14 janvier 2020

Entre PCS et Rome : classification automatique des métiers

Repérage de métiers dans l'EAR et l'ECC et classification d'offres d'emploi

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DU TRAVAIL

DARES
direction de l'animation de la recherche, des études et des statistiques

STATISTIQUE PUBLIQUE

dares.travail-emploi.gouv.fr

Alexis Eidelman
Claire de Maricourt

Dares - Département Analyse des métiers et emploi des travailleurs handicapés

DARES

PLAN DE LA PRÉSENTATION

I. Repérer les métiers du numérique
   1. Nomenclature de métiers du numérique
   2. Identification dans l'EAR et l'EEC
   3. Résultats

II. Classer des offres d'emploi par code ROME
   1. Nomenclature et données
   2. Nettoyage et classification
   3. Résultats

DARES

REPÉRER LES MÉTIERS DU NUMÉRIQUE

Dans le cadre de l'Insee Référence : L'économie et la société à l'ère du numérique

1. Nomenclature et données

- Objectif : repérer les **métiers** du numérique dans l'EAR (recensement) et l'EEC (enquête emploi). Ce n'est pas une approche secteur.

DARES

1. Nomenclature et données

- Objectif : repérer les **métiers** du numérique dans l'EAR (recensement) et l'EEC (enquête emploi). Ce n'est pas une approche secteur.
- Les métiers sont plus ou moins concernés par le numérique

DARES

1. Nomenclature et données

- Objectif : repérer les **métiers** du numérique dans l'EAR (recensement) et l'EEC (enquête emploi). Ce n'est pas une approche secteur.
- Les métiers sont plus ou moins concernés par le numérique

Segment cœur du numérique

Développeur informatique

Ingénieur télécoms

Technicien informatique

DARES

1. Nomenclature et données

- Objectif : repérer les **métiers** du numérique dans l'EAR (recensement) et l'EEC (enquête emploi). Ce n'est pas une approche secteur.
- Les métiers sont plus ou moins concernés par le numérique

➡ Segment cœur du numérique

Développeur informatique

Ingénieur télécoms

Technicien informatique

➡ Segment diffus

Infographiste ***Data scientist***

Webmaster ***Community manager***

DARES

1. Nomenclature et données

- Objectif : repérer les **métiers** du numérique dans l'EAR (recensement) et l'EEC (enquête emploi). Ce n'est pas une approche secteur.
- Les métiers sont plus ou moins concernés par le numérique

➡ Segment cœur du numérique

Développeur informatique

Ingénieur télécoms

Technicien informatique

➡ Segment diffus

Infographiste ***Data scientist***

Webmaster ***Community manager***

➡ Segment périphérique

Secrétaire

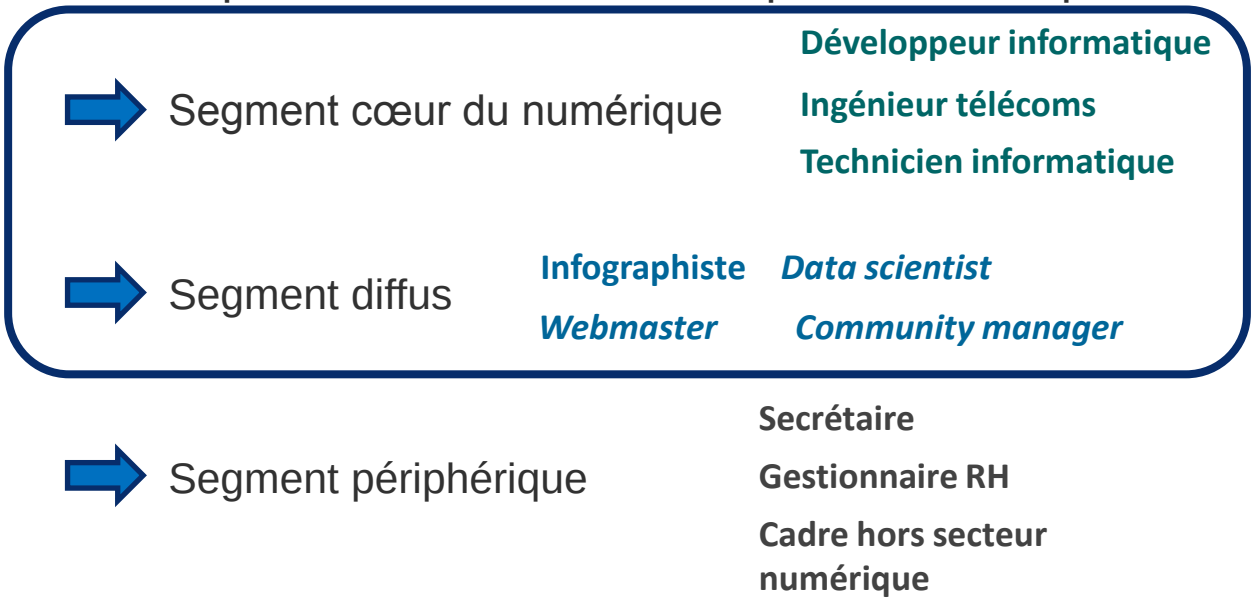
Gestionnaire RH

Cadre hors secteur numérique

DARES

1. Nomenclature et données

- Objectif : repérer les **métiers** du numérique dans l'EAR (recensement) et l'EEC (enquête emploi). Ce n'est pas une approche secteur.
- Les métiers sont plus ou moins concernés par le numérique



DARES

1. Nomenclature et données

- Métiers du numérique dans l'EAR

Verbatim	Code PCS	PCS	Numérique
Chargé de communication digitale	464a	Assistants de la publicité, des relations publiques (indépendants ou salariés)	True
Web reporter	352a	Journalistes (y. c. rédacteurs en chef)	True
Gestionnaire de stock	551a	Employés de libre service du commerce et magasiniers	False
Attaché de presse	464a	Assistants de la publicité, des relations publiques (indépendants ou salariés)	False
Employée e-commerciale	551a	Employés de libre service du commerce et magasiniers	True
Secrétaire de rédaction	352a	Journalistes (y. c. rédacteurs en chef)	False

DARES

1. Nomenclature et données

- Métiers du numérique dans l'EAR

Verbatim	Code PCS	PCS	Numérique
Chargé de communication digitale	464a	Assistants de la publicité, des relations publiques (indépendants ou salariés)	True
Web reporter	352a	Journalistes (y. c. rédacteurs en chef)	True
Gestionnaire de stock	551a	Employés de libre service du commerce et magasiniers	False
Attaché de presse	464a	Assistants de la publicité, des relations publiques (indépendants ou salariés)	False
Employée e-commerciale	551a	Employés de libre service du commerce et magasiniers	True
Secrétaire de rédaction	352a	Journalistes (y. c. rédacteurs en chef)	False

→ Nécessité d'un référentiel spécifique

DARES

1. Nomenclature et données

- Nouvelle nomenclature pour les professions du « numérique » dans le cadre de la **refonte des PCS : agrégat *ad hoc* des professions du numérique** réalisé par Christophe Michel (Dares) et Jean Flamand (France Stratégie)

DARES

1. Nomenclature et données

- Nouvelle nomenclature pour les professions du « numérique » dans le cadre de la **refonte des PCS : agrégat *ad hoc* des professions du numérique** réalisé par Christophe Michel (Dares) et Jean Flamand (France Stratégie)

- Utilisation de **répertoires professionnels** existants :
  - Cigref
  - France Stratégie/ Céreq
  - Opiiec

DARES

1. Nomenclature et données

- Nouvelle nomenclature pour les professions du « numérique » dans le cadre de la **refonte des PCS : agrégat *ad hoc* des professions du numérique** réalisé par Christophe Michel (Dares) et Jean Flamand (France Stratégie)

