April 11, 2024

Implementing Standards for Data Catalogs in Research Organisations: A Case Study of the French Institute for Demographic Studies

Conference On Smart Metadata
for Official Statistics, 2024
Julie BARON, Julie LENOIR

Overview

Introduction

Part I: Implementing DataIned

Part II: Rethinking the dissemination process

Conclusion

Introduction

ined
INSTITUT NATIONAL
D'ÉTUDES DÉMOGRAPHIQUES

The French institute for demographic studies

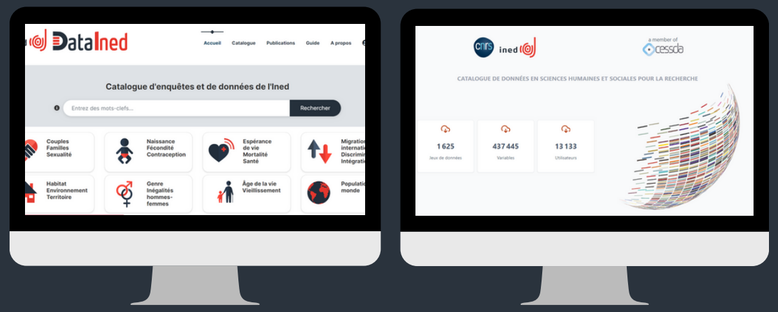
- About 250 people, 11 research units and 7 research support services

- 8 main research themes:

- Survey data as well as demographic and contextual databases

Introduction

QUETELET PROGEDO DIFFUSION



2018 GDPR

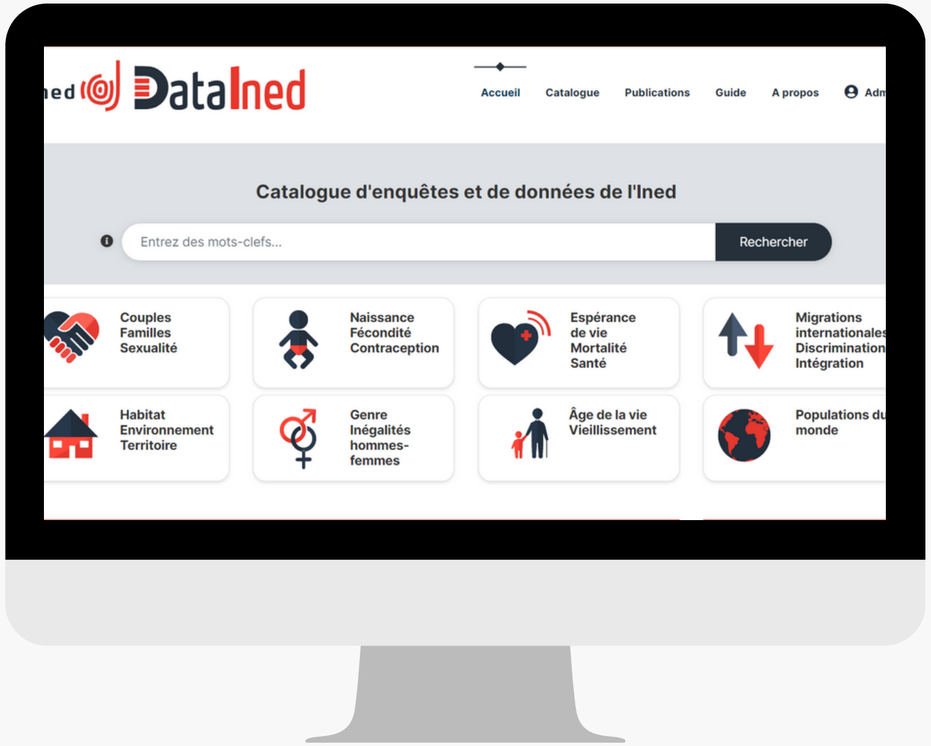
Research data financed from public funds must be disseminated (First National Open Science Plan)

2021 New framework for data dissemination timeline adopted by Ined Scientific Council (3 years after data collection)

2022 INED Open Science Charter

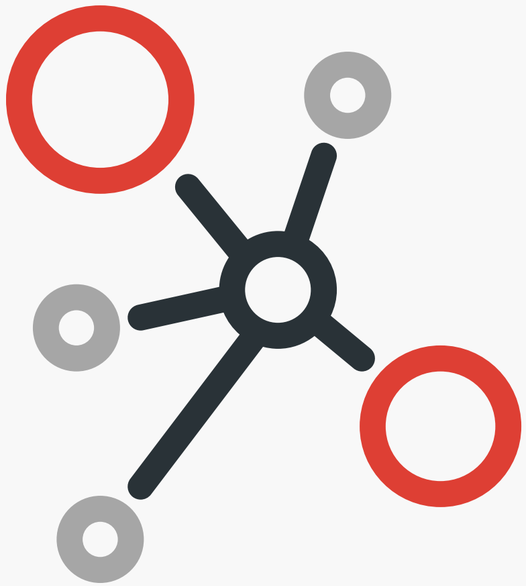
Introduction

- We needed a new catalogue since Nesstar was becoming obsolete
- We needed a more modern tool, that would better showcase surveys and data
- We needed to comply to European standards in order to implement harvesting routines



