation



Mots clés: format des fichiers – appellation des fichiers –  
description des fichiers

# Format des fichiers

**Des formats pérennes assure l'accès aux (et le réusage des) données dans l'avenir.**

- Non propriétaire
- Ouvert
- Utilisé par la communauté de recherche
- Usage de codage de caractères standard (e.g. UTF-8)
- Non compressé



# Formats pérennes (exemples)

| Type de document | Format pérenne (exemples)                      | Format non pérenne (exemples) |
|------------------|--|-------------------------------|
| Texte            | Texte brut (.txt), PDF/A                       | MS Word (.docx)               |
| Tableur          | Tabulator-separated Unicode UTF-8 texte (.txt) | MS Excel (.xlsx)              |
| Image            | TIFF non compressé                             | Windows Bitmap (.bmp)         |
| Son              | WAV  | AAC (.m4a)                    |
| Film             | MPEG-4   | Quicktime (.mov)              |

Cf. e.g. UiT Open Research Data Deposit Guide pour plus d'information:  
<https://site.uit.no/dataverseno/deposit/prepare/>





# Appellation des fichiers

## CONSISTENT & DESCRIPTIF

interview\_Lausanne\_2016-05-01\_notes.txt  
interview\_Nyon\_2016-05-01\_notes.txt  
questionnaire\_Lausanne\_2016-05-01\_notes.txt  
questionnaire\_Nyon\_2016-05-01\_notes.txt

(pourquoi problématique?)

Notes terrain Mai nouveau 171017 Helene



## Retraction Watch

Tracking retractions

### Doing the right thing: Authors retract brain paper with “systematic human error in coding”

with one comment

A group of Swiss neurologists have lost their 2013 article in *Frontiers in Human Neuroscience* after reporting that their data were rendered null by coding errors.

frontiers in  
HUMAN NEUROSCIENCE

2.0

“..a systematic human error in coding the name of the files had been made during the extraction of the EEG template topographic maps best differentiating the two experimental conditions at the single subject level.”

<http://retractionwatch.com/2014/01/07/doing-the-right-thing-authors-retract-brain-paper-with-systematic-human-error-in-coding/>



Research data @ UiT

# Appellation des fichiers

## DATE

2016\_interview\_Loc1  
2016\_questionnaire\_Loc1  
2018\_interview\_Loc1  
2018\_questionnaire\_Loc1

## TYPE

Interview\_Loc1\_2016  
Interview\_Loc2\_2016  
Questionnaire\_Loc1\_2016  
Questionnaire\_Loc2\_2016

