DMP du projet "Project Places"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "Science Europe - DMP template (english)" fourni par Science Europe.

Plan Details

Plan title	DMP du projet "Project Places"
Version	Mid term version
Plan purpose/scope	DPM for storage with corpus generated by project and for disclosure to participants.
Fields of science and technology (from OECD classification)	Social and economic geography, Sociology
Language	eng
Creation date	2021-08-17
Last modification date	2021-10-16

Project Details

Project title	Project Places
Abstract	Le projet de recherche narrative, Places, veut rendre compte des processus d'expatriation et de transplantation vécus par des anglophones qui viennent vivre en France. L'investigation procède par le biais des entretiens qui se déroulent dans un lieu choisi par le participant. La démarche est biographique et autobiographique, car les deux cherch.eur.euse.s responsables sont, eux/elles-mêmes, des expatrié.e.s. Les entretiens visent un échange autant qu'une investigation. L'espace public agit comme une interface entre la biographie des participants et leur socialisation, permettant au projet de s'intéresser aux tactiques déployées pour communiquer, pour apprivoiser les nouveaux tissus urbains, et pour acquérir des nouveaux repères identitaires. Étant donné que les participants sont des anglophones expatrié.e.s qui sont venu.e.s, principalement pour les raisons de travail, un fort aspect du projet concerne le monde des affaires et le rôle qu'y joue l'anglophonie. Les participants sont sélectionnés par un démarchage auprès des associations, des institutions et des entreprises qui s'adressent aux expatriés anglophones. Les entretiens de recherche sont informels, visant à laisser parler les participants librement sur le sujet de leur arrivée et de leurs habitudes en France. Les entretiens ont lieu dehors, dans un lieu choisi par le participant et sont enregistrés. Les entretiens s'accompagnent par une prise de notes et de photographies par le ou la cherche.use.r. L'analyse se fait par

une annotation et une transcription interactionnelle des données audios, et puis par l'investigation des unités de discours narratives. Le paradigme d'investigation sociolinguistique est la recherche narrative, ce qui veut dire que sont analysés à la fois la mise en situation d'interlocution d'un récit et son contenu explicite. Cette analyse peut focaliser sur le langage, la stylistique, les rôles interactionnels, les itérations et les emprunts ainsi que sur les discours qui sont supposés et indexés par ces aspects du récit.

//

The narrative research project, Places, wishes to investigate processes of expatriation and transplantation that affect anglophones in France. Research progresses through interviews that take place in an outside location chosen by participants. The approach is both biographic and autobiographic since the two lead researchers are both expatriates. Place is an interface between the biographies of participants and their new conditions of socialisation, revealing the tactics relied upon to communicate, to integrate the city and to acquire new ways of being. Given that the participants to the research are expatriated anglophones who have primarily come to France for work reasons, a significant focus of the project is on the business community and the role of English within that community. Participants are selected by approaching associations, institutions and companies that are concerned with anglophone expatriates. Research interviews are unstructured and aim to allow participants to speak freely about their arrival and their habits in France. Interviews take place outside, in a location chosen by the participant and are audio recorded. Interviews are accompanied by notes and photographs taken by the researcher. Analysis proceeds by annotation and interactional transcription of the audio data, and then by the parsing of narrative discourse units. The paradigm for the sociolinguistic research is narrative investigation, which means that data is analysed from the perspective of its interactional achievement as well as in terms of overt content. Parsing can concern linguistic features, interactional roles, iteration and borrowings as well as those discourses that are supposed or indexed by certain aspects of the narratives.

Funding

- LIDILE :

Start date 2021-09-01

End date 0027-09-01

Partners

- LISAA 105538

Research outputs :

1. Series of .wav audio segments, ELAN (.eaf) transcriptions and photographs (Dataset)

Contributors

Name	Affiliation	Roles
William Kelleher - https://orcid.org/0000-0001-8368-6404		• DMP manager • Personne contact pour les données • Project coordinator

DMP du projet "Project Places"

1. Data description and collection or re-use of existing data

1a. How will new data be collected or produced and/or how will existing data be re-used?