Part I: Implementing Datained

1. Metadata standardisation and enrichment
2. Choose a software
3. Going from Nesstar to NADA: the migration process

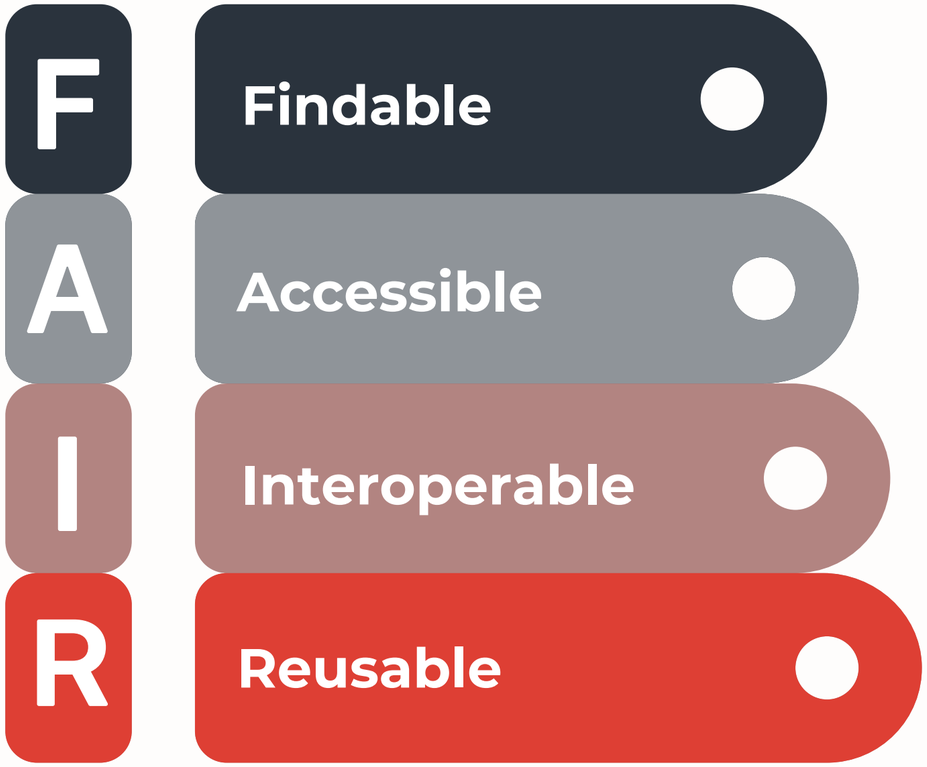


Part I: Implementing Datained

1. **Metadata standardisation and enrichment**
2. Choose a software
3. Going from Nesstar to NADA: the migration process

① Metadata standardisation and enrichment

1. Our Standards



DATA DOCUMENTATION INITIATIVE

DDI-Codebook 2.5

① Metadata standardisation and enrichment

2. Metadata fields we standardised

Field	Standard
Topics	INED / Cessda
Keywords	Cessda (ELSST)
Institutions	INED (PROGEDO)
Sampling procedure	Cessda
Time method	Cessda

Field	Standard
Collection mode	Cessda
Research instrument	Cessda
Unit of analysis	Cessda
Country	Cessda (ISO-3166-1)
Data type	DDI Alliance
Identifier	DOI (Datacite)

① Metadata standardisation and enrichment

3. The process

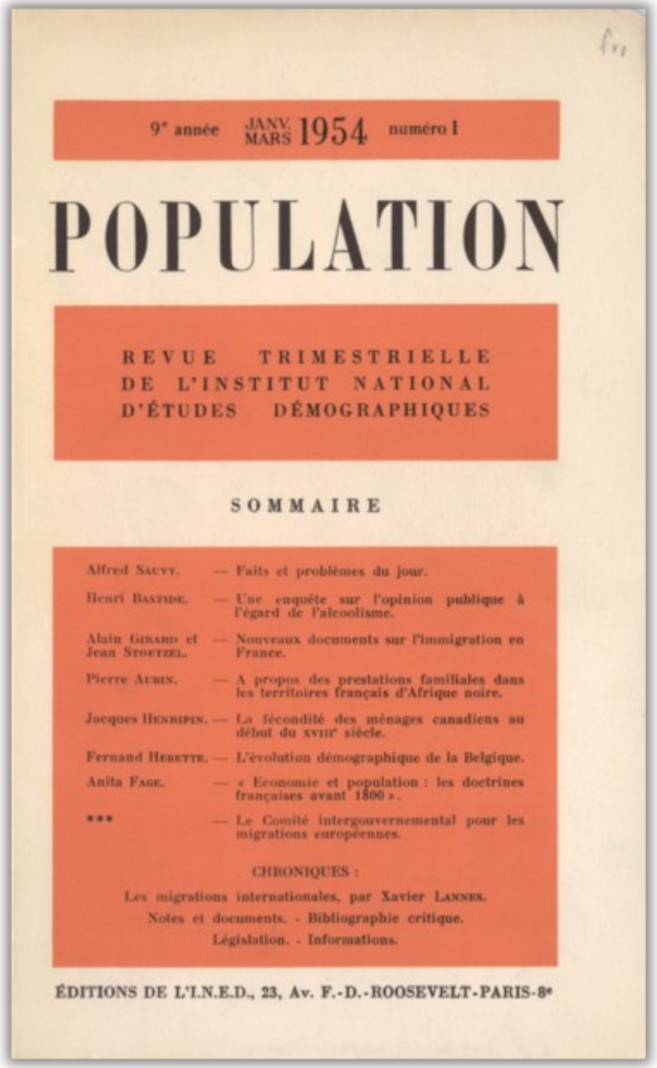
sommaire | Instruments de collecte (une ligne par enquête)