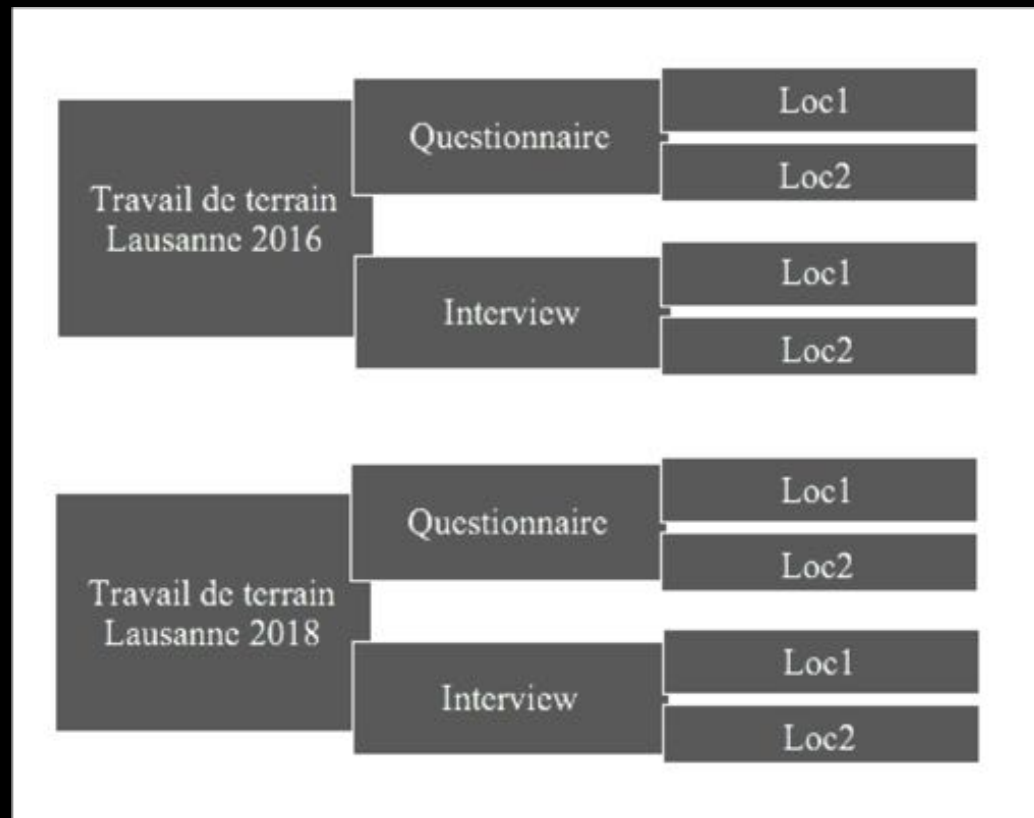
## THÈME OU LOCUTEUR

Loc1\_interview\_2016  
Loc1\_questionnaire\_2016  
Loc2\_interview\_2016  
Loc2\_questionnaire\_2016

## ORDRE FORCÉ

01\_Loc1\_questionnaire\_2016  
02\_Loc1\_interview\_2016  
03\_Loc1\_questionnaire\_2018  
04\_Loc2\_interview\_2018

# Structuration des dossiers



# Le fichier *readme*: la description de vos données

De quoi s'agissent les données?

Vue d'ensemble des fichiers

Méthode (conditions de collecte et de traitement des données, p.ex. enregistrement, transcription, analyse)

Structure des fichiers et conventions d'appellation

En-têtes de colonne dans des tableaux

Abréviations

...



Research data @ UiT

README file for replication data for "Less is More: Why All Paradigms are Defective, and Why that is a Good Thing"

by Laura A. Janda and Francis M. Tyers

This post contains five csv datasets of Russian nouns, plus an R script for their analysis.

The five csv datasets are the following:

procent-I-m.aa.csv This is data on masculine animate I-declension nouns.  
procent-I-m.nn.csv This is data on masculine inanimate I-declension nouns.  
procent-I-nt.nn.csv This is data on neuter I-declension nouns.  
procent-II-f.nn.csv This is data on feminine inanimate II-declension nouns.  
procent-III-f.nn.csv This is data on feminine inanimate III-declension nouns.

This data comes from the SynTagRus corpus

([https://github.com/UniversalDependencies/UD\\_Russian-SynTagRus](https://github.com/UniversalDependencies/UD_Russian-SynTagRus)).

Each dataset has the same structure.

Column 1 "freq" lists the frequency of the lemma in the corpus.

Column 2 "lemma" lists the lemma in question (in Cyrillic).

Column 3 "gramm" lists the type of noun and is the same throughout each file. For the file procent-I-m.aa.csv, for example, all items are marked "Masc.Anim"

Column 4 "total" is the total frequency and is identical to column 2.

Columns 5 through 16 give the relative frequency (percent) of attestations for each case/number combination for this lemma. sg=singular, pl=plural, nom=nominative, gen=genitive, dat=dative, acc=accusative, ins=instrumental, loc=locative.

For example, the first row of the dataset procent-I-m.aa.csv begins like this:

| freq | lemma   | gramm     | total | sg.nom | sg.gen |
|------|---------|-----------|-------|--------|--------|
| 2651 | человек | Masc.Anim | 2651  | 14.79  | 14.22  |

This means that 2651 forms of the word человек 'person' appear in the corpus, and that 14.79% of them are Nominative Singular forms, 14.22% are Genitive Singular forms, etc.