- Utilisation de **répertoires professionnels** existants :
  - Cigref
  - France Stratégie/ Céreq
  - Opiiec

- Enrichissement avec les **libellés vus dans les enquêtes**

DARES

1. Nomenclature et données

- 885 appellations identifiées (dont des libellés « vagues » : architecte, chargé de communication, …)

DARES

1. Nomenclature et données

- 885 appellations identifiées (dont des libellés « vagues » : architecte, chargé de communication, …)
- 69 sous-familles

DARES

1. Nomenclature et données

- 885 appellations identifiées (dont des libellés « vagues » : architecte, chargé de communication, …)
- 69 sous-familles
- 7 familles du numérique

1. Nomenclature et données

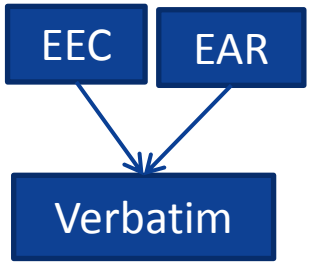
- 885 appellations identifiées (dont des libellés « vagues » : architecte, chargé de communication, …)
- 69 sous-familles
- 7 familles du numérique

- 90 libellés black-listés (« ingénieur réseaux eau », « technicien de réseau SNCF », « développeur de formations », …)

2. Identification dans l’EAR et l’EEC

DARES

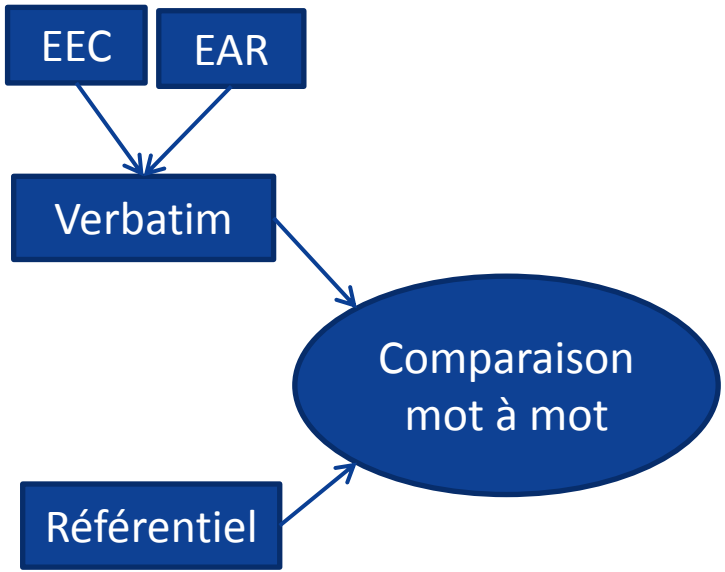
2. Identification dans l'EAR et l'EEC



Référentiel

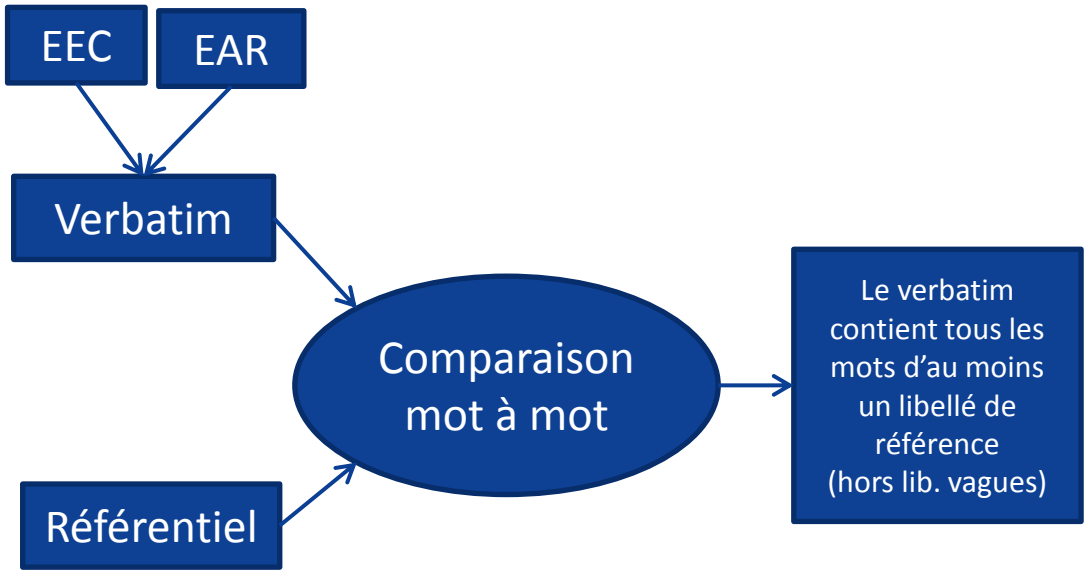