ID_enq	titre_enquête	Description	Instrument_1	Instrument_2
IE0001	Enquête natalité - Ensemble du pu	Questionnaire directif comprenant 13 questions en plus de celles su	Questionnaire structuré	
IE0002	Enquête natalité - Ensemble du pu	Questionnaire directif comprenant 12 questions en plus de celles su	Questionnaire structuré	
IE0003	Enquête risque et prévoyance (194	Questionnaire directif avec possibilité d'émettre un commentaire s	Questionnaire structuré	
IE0004	Attitude des jeunes filles et des jeu	Questionnaire semi-directif avec possibilité d'émettre un comment	Questionnaire semi-structuré	
IE0005	Enquête orientation professionne	Questionnaire semi-directif comprenant 11 questions en plus de ce	Questionnaire semi-structuré	
IE0006	Enquête influence sur le sommeil (	Questionnaire semi-directif avec possibilité d'émettre un comment	Questionnaire semi-structuré	
IE0007	Les conditions d'existence des fan	Les carnets de compte recueillis pour l'enquête "Les conditions d'ex	Questionnaire semi-structuré	Consignes de collecte de données : guide des écrits auto-administrés
IE0008	Désirs des Français en matière d'h	Questionnaire semi-directif comprenant 45 questions en plus de ce	Questionnaire semi-structuré	
IE0009	Les conditions d'existence des fan	3 documents de collecte ont été utilisés, similaires à ceux de l'enqu	Questionnaire semi-structuré	Consignes de collecte de données : guide des écrits auto-administrés
IE0010	Une possibilité d'immigration itali	Questionnaire semi-directif avec possibilité de laisser un commenta	Questionnaire semi-structuré	
IE0011	Résultat d'une enquête préliminai	Le questionnaire a été établi et expérimenté en 1945 par Mme le D	Questionnaire semi-structuré	
IE0012	Enquête par sondage sur l'âge de l	Deux questionnaires distincts ont été rédigés : le questionnaire A et	Questionnaire semi-structuré	
IE0013	Le niveau intellectuel des enfants	Deux instruments ont servi à la collecte des données : - Le test "mo	Questionnaire non-structuré	Tâches des participants
IE0014	Le temps de travail des femmes m	Les documents de collecte des données étaient au nombre de 2 : -	Questionnaire non structuré	Consignes de collecte de données : guide des écrits auto-administrés
IE0015	Les conditions d'existence des fan	L'instrument utilisé pour la collecte des données est un carnet de c	Consignes de collecte de données : guide des écrits auto-administrés	
IE0016	Les conditions d'existence des fan	3 documents de collecte ont été utilisés : - Un carnet de compte co	Consignes de collecte de données : guide des écrits auto-administrés	
IE0017	Les conditions d'existence des fan	La présente étude a été réalisée à partir des données récoltées par	Consignes de collecte de données : guide des écrits auto-administrés	
IE0018	Les conditions d'existence des fan	Un document de collecte unique a été utilisé : un carnet de compte	Consignes de collecte de données : guide des écrits auto-	

① Metadata standardisation and enrichment

4. Adding to our metadata

- Making sure important information are available
- Searching for and referencing of bibliographic resources related to the surveys
- Adding links to these resources in all relevant references