The R script shows the code needed to read these files into R and perform the correspondence analysis.

# Le fichier *readme*: la description de vos données

Créer un log au fur et à mesure, avec toute décision prise, facilitera le développement du fichier *readme*.

This archive contains files containing data and analysis from

Project: ES/K006495/1

**Marriage Migration and Integration**

The following files have been archived:

| File name   | Description  |
|---|--|
| Interview Transcripts                             | Folder containing interview transcripts, in six sub folders by category of participant:<br>BRITISH PAK MARRIED INTRANATIONALLY<br>BRITISH PAK MARRIED TRANSNATIONALLY<br>PAKISTANI MIGRANT SPOUSES<br>BRITISH SIKHS MARRIED INTRANATIONALLY<br>BRITISH SIKHS MARRIED TRANSNATIONALLY<br>SIKH MIGRANT SPOUSES |
| Guide to the data archived                        | This file provides background on the project, an overview of the data archived, including a list of interview transcripts with a key for transcript titles   |
| Interview Schedule                                | This file contains the interview schedule used for the semi-structured interviews  |
| Consent Form                                      | Consent form for interview participants  |
| Marriage Migration and Integration Project Report | Project Report providing overview of findings from the project as a whole, and more information on the project including methods and sampling.   |

The project report can also be found at: <http://www.bristol.ac.uk/ethnicity/projects/mmi/>



# Le fichier *readme*: la description de vos données

Le type de banque de données et les champs de métadonnées qui y sont disponibles aide à déterminer ce qui doit aller dans le fichier *readme*.

## USER GUIDE TO ESDS DATA DEPOSIT

Project title: **Social networks, target language interaction and second language acquisition during the year abroad: a longitudinal study**

ESRC award number: RES-062-23-2996

Principal Investigator/ contact for inquiries: Professor Rosamond Mitchell, Modern Languages, University of Southampton, Southampton SO17 1BJ, UK. Email: [R.F.Mitchell@soton.ac.uk](mailto:R.F.Mitchell@soton.ac.uk)

Subject area: Linguistics

## GENERAL DESCRIPTION OF DATA

These data comprise a corpus of advanced learner French and advanced learner Spanish and will be of interest to researchers into second language acquisition for these two languages.

The participants were Anglophone university students majoring in languages at a UK university (29 for French, 27 for Spanish). The participants completed a battery of assessments on 6 occasions over a period of 18 months, before, during and after spending a compulsory academic year abroad, in France, Spain or Mexico. The data collection schedule was as follows:

|                                |               |
|--------------------------------|---------------|
| Pre-test (Letter code a)       | May 2011      |
| Visit 1 abroad (Letter code b) | November 2011 |
| Visit 2 abroad (Letter code c) | February 2012 |
| Visit 3 abroad (Letter code d) | June 2012     |
| Post-test 1 (Letter code e)    | October 2012  |
| Post-test 2 (Letter code f)    | February 2013 |

The data being made available to other researchers comprises learner outputs in French or Spanish, collected on all 6 occasions, using a set of 3 tasks:

### 1. Oral interview

Participants took part in a semi-structured interview in French or Spanish (Task O) on each occasion, with a member of the research team. The interviews varied in detail on each occasion but all centred on participants' hopes, plans and actual experiences of residence abroad. (Interview schedules are available by request to the research team.)



# Activité 2: Rendre nos données réutilisables

(voir polycopié)





# Activité 2: Rendre nos données réutilisables

- Nom du dossier : Caen classe F
- Fichiers dans le dossier :
  - Notes Dec2018.docx
  - Notes Nov2018\_nouveau.docx
  - 2018 octobre avec notes.docx
  - Protocole d'enquête original.doc
  - Données\_classeF.xlsx
  - Données\_classeF\_traité.nvp
  - Données\_classeF\_traité\_V2.nvp
  - Données\_classeE\_traité\_avec\_classeF\_ÅseØstmo.nvp