DARES

2. Identification dans l'EAR et l'EEC



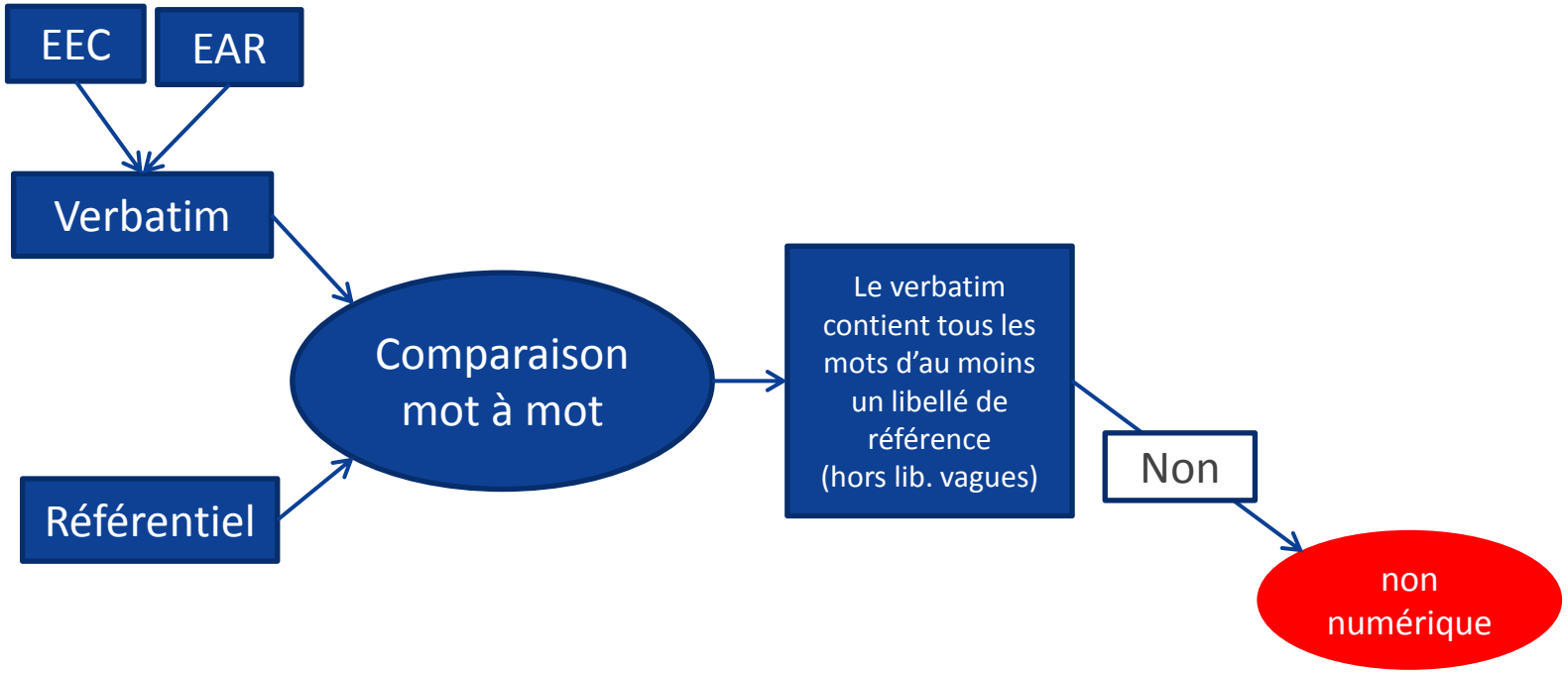
DARES

2. Identification dans l'EAR et l'EEC



DARES

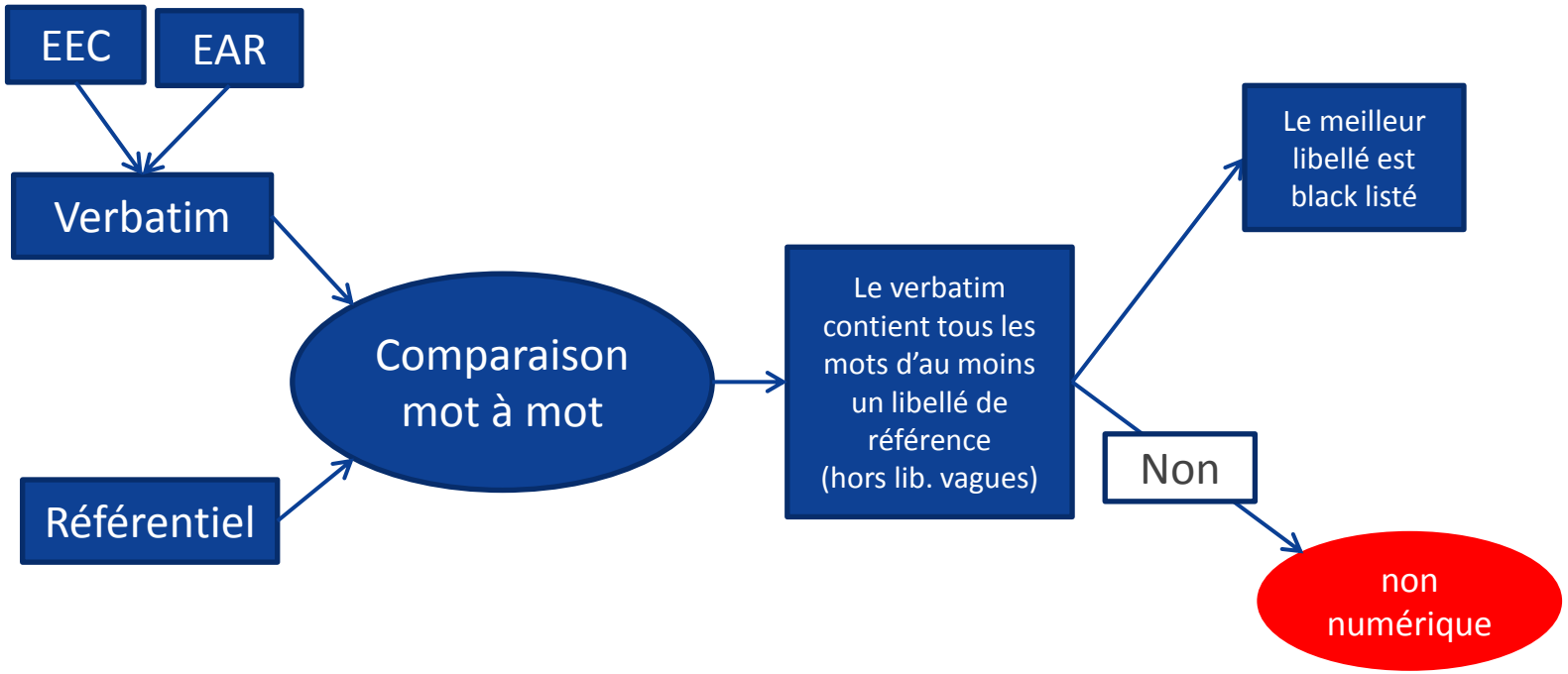
2. Identification dans l'EAR et l'EEC



Secrétaire
Chargé de communication
Ingénieur
Ingénieur réseaux eau

DARES

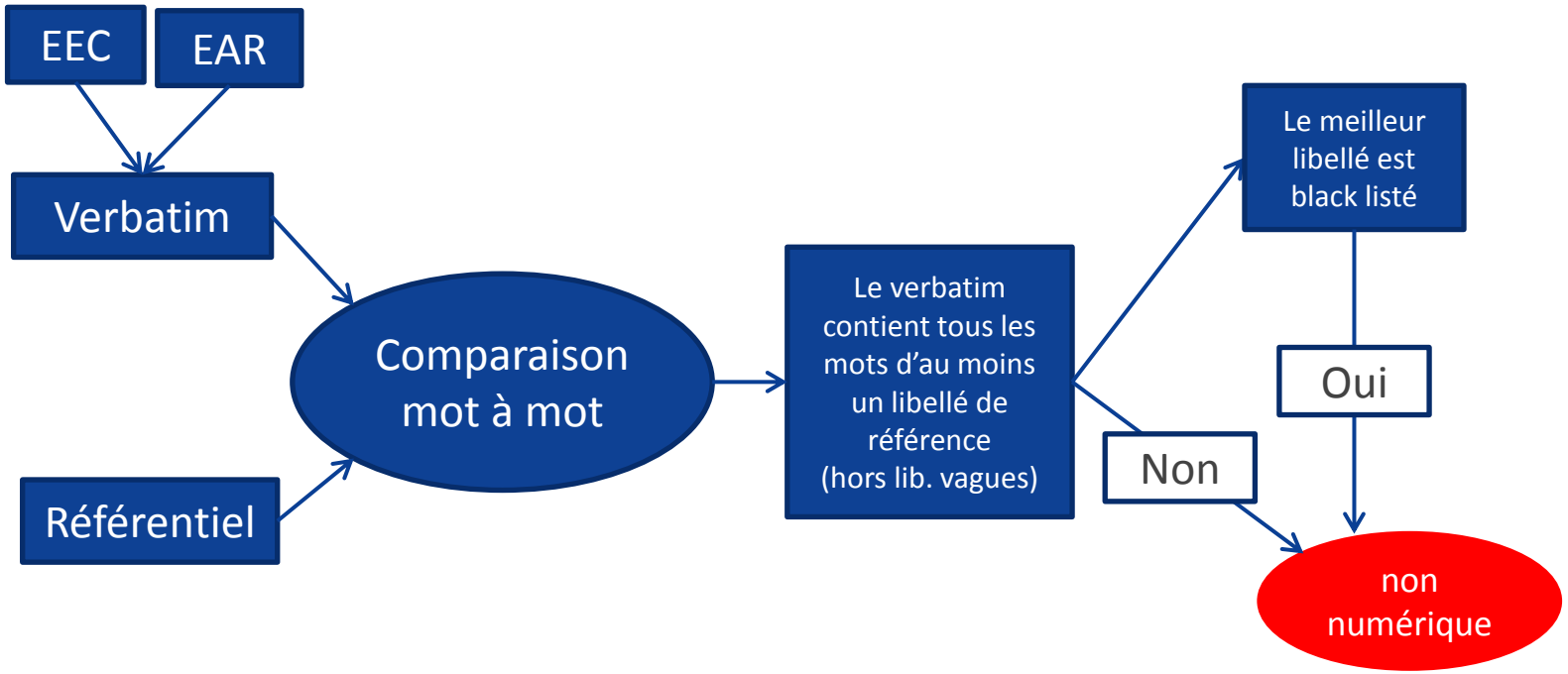
2. Identification dans l'EAR et l'EEC



Secrétaire
Chargé de communication
Ingénieur
Ingénieur réseaux eau

DARES

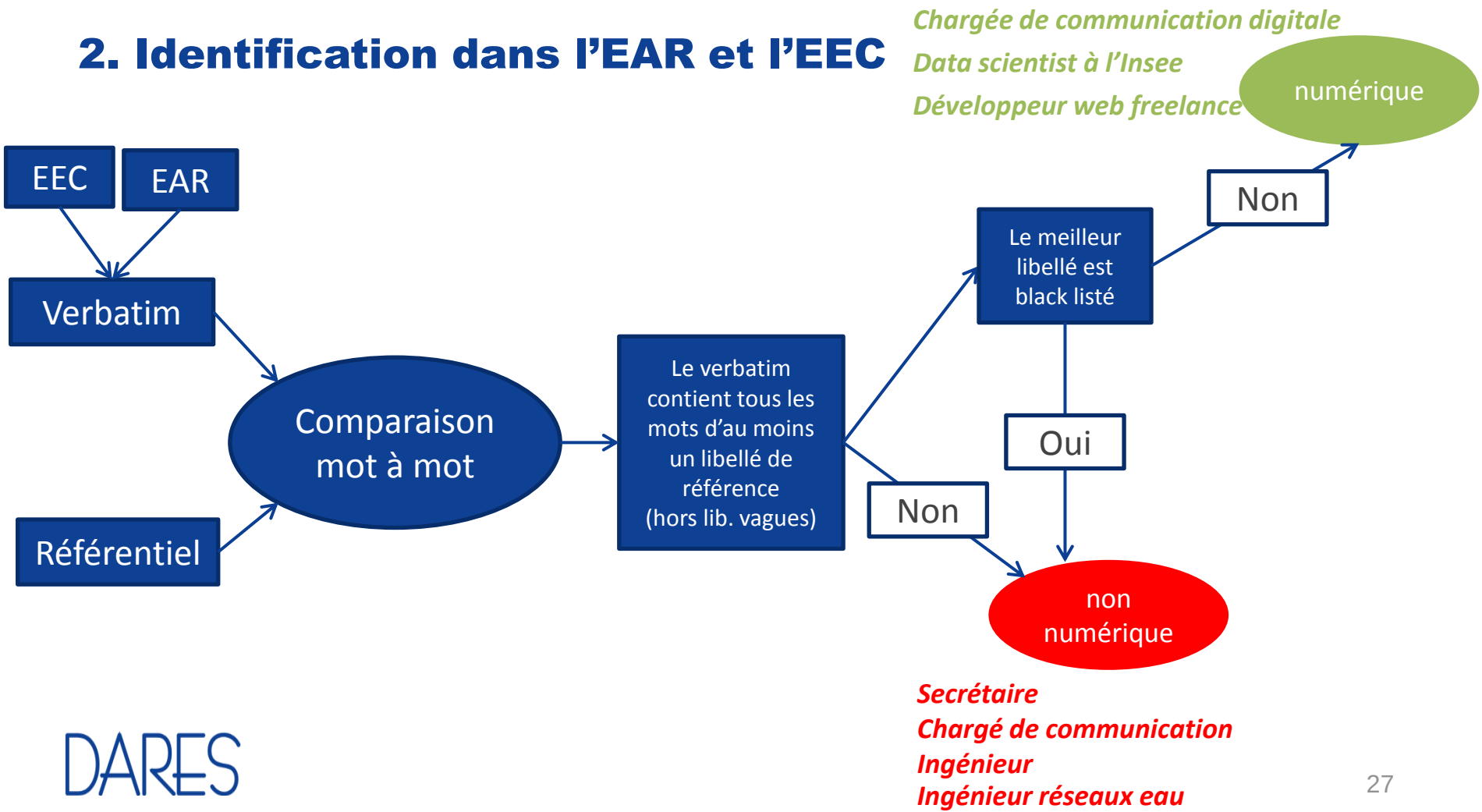
2. Identification dans l'EAR et l'EEC



Secrétaire
Chargé de communication
Ingénieur