Project Places corresponds to newly created data. All data elements are created as part of the project. Data is created through recorded interviews with research participants.

The project includes three types of data:

1. Documentation: authorisations, corpus documentation, metadata, information notices and technical documents;

2. Corpus: corpus of participant-related data such as recordings, transcription, notes and photographs;

3. Data: analytic data produced through analysis of the participant-related data.

Re-use: The corpus of Project Places research project data (interviews, photographs and transcripts) will be uploaded to an open data-sharing platform called Nakala that is run as part of the Huma-Num Very Large Research Infrastructure (“Très Grande Infrastructure de
Recherche”, TGIR). This corpus will contain interview data and transcripts that may be re-used by other scientific investigators for studies into, for example, language varieties, narrative structure, experience of expatriation and interactional pragmatics. When re-used, the data and their source must be quoted (using the doi provided by the data storage platform Nakala, Huma-num).

1. Documentation, a) primary documents: Data provenance is documented at the time of recording via an authorization form provided by the researcher and filled in by the participant. Authorisations concern participation in interviews, pseudonymisation and data storage. Metadata concerns participant gender, country of birth, language, occupation, socio-economic information, place, time and conditions of the interview. Primary documentation also covers how the corpus is collected and the data organised and structured. A consent form and a metadata collection form are part of the project. b) secondary documents: Information provided by local government instances, statistical bureaux, civil society instances, bibliographic references etc.

2.Corpus: Interviews with participants give rise to a corpus of audio files, anonymised photographs of the environment surrounding the interview location, researcher notes and ELAN transcripts. ELAN is annotation software made available by the Max Planck institute. These data are stored in several formats corresponding to open standards (ODT, EAF, CSV, WAV and UTF-8 txt) and some that are proprietary (JPEG and DOCX). Participants are selected by approaching associations, institutions and companies that are concerned with anglophone expatriates. Research interviews are informal and (auto-)biographical, with a free exchange with the researchers who are, themselves, expatriates. The aim of an informal format is to allow participants to speak freely about their arrival and their habits in France. Interviews take place outside, in a location chosen by the participant and are audio recorded. Interviews are accompanied by notes and photographs taken by the researcher. A corpus of photographs and audio recordings accompanied by ELAN annotations, will be stored on a data-sharing platform Nakala that falls under the aegis of the Huma-Num Very Large Research Infrastructure (“Très Grande Infrastructure de Recherche”, TGIR). This corpus will be localisable and available to other researchers through metadata and project description. The data on Nakala is accompanied by a concept note for the project, full ethical and DMP documentation and a ReadMe document that set out the contacts for the project coordination, the methodology, the interview guide and the relative agreement for citation and re-use, the Creative Commons share-alike non commercial licence. The file that is uploaded to Nakala will be a 7-zip file, within each 7-zip file will be .txt files for notes and ReadMe, .csv files for metadata and annotations, .jpg files for photographs of the site, a .wav file for the audio recording and an .eaf ELAN transcription file.

3. Data: Analysis proceeds by annotation and interactional transcription of the audio data with ELAN (.eaf), and then by the parsing of narrative discourse units (.csv, .doc, .txt and .xml formats). The paradigm for the sociolinguistic research is narrative investigation, which means that data is analysed from the perspective of its interactional achievement as well as in terms of overt content. Parsing can concern linguistic features, interactional roles, iteration and borrowings as well as those discourses that are supposed or indexed by certain aspects of the narratives. Analytic findings will be made available in the form of scientific publications that may also reference the Digital Object Indicator (DOI) of the relevant Nakala corpus data.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

These data are stored in several formats corresponding to open standards (ODT, EAF, CSV, WAV and UTF-8 txt) and some that are proprietary (JPEG, PDF and DOCX). The total size of the corpus will be approximately 15GB.