# Activité 2: Rendre nos données réutilisables

Extrait du contenu du fichier Données\_classeF.xlsx :

|        | Groupe | Lng | Date        | Score |
|--------|--------|-----|-------------|-------|
| Caen-1 | 1      | Ang | 8 Nov 2018  | 0.25  |
| Caen-2 | 1      | Ang | 8 Nov 2018  | 0.75  |
| Caen-3 | 1      | Ang | 8 Nov 2018  | 1.25  |
| Caen-1 | 1      | Fra | Nov 17 2018 | 1.75  |
| Caen-1 | 2      | Esp | 2018-11-22  | 2.25  |
| Caen-2 | 2      | Esp | 2018-11-22  | 2.75  |



# La phase d'archivage et de publication

Sélectionner une licence

Déterminer le niveau d'accès aux données

Archiver les données, les promouvoir (p.ex. à travers une citation dans la publication)

Réutiliser les données dans d'autres projets



# La sélection d'une banque de données



# Conseil 1

Sélectionner une banque de données dédiée à la linguistique.

- Equipe avec expertise dans le domaine.
- Services qui peuvent plus facilement améliorer la valeur de vos données.
- Exemple: COCOON, qui accueille également des données avec accès restreint.

<https://cocoon.humanum.fr/exist/crdo/>

The screenshot displays the COCOON website interface. At the top, the logo 'COCOON' is visible, along with navigation links: RECHERCHER, DÉPOSER, DOCUMENTATION, and À PROPOS. A search bar contains the text 'Corpus'. The main content area is titled 'Notice de collection' and features the heading 'Enquête en Corse'. Below this, it lists the depositor (Dalbera-Stefanaggi, Marie-José) and the interviewer (Tognotti, Aurelia Ghjacumina). A metadata table follows, detailing the collection's title, subject, language, rights, and identifiers. On the right side, there are sections for 'Partage' (social media icons), 'Téléchargement' (a 'Télécharger la collection' button), 'Lien(s) de navigation' (a list of 40 recordings), 'Métadonnées' (Linked Open Data and OLAC Dublin-core), and 'Statistiques' (showing 3 consultations).



# Conseil 2

Sélectionner une banque de données qui travaille selon les principes FAIR.

- Métadonnées
- Identifiants pérennes
- Exigences format de fichiers
- Contrôle de versions
- Lien à d'autres publications
  
- Exemple: TROLLing, qui contient uniquement des données structurées et ouvertes. (souvent des « replication data for »)



The screenshot shows a DataVerse dataset page. At the top, there is a logo for TROLLing and the breadcrumb path: DataVerseNO > UiT Open Research Data > TROLLing (The Tromsø Repository of Language and Linguistics) > Replication Data for: Les expressions spatiales en français médiéval: particules et formes préfixées en de-. Below the breadcrumb, there are buttons for Metrics, Downloads (8), Contact, and Share. The main title is 'Replication Data for: Les expressions spatiales en français médiéval: particules et formes préfixées en de-' with a version indicator 'Version 1.0'. Below the title, there is a citation: Rainford, Thomas, 2019, "Replication Data for: Les expressions spatiales en français médiéval: particules et formes préfixées en de-", <https://doi.org/10.18710/X5Z9XZ>, DataVerseNO, V1. There is a 'Cite Dataset' button and a link to 'Learn about Data Citation Standards'. The 'Description' section contains a detailed text about the dataset's content and the publication it replicates. The 'Subject' is 'Arts and Humanities', the 'Keyword' is 'Old French, prepositions, adverbs, locatives, particles', and the 'Related Publication' is cited. At the bottom, there are tabs for 'Files', 'Metadata', 'Terms', and 'Versions'. A search bar is present with the text 'Search this dataset...' and a 'Find' button. Below the search bar, it says '1 to 10 of 15 Files'. A list of files is shown, including '00\_README.txt' (Plain Text - 10.0 KB - Feb 8, 2019 - 4 Downloads) and 'arriere.txt'. Each file has a 'Download' button.