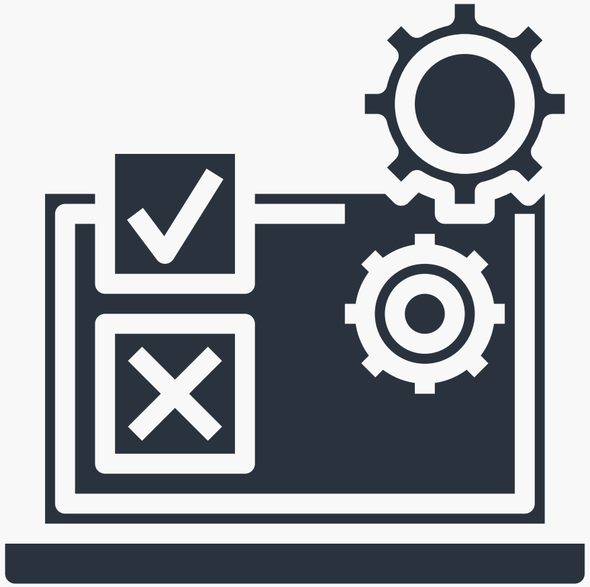
9e année JANV. MARS 1954 numéro 1

POPULATION

REVUE TRIMESTRIELLE
DE L'INSTITUT NATIONAL
D'ÉTUDES DÉMOGRAPHIQUES

SOMMAIRE

ÉDITIONS DE L'I.N.E.D., 23, Av. F.-D.-ROOSEVELT-PARIS-8e



Part I: Implementing Datained

1. Metadata standardisation and enrichment
2. **Choose a software**
3. Going from Nesstar to NADA: the migration process

② Choose a software

The Dataverse® Project

- Developed by Harvard
- More and more widely used in social sciences
- Used by Center for socio-political data

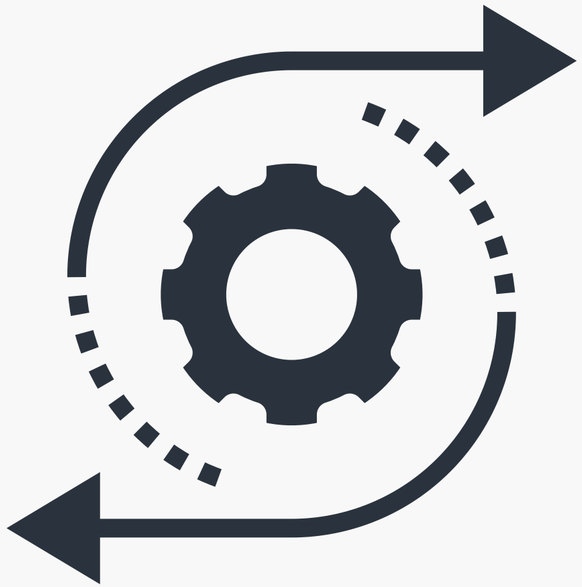
VS

- Developed by the World Bank Group
- Already used at INED for Demostaf

NADA
Microdata Cataloging Tool

2 Choose a software

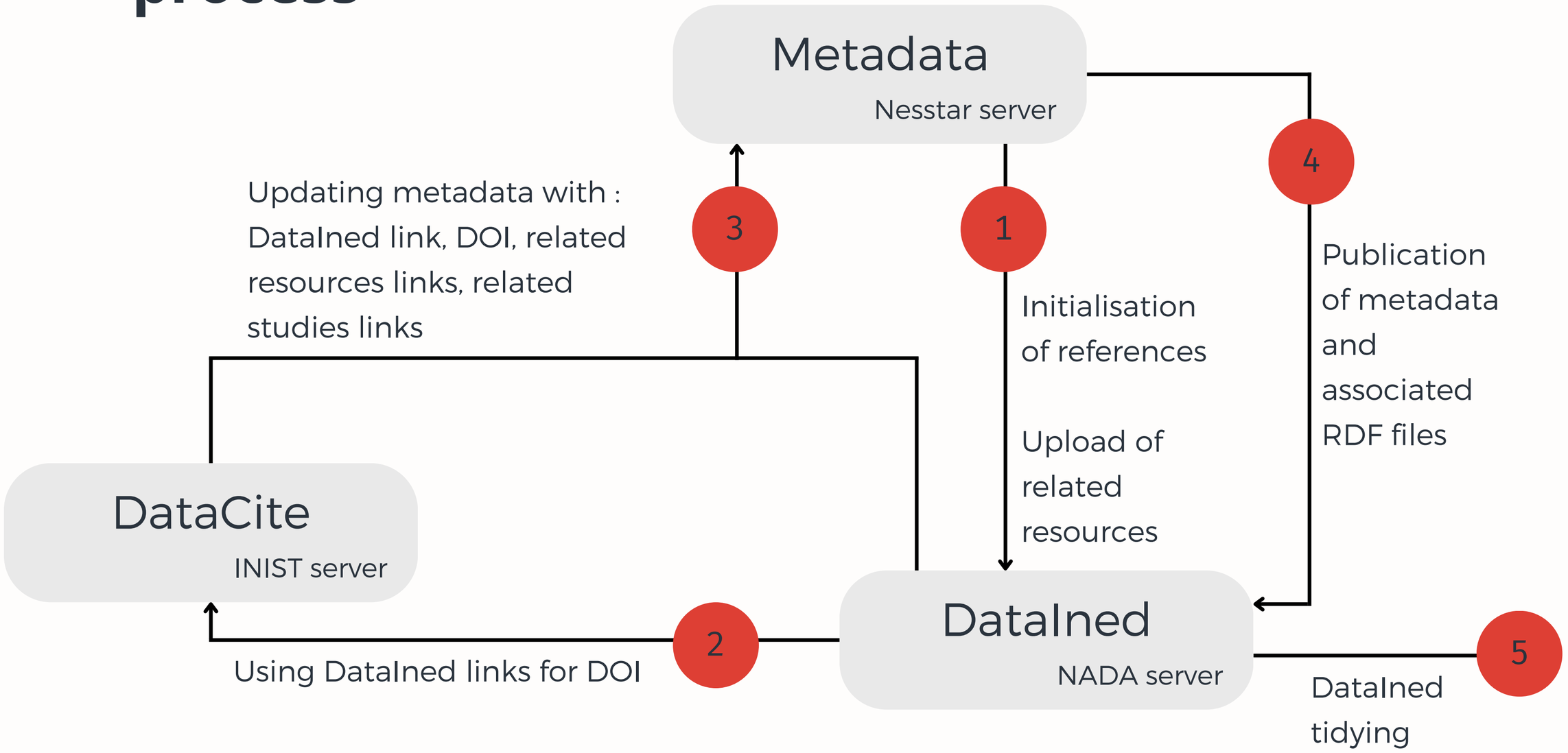
	Dataverse	Nada
Compatibility with / maintainability on INED's servers	-	++
Based on DDI	+	++
Variable Metadata (visible and searchable)	-	+
Used in Social Sciences	++	-
Allow to be harvested by CESSDA	+	+
User-friendly back-end	-	+
DOI integration	+	+