The collected documentation includes:

1. Ethics and data storage authorisation forms filled in and signed by the participant and that provide for confidentiality by attribution of a pseudonym.

2. Metadata form filled in by the researcher in .csv format and later used to constitute the corpus metadata file) that

respects pseudonymisation.

3. Researcher notes (these will be handwritten in a researcher field book, scanned and stored in PDF format) that respect pseudonymisation.

4. Secondary documentation in .pdf reader files.

The collected data that will constitute the corpus includes:

1. An audio recording of a semi-guided interview in either MP3 or WAV formats (.mp3 formats are light and transportable. .wav formats allow a waveform that can be incorporated into transcription software such as ELAN) that respects pseudonymisation.

2. Anonymised photographs of the surroundings of the interview location in JPEG format.

3. ELAN annotation files (.eaf) that respect pseudonymisation. ELAN software is chosen for the quality and standardised nature of resulting transcripts. ELAN is developed by the Max Planck institute and supported by the United Nations. It is also suitable for storage in a corpus collection.

Based on these two types of data and their pseudonymisation, further data may be produced and stocked in addition to the documentation and corpus data:

1. Spreadsheets with annotations (CSV). These allow transcriptions to be prepared and then imported into ELAN, they also allow export into other formats such as TXT or XML that is useful for corpus analysis.

3. Parsing of data in DOCX files.

---

2. Documentation and data quality

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

The corpus will be stored on the Nakala server provided by the Huma-Num Very Large Research Infrastructure (“Très Grande Infrastructure de Recherche”, TGIR).

Storage on Nakala will be under a project 'collection' and accompanied by full project documentation for ethics and data management, a project concept note as well as a ReadMe form that sets out researcher contacts, methodology, interview schedule and data re-use policy - the Creative Commons share-alike non commercial licence. Nakala uses metadata included in the Dublin Core Metadata Initiative (DCMI) standard to help others identify and discover the data. This includes:

Concerning the participant (the confidentiality of whose data is protected by pseudonymisation and editing out of identifying information):

- participant gender, country of birth, language, occupation, socio-economic information, place, time and conditions of the interview.

Concerning the data files:

- title, subject, description, creator, contributor, date, type, format, identifier, source, language, rights.

A metadata entry protocol is distributed to participating researchers. It's contents are as follows:

Metadata is necessary to the CONSENT FORM and for the NAKALA data sharing platform.

Language: ENGLISH, FRENCH, SPANISH etc

Dates for interviews: DD – MM - YYYY

Length of interviews: hh h mm’ secs’’ eg 2h34’52’’

Pseudonym: ONE SINGLE NAME

Dublin Core Metadata Initiative (DCMI) - https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#http://purl.org/dc/terms/RFC5646

For the interview metadata:

- date: DD – MM - YYYY----------
- language spoken: ENGLISH, FRENCH, SPANISH etc----------
- declared gender: MALE / FEMALE / LGBTIQ----------
- age: IN NUMBERS eg. 39----------
- education level: PHD, MSC, BA, SEC CERT, VOC----------
- socio-economic status: UPPER/MIDDLE/LOW----------
- occupation: ONE OR TWO WORDS----------
- country of origin (birth): FULL SPELLING eg. UNITED STATES OF AMERICA----------
- theme (work/retirement/other reason): SHORT PHRASE----------
- city of residence: FULL SPELLING----------

- neighbourhood of residence: FULL SPELLING----------
- city of work: FULL SPELLING----------

neighbourhood of work: FULL SPELLING

For Nakala files under collections:

- Data type: dataset
- Title: pseudonym_segment_1
- Author: researcher
- Creation date: interview date
- Licence: Creative commons attribution non commercial 4.0 international
- Description: English + Research interview data. Transcript and audio for research participant.
- Keywords: data, narrative, interview, ELAN, transcription, expatriation, anglophone, France
- Languages: English
- Subject: interview data
- Format: .pdf, .docx, .eaf, .jpg
- Source: Project Places
- Rights holder: University of Rennes