Research data @ UiT

<https://dataverse.no>  
[/dataverse/trolling](https://dataverse/trolling)

# La sélection d'une banque de données

Questions à poser, indépendamment de la nature des données et des contraintes sur l'accès aux données.

1. Est-ce qu'elle a une bonne réputation?
2. Est-ce qu'elle accepte le type de données que vous avez?
3. Est-ce que les données seront en sécurité au niveau légal?
4. Est-ce qu'elle va soutenir la valeur des données?
5. Est-ce qu'elle va faciliter l'analyse de l'usage des données?

(Whyte, 2015)



# Archivage: Points importants

- Métadonnées
  - Entrez le plus de métadonnées possible.
  - N'oubliez pas de mentionner des logiciels nécessaires pour lire & analyser les données (p.ex. Praat, Dolmen). Publier des scripts.
- Licences
  - Cherchez à comprendre la licence proposée par la banque de données.
  - <https://creativecommons.org/licenses/>
  - Pour plus d'info: [How to select the appropriate license for research data](#)
- Partager les données avant publication
  - Quand vous soumettez un papier, envoyez un lien à celles-ci dans les cas où elles ne sont pas encore publiées (TROLLing: Private URL).
  - Citez les données dans le manuscrit.





# La citation des données

## Modèle d'une référence bibliographique:

Auteur, date, titre, éditeur, localisateur (version, date d'accès, étiquette)

Kamoen, Naomi, Bregje Holleman, Pim Mak, Ted Sanders & Huub Van den Bergh. 2017. *Why are negative questions difficult to answer? On the processing of linguistic contrasts in surveys*. DataverseNL. <https://hdl.handle.net/10411/20857>. V5, dataset.

Hauk, Bryn. 2016-2018. *Batsbi (Tsova-Tush)*. Kaipuleohone University of Hawai'i Digital Language Archive. <http://hdl.handle.net/10125/42581>. (accessed 2019-06-21).

Hauk, Bryn (researcher, depositor), Omar P'ap'ashvili (speaker), Rezo Orbetishvili (consultant). 2018. BH2-074. In: *Batsbi (Tsova-Tush)*. Kaipuleohone University of Hawaii Digital Language Archive. <http://hdl.handle.net/10125/58935>.



# La citation des données

Le formatage détaillé des citations varie d'une revue à une autre.

Si l'archive propose un format de référence, gardez tous ces éléments et ajoutez ce qui manque.



 **Replication Data for: Les expressions spatiales en français médiéval: particules et formes préfixées en de-** Version 1.0

Rainsford, Thomas, 2019, "Replication Data for: Les expressions spatiales en français médiéval: particules et formes préfixées en de-", <https://doi.org/10.18710/X5ZFXZ>, DataverseNO, V1

[Learn about Data C](#)



**Citation** X

Tognotti, Aurelia Ghjacumina (interviewer); Dalbera-Stefanaggi, Marie-José (depositor). Editeur: Lieux, Identités, eSpaces, Activités. *Enquête en Corse*. 2008. Collection. Récupéré sur la plateforme COCOON, <<http://purl.org/poi/crdo.vjf.cnrs.fr/cocoon-040a0841-6ea2-35bf-bdbd-aa660385c8c0>>. (Accessed July 14, 2019.)

Collection



# La citation des données

## Modèle d'une citation dans le texte:

Auteur, date (locateur, horodatage/numéro de ligne, autre attribution (rôle))

(Hauk 2018)

(Hauk 2018, BH2-076)

(Hauk 2018, <http://hdl.handle.net/10125/58937>)

(Hauk 2018, BH2-081, 00:00:01–00:00:03, Rezo Orbetishvili (Speaker))

Basé sur:

Linguistics Data Interest Group. (à paraître). *Recommendations for citation of research data in linguistics*,

[https://docs.google.com/document/d/1\\_r2D9ReMTin3\\_qUlfEO48z5X54otTX33m9Sujso68tl/edit#heading=h.vew8lijfs9wn](https://docs.google.com/document/d/1_r2D9ReMTin3_qUlfEO48z5X54otTX33m9Sujso68tl/edit#heading=h.vew8lijfs9wn)



# Activité 3: La citation des données

1. Allez sur <https://public.projet-pfc.net/> et chercher le point d'enquête RCA Bangui. Proposer une référence bibliographique et justifier vos choix.
2. Sélectionner un des fichiers de ce point d'enquête et proposer une citation dans le texte. Justifier vos choix.