Part I: Implementing Datained

1. Metadata standardisation and enrichment
2. Choose a software
3. **Going from Nesstar to NADA: the migration process**

3 Going from Nesstar to Nada : the migration process



3 Going from Nesstar to NADA : the migration process

Migrating from Nesstar to NADA was a 5-steps process

1) Initialising NADA with empty references and migrating related resources in order to get their new urls

2) Creating the DOI for the relevant references (using the links obtained in the last step)

3) Adding back into the metadata the information obtained in the former steps, i.e. NADA links and DOI

3 Going from Nesstar to NADA : the migration process

Migrating from Nesstar to NADA was a 5-steps process

4) Publishing the updated metadata to NADA and an RDF file by references in order for NADA to link a reference to its related ressources

5) Doing all the NADA specific tidying up :

- Adding survey logos
- Linking surveys together
- Adding the DOI in the proper place
- Update data availability status
- Check countries

3 Going from Nesstar to NADA : the migration process

Softwares used at each step :

1. **NADA API** using the dedicated **R** package
2. **DataCite Fabrica**
3. **R** scripts and **XSLT**
4. **NADA API** using the dedicated **R** package
5. **NADA** administrator interface

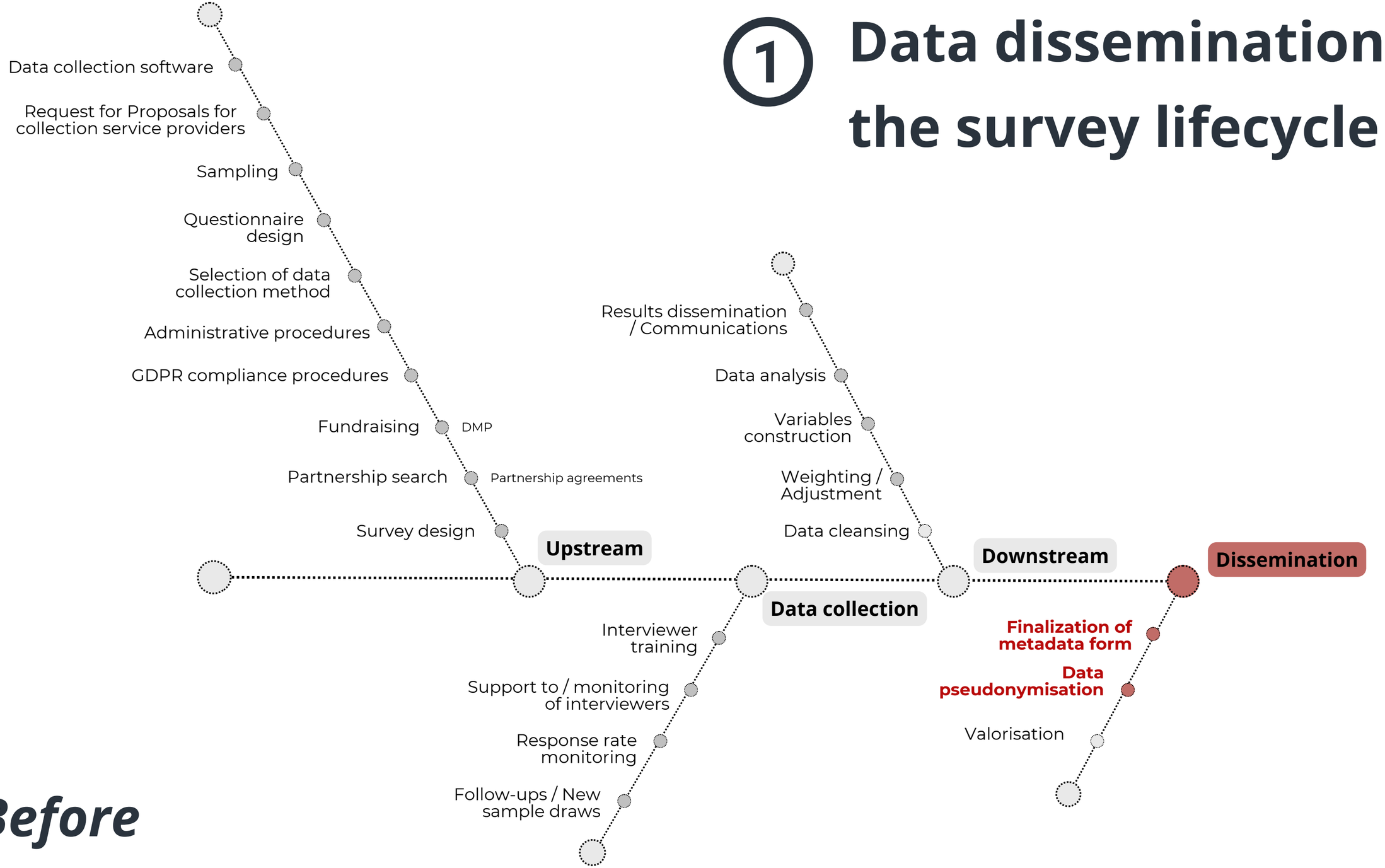
Part II: Rethinking the dissemination process