Ingénieur réseaux eau

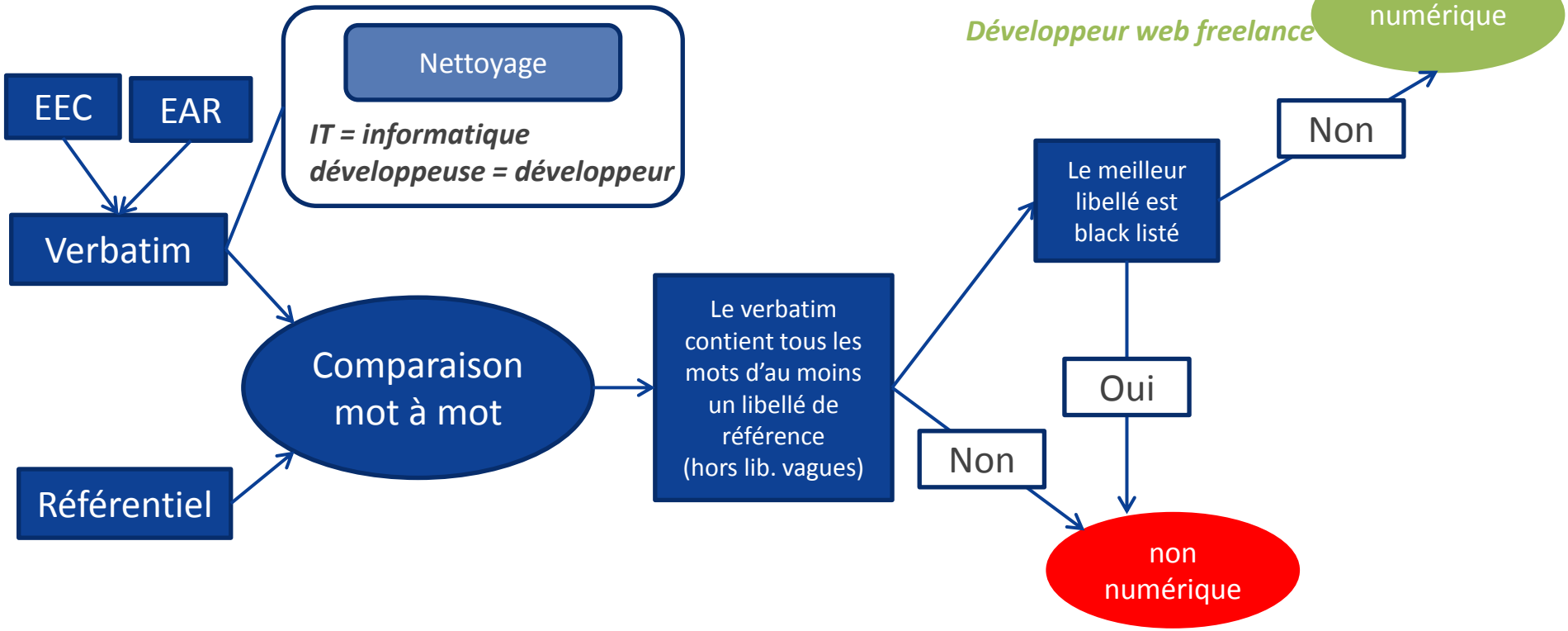
DARES

2. Identification dans l'EAR et l'EEC



DARES

2. Identification dans l'EAR et l'EEC

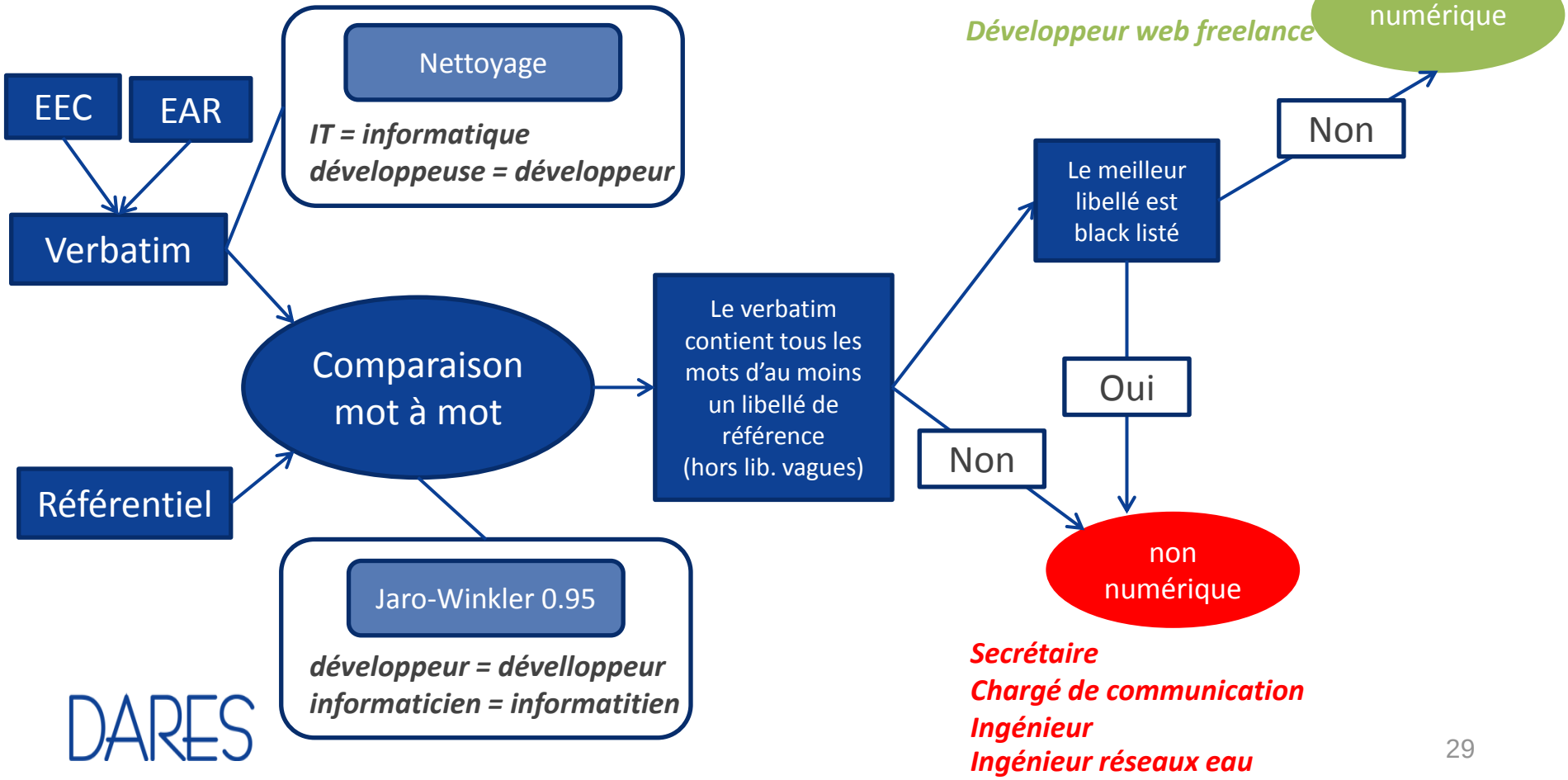


Chargée de communication digitale
Data scientist à l'Insee
Développeur web freelance

Secrétaire
Chargé de communication
Ingénieur
Ingénieur réseaux eau

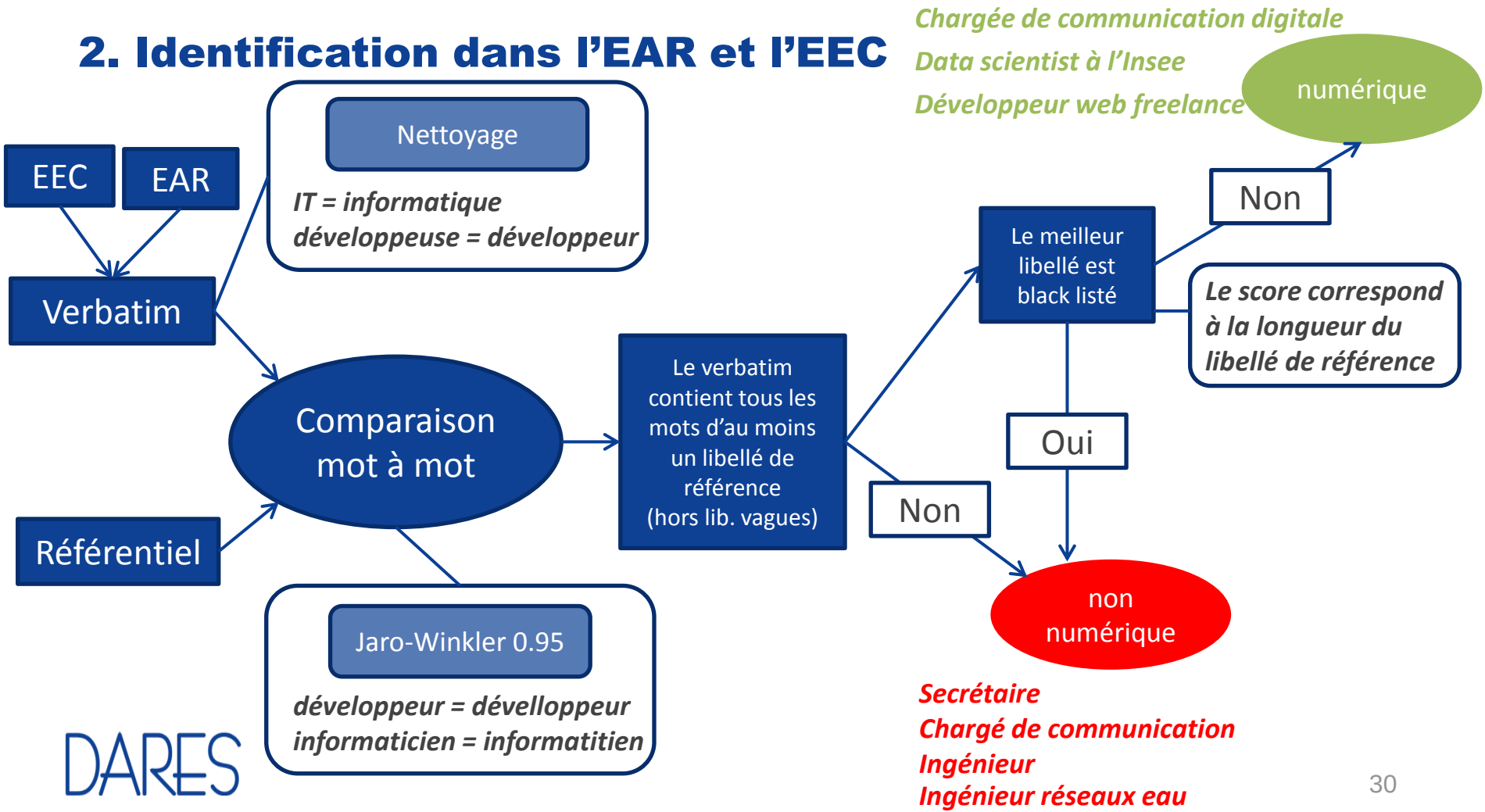
DARES

2. Identification dans l'EAR et l'EEC



DARES

2. Identification dans l'EAR et l'EEC



DARES

2. Identification dans l'EAR et l'EEC

« Ingénieur réseaux télécoms chez Orange »