# Activité 3: La citation des données

## 1. Référence bibliographique

Modèle: Auteur, date, titre, éditeur, localisateur (version, date d'accès, étiquette)

**Lyche, C. (responsable) & Steien, G. B. (travail de terrain, transcription, codage). 2008 (date de collecte). *Enquête RCA Bangui*, Phonologie du français contemporain, [www.projet-pfc.net](http://www.projet-pfc.net), accédé 2019-07-22.**

## 2. Citation dans le texte

Modèle: Auteur, date (locateur, horodatage/numéro de ligne, autre attribution (rôle))

**Lyche & Steien, 2008, rcark1, liste de mots**

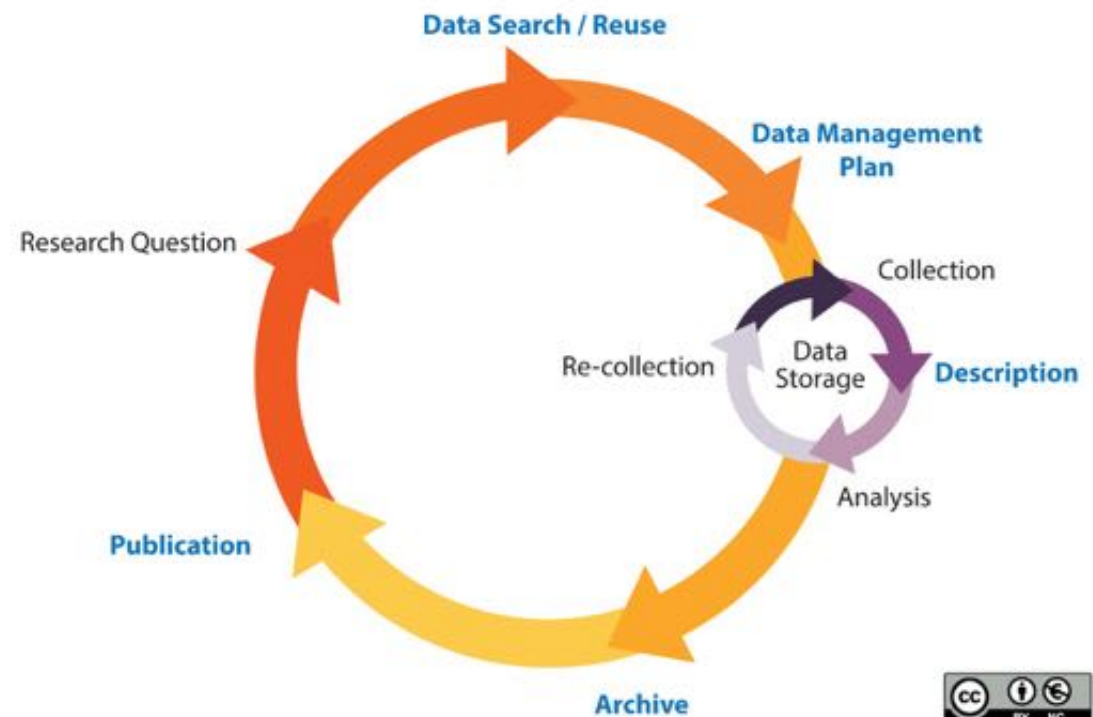
**Lyche & Steien, 2008, rcark1, conversation guidée, 00:03:41–00:04:03**