For ELAN:

- https://datcatinfo.termweb.eu/search/terms
  - **ISO 12620:2019**
- **annotation**: http://datcatinfo.termweb.eu/datcat/DC-2462
- **speaker turn**: http://datcatinfo.termweb.eu/datcat/DC-2976
- **narrative**: http://datcatinfo.termweb.eu/datcat/DC-4634
- **discourse unit**: http://datcatinfo.termweb.eu/datcat/DC-4577
- **discourse**: http://datcatinfo.termweb.eu/datcat/DC-2608
- **utterance**: http://datcatinfo.termweb.eu/datcat/DC-1409
- **morphosyntactic annotation**: http://datcatinfo.termweb.eu/datcat/DC-2320
- **lexeme**: http://datcatinfo.termweb.eu/datcat/DC-1325

The data hosted on Nakala are organised around the concept of participant (whose confidentiality is protected by anonymisation). Each participant data is uploaded in a .zip file . Participant data consists of photographs of the site of the interview in .jpeg, .wav audio recording of interview, ELAN .eaf transcription of interview, metadata in .csv, notes in .txt. The data is attributed a DOI. The ReadMe document is uploaded and accompanies each set of participant data.

During the project, data that is in the process of being collected will be stored with the following folder structure: project name > participant data > participant identifier > visual data / audio data / transcription data. Full details on folder structure are included in a protocol that will be given to participating researchers. The contents of this protocol are as follows (see also: https://doranum.fr/stockage-archivage/comment-nommer-fichiers/):

- Initial participant folder by pseudonym (initial capitalisation) + underscore + place of interview: **Pauline_Place de la Republique**
  1. Within initial participant folder under participant pseudonym: **Pauline**
     1. Checklist and flow sheet to be filled in per participant: **Pauline_Checklist + flow**
     2. Metadata sheet with metadata to be used on Consent forms + Nakala: **Pauline_metadata**

should contain: participant pseudonym, date and place of interview, length of interview, metadata necessary for Nakala (use Metadata sheet proforma)

- Notes: **Pauline_Notes**

1. pseudonym + underscore + photos: **Pauline_photos**
   1. Photographs of interview location with original numbering from camera
   2. Edited photographs (faces blanked out) by name and number: **Pauline_1**, **Pauline_2**, **Pauline_3**, etc
2. pseudonym + underscore + audio: **Pauline_audio**
   1. Mp3 files of interview by pseudonym + underscore + date + version number: **Pauline_2021-10-05_1**
   2. Edited segments that will be uploaded to sharedocs and Nakala by pseudonym + underscore + segment + version number: **Pauline_segment_1**
3. pseudonym + underscore + transcription: **Pauline_transcription**
   1. ELAN .eaf files (that need to match the mp3 file extensions) by pseudonym + underscore + segment + version number: **Pauline_segment_1**
   2. MAKE SURE THAT IN ELAN the NAME OF THE ANNOTATOR is given in tier attributes, when adding or changing annotations ADD THE NAME OF THE MOST RECENT ANNOTATOR etc: **William + Hillary + Shannon ...**

- All versions are in rising numerical order from **_1**

2b. What data quality control measures will be used?

Consistency: Data are collected by a LIDILE researcher. The audio recordings are transcribed by the researcher or by a transcription service compatible with ELAN. ELAN transcription involves the use of controlled vocabularies. The files of audio recordings, photographs and notes are verified in their structure and annotation by the researcher specialised in linguistics in consultation with a colleague, also specialised in linguistics.

Methodology: Participants are invited to an informal interview that takes place at a location of their choosing. The interview contains questions that concern: the biography of the participant, their arrival in France, their adaptation to the new environment, questions of language learning and adaptation. Interviews are accompanied by researcher notes and photographs of the location of the interview. These additional data types serve to give depth and contextualisation to the interview. A guide to possible interview questions is provided for both researcher and participant.