1. Data dissemination in the survey lifecycle
2. Our new workflow

Part II: Rethinking the dissemination process

1. **Data dissemination in the survey lifecycle**
2. Our new workflow

① Data dissemination in the survey lifecycle



Before

① Data dissemination in the survey lifecycle

Before

- An unanticipated and unplanned dissemination within conventions and administrative procedures
- Back and forth, redundancies, and information loss over time
- Scattered documentation
- Data constructed by multiple people, without harmonization

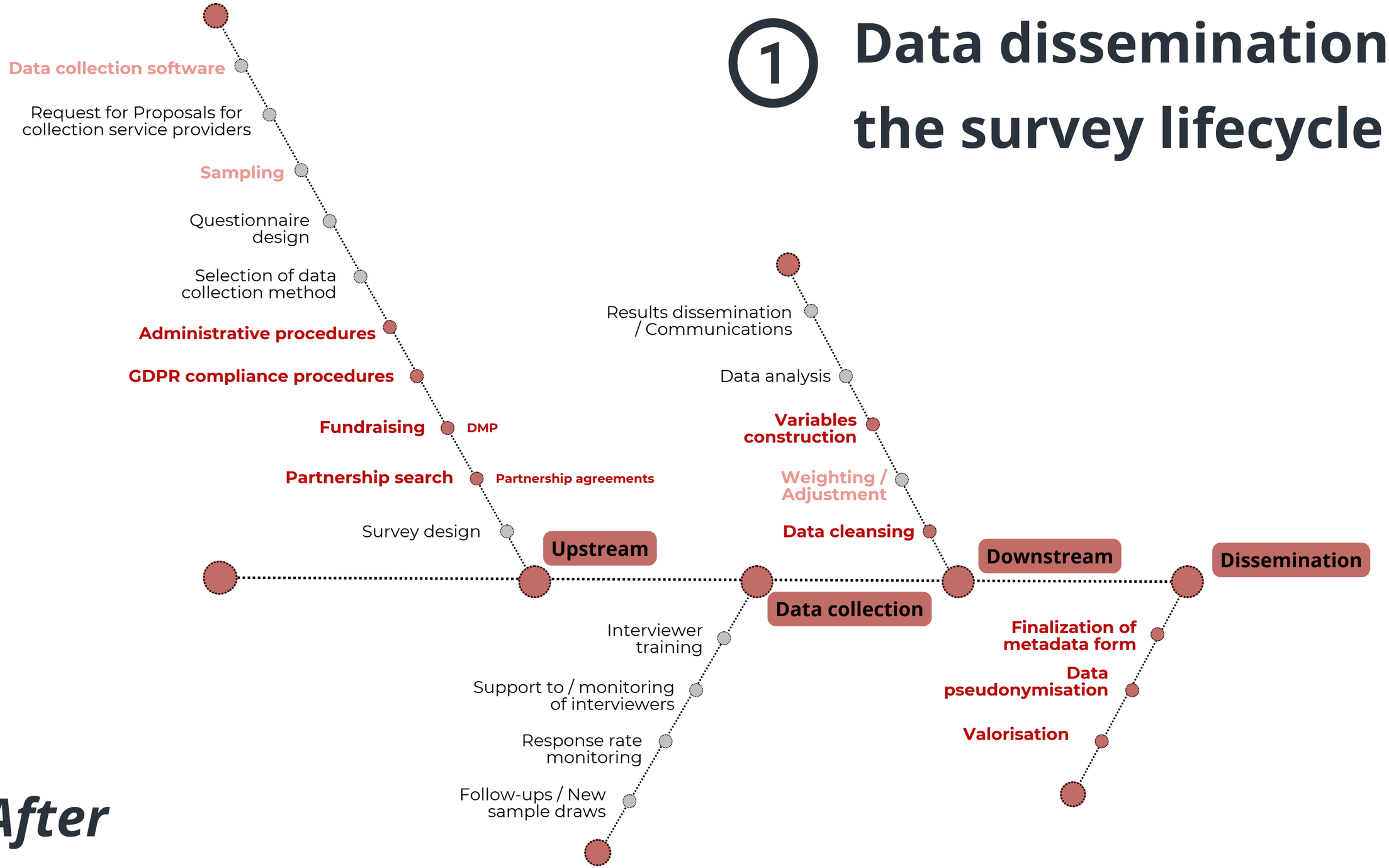
[g_ori_soc] Origine sociale - Groupe PCS Ménage (pcs ménage père et mère)	
I. Ménages à dominante cadre	278
II. Ménages à dominante intermédiaire (ou cadre)	520
III. Ménages à dominante employée (ou intermédiaire)	395
IV. Ménages à dominante petit indépendant	385
V. Ménages à dominante ouvrière	270
VI. Ménages monoactifs d'un-e employé-e ou ouvrier-ère	557
VII. Ménages d'inactif-ves	108