# Take-home message du cours

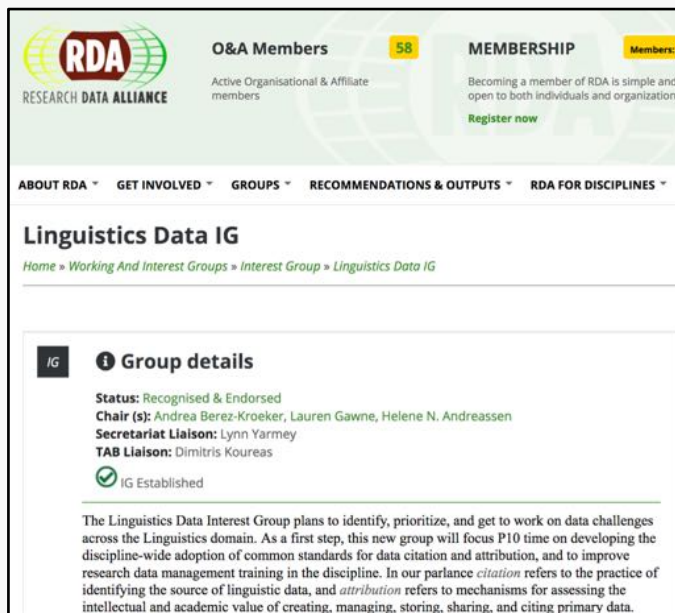
- Allez regarder de différentes banques de données.
- Profitez du matériel d'apprentissage mis à disposition en ligne.
- Demandez de l'aide et/ou des conseils.

## The Research Data Management Lifecycle



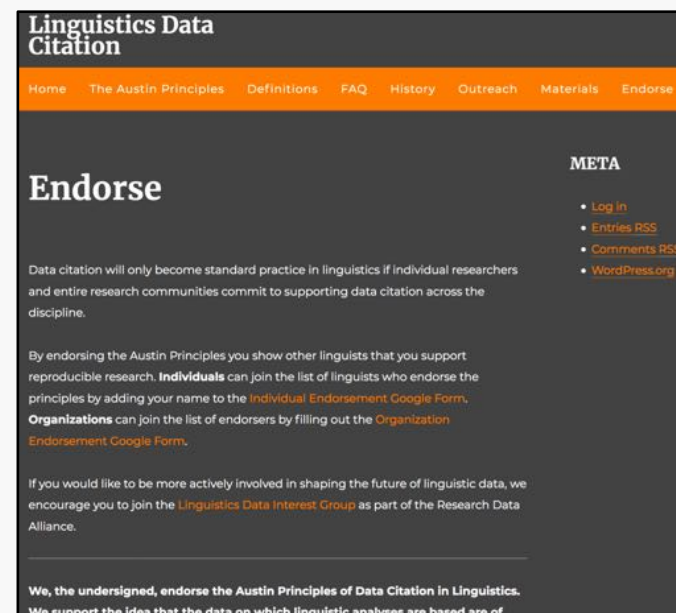
# Participer à la discussion

## Devenir membre du Linguistics Data IG



The screenshot shows the 'Linguistics Data IG' page on the Research Data Alliance website. At the top, there is a navigation bar with 'O&A Members' (58) and 'MEMBERSHIP' (Members). Below this is a navigation menu with 'ABOUT RDA', 'GET INVOLVED', 'GROUPS', 'RECOMMENDATIONS & OUTPUTS', and 'RDA FOR DISCIPLINES'. The main heading is 'Linguistics Data IG' with a breadcrumb trail: 'Home » Working And Interest Groups » Interest Group » Linguistics Data IG'. Under 'Group details', it lists: Status: Recognised & Endorsed; Chair(s): Andrea Berez-Kroeker, Lauren Gawne, Helene N. Andreassen; Secretariat Liaison: Lynn Yarmey; TAB Liaison: Dimitris Koureas; and a checkmark indicating 'IG Established'. A paragraph below describes the group's focus on data challenges, standards, and training in the field.