Possible interview questions:

Informal and (auto-) biographic interview questions

The informal interviews with participants that are envisioned as an exchange with the researchers, who are also expatriates, concern, essentially a) their biography, and its relation to space/city, b) the move to France and its implications personally, professionally, in terms of new habits and ways of being, and c) the places/spaces with which these changes are concerned.

The following questions are merely indicative serving as a guide for the kinds of prompts that the researcher might give a participant.

1. Biography
    - Where were you born?
    - Can you tell me a bit about that place?
    - What did it mean to you?
    - Have you moved since? Where to?
    - Can you remember any places to which you have/had a particular affective link or which evoke memories for you?
    - Where did you go to primary school/secondary school/university …?
    - Tell me more about your family
    - Where do they live?
    - Do you visit often?
    - Where would you say your 'home' is?
    - Can you think of an anecdote that concerns your home/place where you grew up?
    - What kind of person would you say you are ? (outgoing, talkative etc)
    - What languages do you speak?
    - What is your job/position/responsibilities?
    - Where else have you worked? Which workplace had the most significance to you?

1. Move to France
    - When did you move to France?
    - Where did you move to France from?
    - What was it like leaving that place?
    - What was the move/process of leaving and arriving like?
    - Have you been overseas (for work purposes) before?
    - Why did you come to France?
    - How long will you be staying?
    - What does this place mean for you?
    - How did/do you go about feeling at home?
    - How do you get around?
    - What do you like/dislike about this place/city/country?
    - If you had to compare your home/last residence with this one, what would you say?
    - Are you learning French? / Do you speak French? / Where did you learn French?
    - How did you find your current residence?
    - What were the implications of this move for your family?
    - Can you think of an anecdote that sums up this experience (of living in/moving to France) for you?
    - How has it been to integrate your new workplace?
    - Have there been any obstacles/advantages/differences specifically related to working conditions here?

1. Spaces
    - What was the reason for choosing (the place in which the interview is conducted)?
    - What kinds of new habits do you have here?
    - Are there any places (like coffee shops, (super)markets …) that your new life rhythms involve?
    - What can you say of these places?
    - Can you think of an anecdote that sums up what these new places are like?
    - Can you compare your new places with your old in terms of:
        1. Weather

2. How people behave

- Urban form

1. Distance
2. Movement

- Have you been involved in any activities (sporting, social, associative…) since you moved here?
- Do you think that this is a place in which you could be happy? Why/why not?
- Can you think of any kinds of events that defined your birthplace/home/growing up/change/expatriation/work and where did they happen?
- What are the differences between your previous residence and your new in terms of the relationship between home and work?

Prior to the beginning of all recording sessions, participants are asked to read and sign an information form and a consent form. They are also asked to agree to a pseudonym that will then be used in all data conservation, treatment and analysis. The initial questions of the interview concern a metadata questionnaire. The consent form and metadata questionnaire were created as part of the project and are based on the documents available from https://msh-lorraine.fr/nos-services/droits-et-obligations/

For quality and respect of protocols elaborated for this data management plan, a workflow checklist is given to all participating researchers. Its contents are as follows:

SHEET FOR – PSEUDONYM: _____

SHEET OPENED ON: ___ / ___ / _____

	Action/Event	**Yes**	**Dates (dd/mm/yyyy)**
	First contact		
1	Give out information sheet + project concept note		___ / ___ / _____
	Interview		
2	Meetings: 1) 2) 3) 4)		___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____
3	Give out ethics consent forms (making available Data Management Plan)		___ / ___ / _____
4	Complete meta-data for Nakala corpus (respect meta-data entry protocol)		___ / ___ / _____
5	Choose pseudonym for data storage and treatment		___ / ___ / _____
6	Save photographs and audio recording on devices under pseudonym (respect file extension naming and organisation protocol) Save metadata sheet under pseudonym Save checklist sheet under pseudonym		___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____
	Data storage		
7	Create folder for participant + correctly name file extensions for data (respect file extension naming and organisation protocol)		___ / ___ / _____
8	Edit audio and photographs to remove all identifying or sensitive data (respect ethical guidelines documentation)		___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____