Libellés	Match	Score
libéllés numériques		
Chargé de com digitale	0	0
Ingénieur réseaux télecoms	1	3
Ingénieur télécoms	1	2
Ingénieur réseaux	1	2
libellés vagues		
Ingénieur	1	0
Informaticien	0	0
libellés black listés		
Ingénieur btp	0	0
Ingénieur réseaux eau	0	0

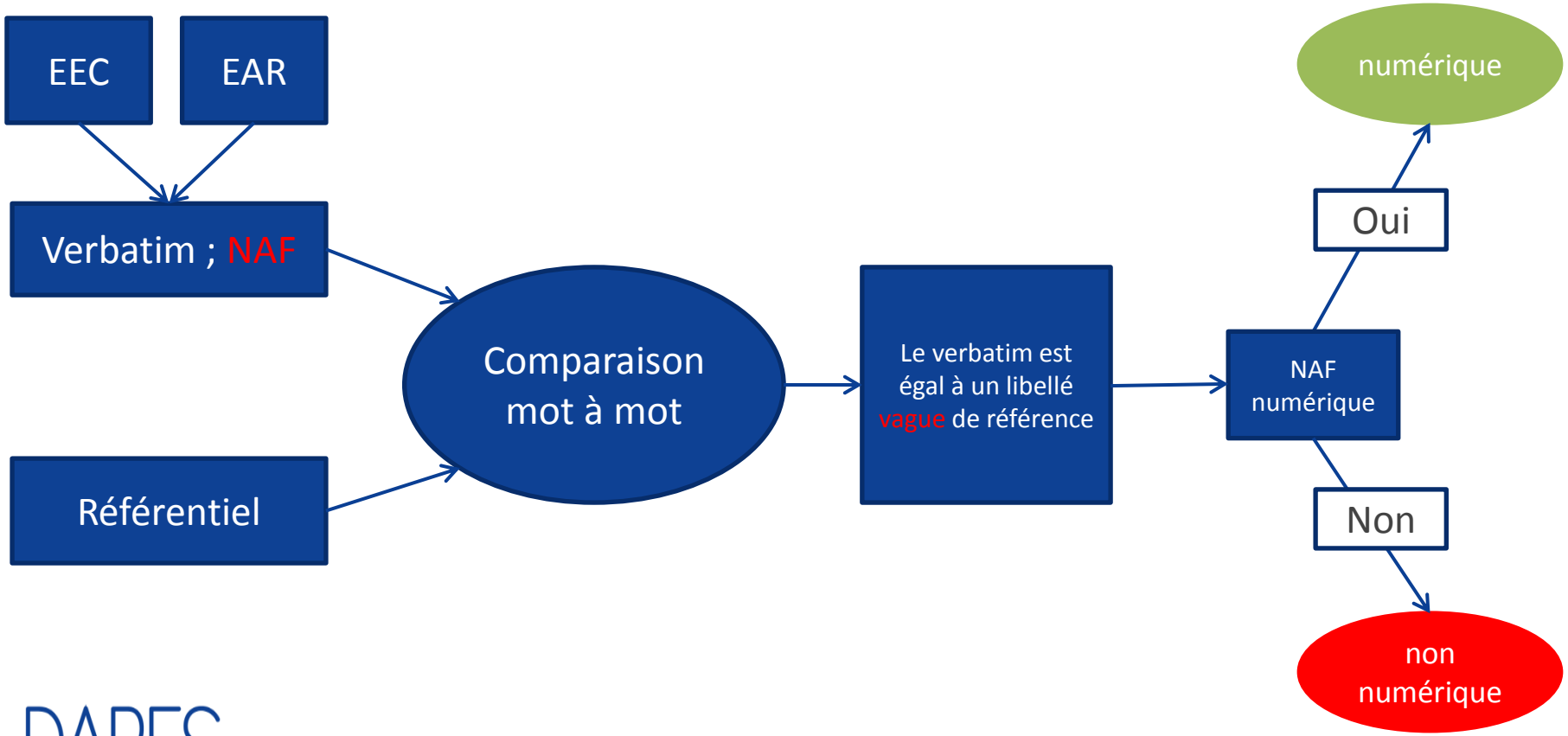
« Ingénieur réseaux eaux usées »

Libellés	Match	Score
libéllés numériques		
Chargé de com digitale	0	0
Ingénieur réseaux télecoms	0	0
Ingénieur télécoms	0	0
Ingénieur réseaux	1	2
libellés vagues		
Ingénieur	1	0
Informaticien	0	0
libellés black listés		
Ingénieur btp	0	0
Ingénieur réseaux eau	1	3

DARES

2. Identification dans l'EAR et l'EEC

Architecte ***dans une SSII***
Chef de projet ***dans les Télécoms***



DARES

Ingénieur ***dans le BTP***

3. Résultats

- Pas de benchmark mais …

DARES

3. Résultats

- Pas de benchmark mais …

- Sous-estimation probable :

DARES

3. Résultats

- Pas de benchmark mais …

- Sous-estimation probable :
  - peu de correction de fautes d'orthographe

DARES

3. Résultats

- Pas de benchmark mais …

- Sous-estimation probable :

  - peu de correction de fautes d'orthographe

  - pas de référentiel spécifique pour la fonction publique (au-delà des cas de verbatims qui correspondent à des grades)

DARES

3. Résultats

- Pas de benchmark mais …

- Sous-estimation probable :

  - peu de correction de fautes d'orthographe

  - pas de référentiel spécifique pour la fonction publique (au-delà des cas de verbatims qui correspondent à des grades)

  - vocabulaire de certains de ces métiers qui évoluent rapidement

DARES

3. Résultats

- Pas de benchmark mais …

- Sous-estimation probable :
  - peu de correction de fautes d'orthographe
  - pas de référentiel spécifique pour la fonction publique (au-delà des cas de verbatims qui correspondent à des grades)
  - vocabulaire de certains de ces métiers qui évoluent rapidement
  - référentiel féminisée « à la main »