## Soutenir les Austin Principles



The screenshot shows the 'Endorse' page on the Linguistics Data Citation website. The page has a dark background with an orange navigation bar at the top containing 'Home', 'The Austin Principles', 'Definitions', 'FAQ', 'History', 'Outreach', 'Materials', and 'Endorse'. The main heading is 'Endorse'. To the right, there is a 'META' section with links for 'Log in', 'Entries RSS', 'Comments RSS', and 'WordPress.org'. The main text explains that data citation will become standard practice if researchers and communities commit to it. It encourages endorsing the Austin Principles and provides links for 'Individual Endorsement Google Form' and 'Organization Endorsement Google Form'. At the bottom, there is a statement: 'We, the undersigned, endorse the Austin Principles of Data Citation in Linguistics. We support the idea that the data on which linguistic analyses are based are of...'



## Activité 4: Préparer des données pour publication

Examinez le fichier *liaison IPFC-guide\_libre* et déterminez ce qu'il faut faire pour le rendre réutilisable.

Mots clés: format, appellation, description.

liaison IPFC-guide\_libre:

1ère version d'un fichier publié dans Andreassen, Helene N.; Lyche, Chantal, 2015, "Replication data for: Enchaînement, liaison, accentuation chez les apprenants norvégiens", <https://hdl.handle.net/10037.1/10176>, DataverseNO, V1.







# Données de recherche: La transparence et les meilleures pratiques

[helene.n.andreassen@uit.no](mailto:helene.n.andreassen@uit.no)

*Ecole d'été « Phonologie de corpus »*

*Lausanne, 22-26 juillet 2019*



Research data @ UiT

# Bibliographie

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, 533, 452-454.

Charsley, K. (2017). *Marriage migration and integration - Interview data*. [data collection]. UK Data Archive. <http://reshare.ukdataservice.ac.uk/852418/>

Corti, L., Van den Eynden, V., Bishop, L. & Woollard, M. (2014). *Managing and Sharing Research Data: a Guide to Good Practice*. Los Angeles: Sage.

European Commission. (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Version 3.2. Retrieved from [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-pilot-guide\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf)

Gawne, L., Kelly, B. F., Berez-Kroeker, A. L., & Heston, T. (2017). Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation*, 11, 157-189.

Janda, L. A. & Tyers, F. M. (2018). *Replication Data for: Less is More: Why All Paradigms are Defective, and Why that is a Good Thing*, <https://doi.org/10.18710/VDWPZS>, DataverseNO, V1, UNF:6:hiDJKeFSs1ZOcEs0dU0gw== [fileUNF]

Kirilova, Dessi, and Sebastian Karcher. 2017. Rethinking data sharing and human participant protection in social science research: Applications from the qualitative realm. *Data Science Journal* 16(43), 1-7. <https://doi.org/10.5334/dsj-2017-043>.

Lindsay, G. (2015, October 14). The latest medical breakthrough in spinal cord injuries was made by a computer program. Retrieved from <https://www.fastcompany.com/>

Mitchell, R. (2014). *Social networks, target language interaction and second language acquisition during the year abroad: a longitudinal study*. [data collection]. UK Data Service. <http://doi.org/10.5255/UKDA-SN-851169>



# Bibliographie

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1, e175.

<https://doi.org/10.7717/peerj.175>

Thomason, S. G. (1994). The Editor's Department. *Language*, 70(2), 409-413.

UiT The Arctic University of Norway. (2017). *Principles and guidelines for research data management at UiT*. Retrieved from [https://uit.no/forskning/art?p\\_document\\_id=521580](https://uit.no/forskning/art?p_document_id=521580)

Vines, Timothy H., Albert, Arianne Y. K., Andrew, Rose L., Débarre, F., Bock, Dan G., Franklin, Michelle T., . . . Rennison, Diana J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, 24(1), 94-97.

<https://doi.org/10.1016/j.cub.2013.11.014>

Whyte, A. (2015). *Where to keep research data. DCC checklist for evaluating data repositories*. V1. Edinburgh: Digital Curation Centre. Retrieved from <http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data/where-keep-research-data>

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>

Wouters, P., & Haak, W. (2017). *Open data: The researcher perspective*. Retrieved from <https://www.elsevier.com>