9	Upload edited and pseudonymised audio and photographs to sharedocs	___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____
	Annotation	
10	Import audio to ELAN and use template with tiers and controlled vocabularies	___ / ___ / _____
11	Annotate 1) 2) 3) 4) 5) 6) 7)	 ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____
12	Upload versions to sharedocs with extension _1, _2 etc 1) 2) 3) 4) 5) 6) 7)	 ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____ ___ / ___ / _____
13	Check annotations with pair researcher	___ / ___ / _____
	Corpus	
14	Indicate to head researcher that participant data is ready for uploading to Nakala	___ / ___ / _____
15	Head researcher opens a folder on Nakala + completes metadata + publishes	___ / ___ / _____ ___ / ___ / _____

---

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during the research?

The Places data set has its specific DOI persistent identifier. The Project Places also serves as a data collection node on the Nakala data storage platform with its own DOI. This collection will be associated with a LIDILE collection, also hosted on Nakala, so that the individual data set is associated with the other data sets produced by the LIDILE research unit.
The data and metadata are stored on servers belonging to the Human-Num Nakala server provided by the Huma-Num Very Large Research Infrastructure (“Très Grande Infrastructure de Recherche”, TGIR). The
service includes backup on a daily basis.

Raw data are stored on the hard drive of the LIDILE researcher. Only the researcher has access to this hard drive that is access protected.

As soon as possible the researcher edits the raw data to remove any identifying information or any sensitive information. The researcher then uploads edited and pseudonymised data to the Sharedocs platform hosted by Huma-Num the Very Large Research Infrastructure (“Très Grande Infrastructure de Recherche”, TGIR).

Sharedocs is used for collaborative work between researchers who in turn store versions of data on their hard drives using the data storage and file extension naming protocol referred to above.

Once data is annotated and checked the pseudonymised and edited data are then uploaded into the corpus hosted on Nakala. This corpus includes metadata, concept notes full project documentation and project ReadMe notice. The Huma-num data protection policy applies to the data stored on Nakala and on ShareDocs. These platforms are secured by institutional firewalls, identification, passwords and charters of good usage. Backup solutions are provided as part of the service. The data uploaded to Nakala will be a 7-zip file, within each 7-zip file will be .txt files for notes and ReadMe, .csv files for metadata and annotations, .jpg files for photographs of the site, a .wav file for the audio recording and an .eaf ELAN transcription file. This offers the advantage of being able to protect data 7-zip files with a password. This password can be made available in the ReadMe. It does not alter the accessibility of the data but it does mean that those accessing the data must read the ReadMe and take cognizance of citation norms, researcher contact information, project documentation and methodology. It is a significant and additional level of protection.

A good practice participant checklist and workflow was given above in 2b and gives relevant storage instructions.

Head researcher will regularly check backup, ShareDocs and Nakala.

3b. How will data security and protection of sensitive data be taken care during the research?

The aim of Project Places is not to collect sensitive data.

The confidentiality of non-sensitive participant data is protected by pseudonymisation.

Before beginning an interview, and when reading information and consent forms, participants are requested to choose a pseudonym. It is then solely this pseudonym that is used to refer to the participant in the interview, in the recording, and then for data storage, editing to remove identifying information and for conservation.

Sensitive data that is removed is defined by the institutional DPO and includes:

racial or ethnic references, political opinions, religious and philosophical convictions, trade union affiliation, genetic and biometric data, health data, sexual data, penal record, identification document data and numbers.

Confidential data that is removed includes:

names and addresses of persons directly related to the participant, proper names such as residential street addresses, telephone numbers, work names and addresses and specific identifying times and events.

Raw data from the interview are stored on hard drive of the LIDILE researcher. Only the researcher has access to this hard drive that is access protected. Only the pseudonymised and edited data are uploaded onto the corpus hosted on Nakala that is destined for the scientific community.