DARES

3. Résultats

- Pas de benchmark mais …

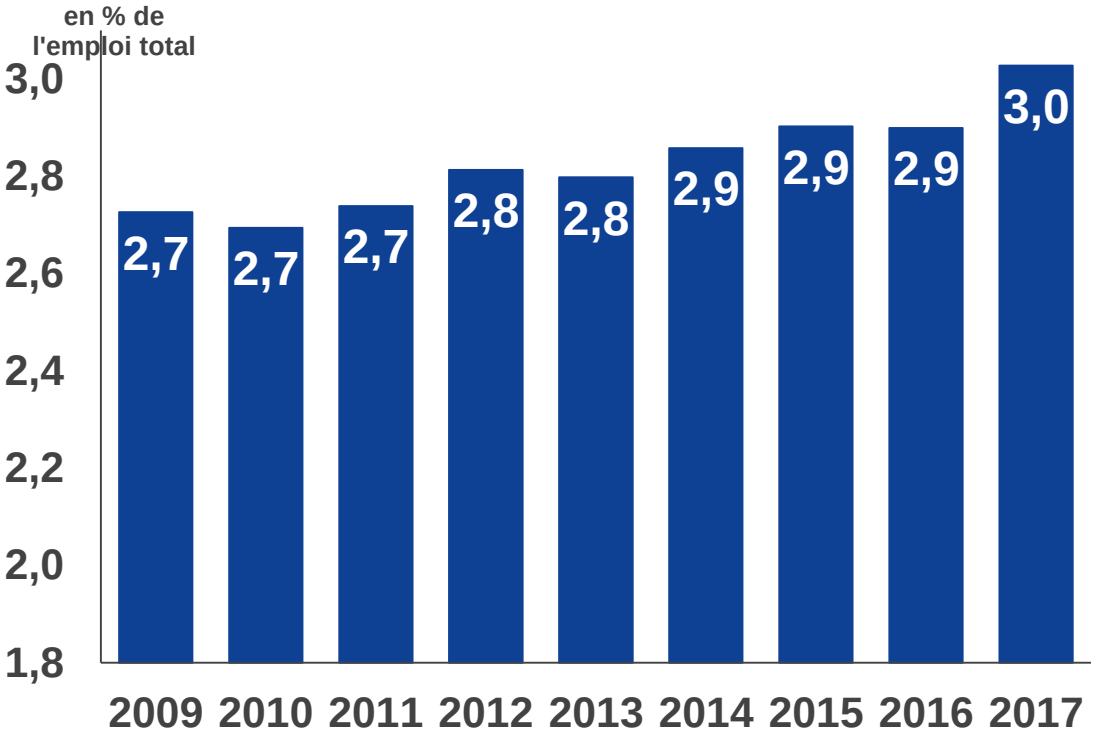
- Sous-estimation probable :
  - peu de correction de fautes d'orthographe
  - pas de référentiel spécifique pour la fonction publique (au-delà des cas de verbatims qui correspondent à des grades)
  - vocabulaire de certains de ces métiers qui évoluent rapidement
  - référentiel féminisée « à la main »

- Plus de diversité de libellés dans l'EAR (vs EEC) : effet enquêteur ?

DARES

3. Résultats

- Analyse de l'évolution des métiers du numérique

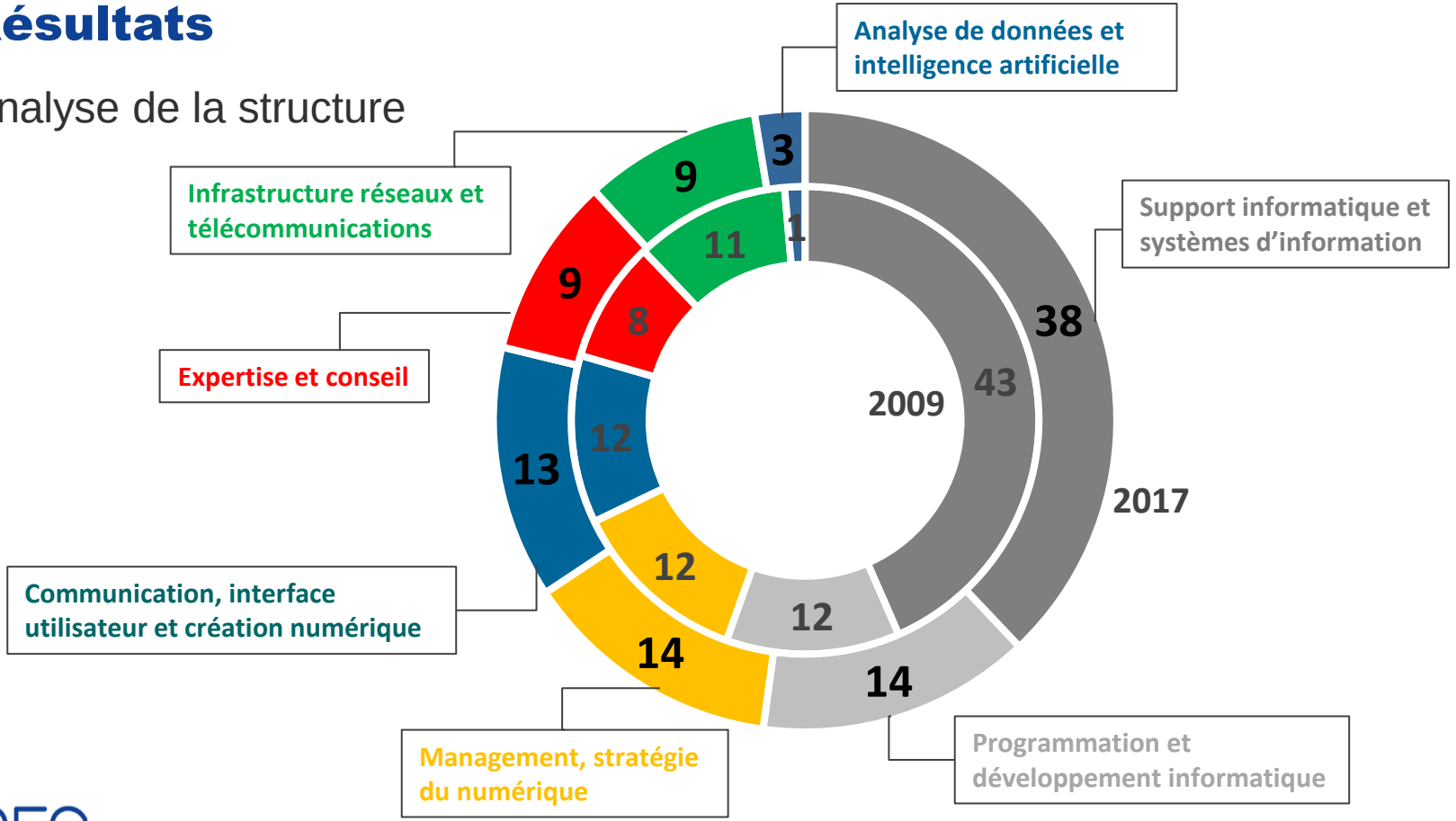


DARES

Insee Référence, L'économie et la société à l'ère du numérique

3. Résultats

- Analyse de la structure



DARES

Insee Référence, L'économie et la société à l'ère du numérique

CLASSER DES OFFRES D'EMPLOI PAR CODE ROME

DARES

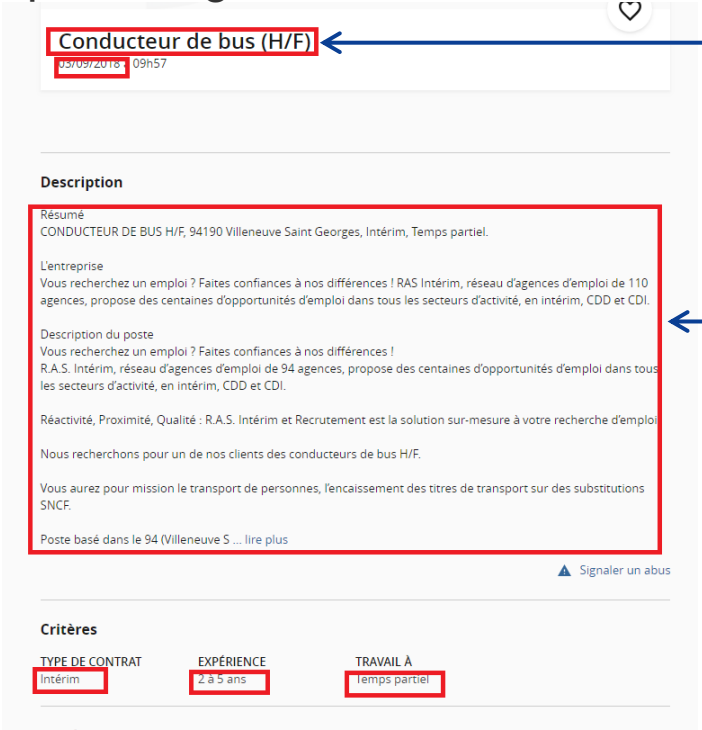
1. Nomenclature et données

- Objectif : annoter par métier des offres scrapées en ligne, pour pouvoir les intégrer aux calculs d'indicateurs statistiques