Expectation

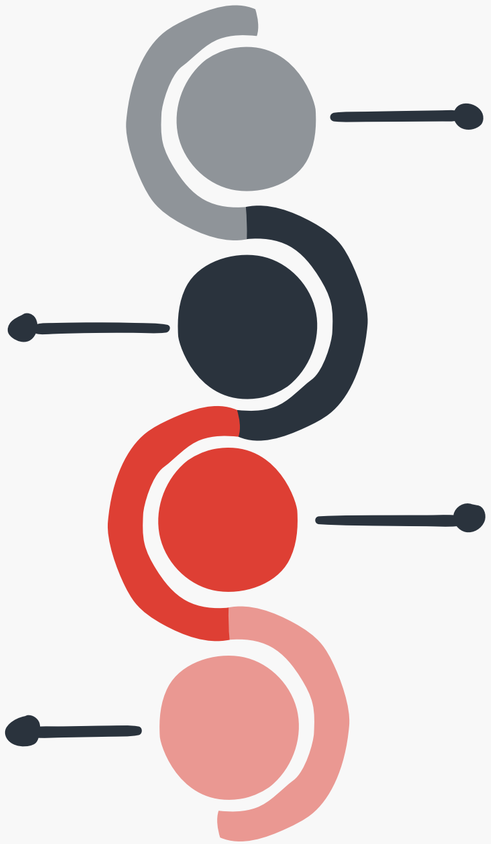
vs. Reality

g_ori_soc	
I	278
II	520
III	395
IV	385
V	270
VI	557
VII	108

① Data dissemination in the survey lifecycle



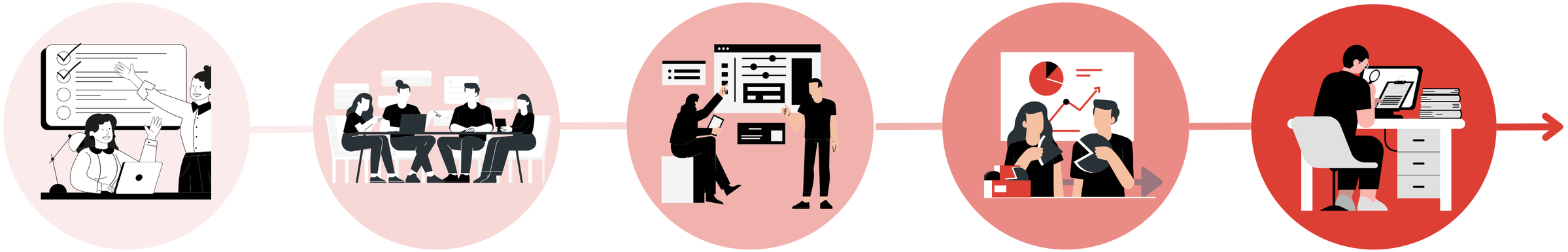
After



Part II: Rethinking the dissemination process

1. Data dissemination in the survey lifecycle
2. **Our new workflow**

2 Our new workflow



Meeting with research teams at the start of the data collection process to introduce data dissemination (workflow, documents, guides, form) and asks for finalised versions of questionnaires (paper and scripts)