Huma-num Very Large Research Infrastructure (“Très Grande Infrastructure de Recherche”, TGIR) data protection policy applies to the data stored on Nakala. Backup solutions are provided as part of the service. The data uploaded to Nakala will be a 7-zip file, within each 7-zip file will be .txt files for notes and ReadMe, .csv files for metadata and annotations, .jpg files for photographs of the site, a .wav file for the audio recording and an .eaf ELAN transcription file. This offers the advantage of being able to protect data 7-zip files with a password. This password can be made available in the ReadMe. It does not alter the accessibility of the data but it does mean that those accessing the data must read the ReadMe and take cognizance of citation norms, researcher contact information, project documentation and methodology. It is a significant and additional level of protection.

---

4. Legal and ethical requirements, codes of conduct

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

Subjects are given an information sheet that details the project aims, methodology and types of data collected as well as the data management plan. They are also given information regarding the pseudonymization of their data and they sign a consent form. This consent form gives the reason for data collection which is the public mission of the university to collect data that serves for scientific research. The consent form also gives details for contacting the data protection officer of the university and the data protection agency (the CNIL - Commission Nationale de l'Informatique et des Libertés - National Commission for Freedoms and Information Treatment). They can contact the researcher to access, correct or delete their data.

Nakala is a research infrastructure, it is destined for the community of researchers and protected by registration and by a creative commons share-alike non commercial licence for re-use and citation. In addition, the data uploaded to Nakala will be a 7-zip file protected with a password. This password can be made available in the ReadMe. It does not alter the accessibility of the data but it does mean that those accessing the data must read the ReadMe and take cognizance of citation norms, researcher contact information, project documentation and methodology. It is a significant and additional level of protection.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

The data is held by the LIDILE research team of Rennes 2 university. Access is controlled by the researcher.
The data on Huma-Num Nakala is available to the scientific community via persistent DOI. It is protected under a Creative Commons share-alike non commercial licence.

Conformity of the project and ethical considerations are checked and verified following a submission of all relevant documentation for the project to the University Rennes 2 data management consultant and legal and ethical committee and its Data Protection Officer.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Conformity of the project and ethical considerations are checked and verified following a submission of all relevant documentation for the project to the University Rennes 2 data management consultant and legal and ethical committee and its DPO.

The University Rennes 2 has a data management support platform SOCle (https://socle.univ-rennes2.fr/) that hosts contacts, and a good usage charter.

Only data necessary for studying sociolinguistically pertinent processes of language and adaptation to expatriation are stored on Nakala. Raw data are stored on the researcher computer and are not shared. Only pseudonymised and edited data will be shared, for questions of research and annotation consistency and reliability with research colleagues. Sharing will be facilited through a sharedoc platform such as the one hosted by Huma-Num the Very Large Research Infrastructure (“Très Grande Infrastructure de Recherche”, TGIR).

Consent forms, that contain personal information, are paper-based and kept seperate from all data either raw or treated and are therefore confidential. Pseudonymisation is robust.

Role	Type of access	Location
Researcher	Full	Hard drive, sharedoc, Nakala
Colleague	Pseudonymised and edited participant data elements	ShareDocs, Nakala

Sensitive data that is removed is defined by the institutional DPO and includes:

racial or ethnic references, political opinions, religious and philosophical convictions, trade union affiliation, genetic and biometric data, health data, sexual data, penal record, identification document data and numbers.

Confidential data that is removed includes:

names and addresses of persons directly related to the participant, proper names such as residential street addresses, telephone numbers, work names and addresses and specific identifying times and events.

---

5. Data sharing and long-term preservation

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Data are shared publicly through query interfaces (Nakal UI and API) linked to data stored in the Nakala database. These interfaces will give access to the metadata of the Nakala collection for Project Places. The pseudonymized and edited participant data is available to the scientific community.
Data are shared under the Creative commons share alike non-commercial licence.

The research calendar for the project is as follows:

Phase	Start	End	Duration
1 : Data collection	01 September 2021	01 May 2024	2 year ½
2 : Corpus collation	01 September 2021	01 September 2024	3 years
3 : Narrative analysis	01 September 2021	01 September 2025	4 years
4 : Publications	01 September 2021	01 September 2027	6 years

The data uploaded to Nakala will be openly accessible to the scientific community and citable through attribution of a DOI. There are two levels of restriction to data sharing. Firstly, Nakala itself forms part of the Huma-Num TGIR and requires a sign-in for full functionality, although it does also have a preview option. Secondly, within Nakala access to the Places data requires specific software. The 7-zip software is necessary for opening the data files. These are protected with a password. This password can be made available in the ReadMe. It does not alter the accessibility of the data but it does mean that those accessing the data must read the ReadMe and take cognizance of citation norms, researcher contact information, project documentation and methodology. It is a significant and additional level of protection. Secondly, the photographs will need a viewer compatible with the .jpeg format. The audio will need a reader compatible with the .wav format. The ELAN transcription .eaf requires ELAN. An office application will be needed to read .txt and .csv files. 7-zip is downloadable here: https://www.7-zip.org/

ELAN is downloadable here: https://archive.mpi.nl/tla/elan/download

---

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Pseudonymised and edited participant data is constituted into a corpus and will remain in Nakala servers for as long as the project is active. This is done because accumulated data will provide insights into participant orientation to the subject of research.
The Project Places corpus that is held on Nakala is available to the scientific research community. The data are stored and available online in the long run in order to support longitudinal research. The data could be reused for different types of analyses focused on different linguistic dimensions.
Following recommendations from the "Association des archivistes français": the data will be kept indefinitely:
"Conservation définitive et intégrale des documents dont l'intérêt historique ou scientifiquele justifie, dans le service public d'archives territorialement compétent."
(Source document)

5c. What methods or software tools are needed to access and use data?

Nakala APIs are available. Queries on the database will be handled directly with the researchers responsible for Project Places. The University Rennes 2 has a data management support service, SOCle (https://socle.univ-rennes2.fr/) that may be contacted at: donnees-recherche@listes.univ-rennes2.fr

The Places Nakala data will also be associated with the LIDILE research unit collection that will include institutional affiliations and contacts.

The data uploaded to Nakala will be openly accessible to the scientific community and citable through attribution of a DOI. There are two levels of restriction to data sharing. Firstly, Nakala itself forms part of the Huma-Num TGIR and requires a sign-in for full functionality, although it does also have a preview option. Secondly, within Nakala access to the Places data requires specific software. The 7-zip software is necessary for opening the data files. These are protected with a password. This password can be made available in the ReadMe. It does not alter the accessibility of the data but it does mean that those accessing the data must read the ReadMe and take cognizance of citation norms, researcher contact information, project documentation and methodology. It is a significant and additional level of protection. Secondly, the photographs will need a viewer compatible with the .jpeg format. The audio will need a reader compatible with the .wav format. The ELAN transcription .eaf requires ELAN. An office application will be needed to read .txt and .csv files. 7-zip is downloadable here: https://www.7-zip.org/

ELAN is downloadable here: https://archive.mpi.nl/tla/elan/download

---

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Nakala provides persistent identification of the data via a unique handle managed by the Corporation for National Research Initiatives
(CNRI).

---

6. Data management responsibilities and resources

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Data management, quality, storage and backup as well as DMP implementation will be overseen by the LIDILE researcher. The DMP is verified with the "Service d'accompagnement à la gestion des données" of the university of Rennes 2.

The university Rennes 2 has a data management support platform SOCle (https://socle.univ-rennes2.fr/) that has contact information and an institutional good usage charter.

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

The LIDILE researcher is responsible for regular data collection, storage and curation tasks. This work time is part of their academic duties.

Project Places will be submitted to calls for financing. Depending on the terms of these calls the project responsabilities might change. This section will be revised in terms of those engagements.