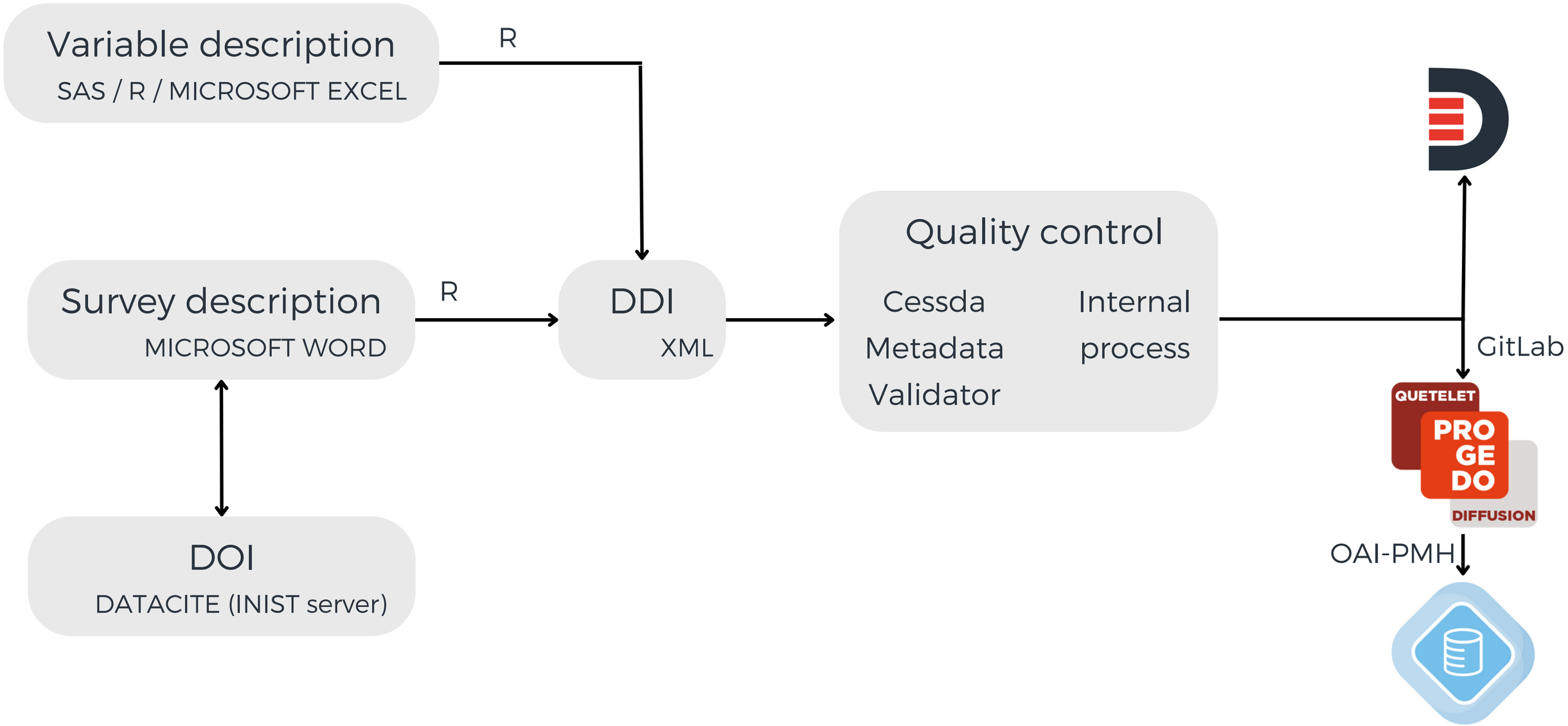
Research teams fill in information available to them (abstract, funding agencies, ...)

Support teams fill in information available to them (questionnaire, data collection process, ...)

Statisticians fill in information available to them (adjustements, weighting, response rate) and promote good practices

The data dissemination team coordinates this process, checks information, pseudonimises the data, ensures standard compliance and publishes data and metadata online

2 Our new workflow



Conclusion

A work in progress

Stay up to date with open science standards

Add new types of data to the catalogue

Make the catalogue accessible to English-speaking users

Useful resources 

PROGEDO Catalogue : https://data.progedo.fr

INED's DataLab website : https://datalab.site.ined.fr/

Access INED scientific use files : https://commande.progedo.fr

Guides and references : https://data.ined.fr/index.php/ressources

NADA's official website : https://nada.ihsn.org/

NADA's github : https://github.com/ihsn/nada

Cessda controlled vocabularies : https://vocabularies.cessda.eu/

Dissemination scripts : https://github.com/JulieLen16/Ined_Scripts

Thank you !

julie.lenoir@ined.fr

julie.baron@ined.fr

License