DARES

1. Nomenclature et données

- Objectif : annoter par métier des offres scrapées en ligne, pour pouvoir les intégrer aux calculs d'indicateurs statistiques
- Les offres d'emploi en ligne



1. Nomenclature et données

- La nomenclature ROME

K			**SERVICES A LA PERSONNE ET A LA COLLECTIVITE**
K	**21**		**Formation initiale et continue**
K	**21**	**08**	**Enseignement supérieur**
K	21	08	Attaché / Attachée temporaire d'enseignement et de recherche -ATER-
K	21	08	Chargé / Chargée de cours
K	21	08	Chargé / Chargée d'enseignement
K	21	08	Chargé / Chargée d'enseignement du supérieur
K	21	08	Chef de département de l'enseignement supérieur
K	21	08	Enseignant-chercheur / Enseignante-chercheuse
K	21	08	Lecteur / Lectrice de langues dans l'enseignement supérieur
K	21	08	Maître / Maîtresse de conférences
K	21	08	Maître / Maîtresse de langues dans l'enseignement supérieur
K	21	08	Moniteur / Monitrice d'initiation à l'enseignement supérieur
K	21	08	Professeur / Professeure de l'enseignement supérieur
K	21	08	Professeur / Professeure des universités
K	21	08	Responsable filière enseignement supérieur
K	21	08	Responsable matière enseignement supérieur
K	21	08	Responsable programme enseignement supérieur

Famille de métiers
Domaine professionnel
Intitulé du code ROME
Appelations

DARES

1. Nomenclature et données

- Les offres de Pôle emploi (via API Pôle emploi)


```
{
      "id": "096VSPD",
      "intitule": "Consultant / Consultante en organisation et
management    (H/F)",
      "appellationlibelle": "Consultant / Consultante en
organisation et management",
      "romeCode": "M1402",
      "romeLibelle": "Conseil en organisation et management
d'entreprise",
      "description": "Société européenne de consulting, leader
en maintenance et asset management intervenant principalement
pour des ETI, des grands groupes internationaux, dans 3 secteurs
: usines, flottes, infrastructures (énergie, port ). A l'issue
d'une période de familiarisation avec la méthodologie et
l'approche de cette société, vos objectifs seront centrés sur le
développement de l'activité en France et en pays francophone
(hors Belgique). Ce poste nécessite d'associer une expérience de
développeur d'affaires acquise dans le consulting ou en
entreprise avec une pratique de la maintenance en milieu
industriel.Vos activités principales : fidéliser le portefeuille
clientèle existant,développer le marché, renforcer l'image et la
notoriété de la société, piloter ou réaliser les projets,
recruter et développer une équipe de salariés et de partenaires
expérimentés, capitaliser les retours d'expérience.",
      "lieuTravail_codePostal": 75006
}
```


DARES

1. Nomenclature et données

- Les offres de Pôle emploi (via API Pôle emploi)


```
{
      "id": "096VSPD",
      "intitule": "Consultant / Consultante en organisation et
management    (H/F)",
      "appellationlibelle": "Consultant / Consultante en
organisation et management",
      "romeCode": "M1402",
      "romeLibelle": "Conseil en organisation et management
d'entreprise",
      "description": "Société européenne de consulting, leader
en maintenance et asset management intervenant principalement
pour des ETI, des grands groupes internationaux, dans 3 secteurs
: usines, flottes, infrastructures (énergie, port ). A l'issue
d'une période de familiarisation avec la méthodologie et
l'approche de cette société, vos objectifs seront centrés sur le
développement de l'activité en France et en pays francophone
(hors Belgique). Ce poste nécessite d'associer une expérience de
développeur d'affaires acquise dans le consulting ou en
entreprise avec une pratique de la maintenance en milieu
industriel.Vos activités principales : fidéliser le portefeuille
clientèle existant,développer le marché, renforcer l'image et la
notoriété de la société, piloter ou réaliser les projets,
recruter et développer une équipe de salariés et de partenaires
expérimentés, capitaliser les retours d'expérience.",
      "lieuTravail_codePostal": 75006
}
```


Offres structurées et annotées :

- Evaluation
- Entraînement

DARES

1. Nomenclature et données

- Séparation des données

  - Offres de Pôle emploi directement publiées sur leur site
    ➡ **Données Pôle emploi**

  - Offres diffusées sur le site de Pôle emploi transmises par les partenaires
    ➡ **Données partenaires**

  - Offres scrapées par la Dares
    ➡ **Données scrapées**

DARES

2. Nettoyage et classification

- Nettoyage des titres d'offres d'emploi

Étapes du nettoyage	Exemple
Tout mettre en minuscules	employé.e commercial au rayon bazar à lyon (69) cdi h/f
Enlever la ponctuation	employée commercial au rayon bazar à lyon 69 cdi hf
Enlever les chiffres	employée commercial au rayon bazar à lyon cdi hf
Enlever les *stopwords*	employée commercial rayon bazar lyon cdi hf
Enlever certains mots	employée commercial rayon bazar lyon
Enlever le nom de la ville	employée commercial rayon bazar
Lemmatiser	employé commercial rayon bazar

DARES

2. Nettoyage et classification

- Nettoyage de la nomenclature pour extraire des mots féminisés et des abréviations

K			**SERVICES A LA PERSONNE ET A LA COLLECTIVITE**
K	**21**		**Formation initiale et continue**
K	**21**	**08**	**Enseignement supérieur**
K	21	08	Attaché / Attachée temporaire d'enseignement et de recherche -ATER-
K	21	08	Chargé / Chargée de cours

DARES

2. Nettoyage et classification

- Nettoyage de la nomenclature pour extraire des mots féminisés et des abréviations

K			**SERVICES A LA PERSONNE ET A LA COLLECTIVITE**
K	**21**		**Formation initiale et continue**
K	**21**	**08**	**Enseignement supérieur**
K	21	08	Attaché / Attachée temporaire d'enseignement et de recherche -ATER-
K	21	08	Chargé / Chargée de cours

Attaché temporaire d’enseignement et de recherche

DARES

2. Nettoyage et classification

- Nettoyage de la nomenclature pour extraire des mots féminisés et des abréviations

K			**SERVICES A LA PERSONNE ET A LA COLLECTIVITE**
K	**21**		**Formation initiale et continue**
K	**21**	**08**	**Enseignement supérieur**
K	21	08	Attaché / Attachée temporaire d'enseignement et de recherche -ATER-
K	21	08	Chargé / Chargée de cours

Attaché temporaire d’enseignement et de recherche
Attachée temporaire d’enseignement et de recherche

2. Nettoyage et classification

- Nettoyage de la nomenclature pour extraire des mots féminisés et des abréviations

K			**SERVICES A LA PERSONNE ET A LA COLLECTIVITE**
K	**21**		**Formation initiale et continue**
K	**21**	**08**	**Enseignement supérieur**
K	21	08	Attaché / Attachée temporaire d'enseignement et de recherche -ATER-
K	21	08	Chargé / Chargée de cours

Attaché temporaire d’enseignement et de recherche
Attachée temporaire d’enseignement et de recherche
ATER

DARES

2. Nettoyage et classification

- Test de plusieurs méthodes :
    - Similarité entre les chaînes de caractères (Levenshtein, Jaro-Winkler)
    - Tf-Idf et similarité cosinus
    - SVM linéaire
    - Word2Vec et SVM linéaire
    - Combinaison de ces méthodes

DARES

2. Nettoyage et classification

- Test de plusieurs méthodes :
  - Similarité entre les chaînes de caractères (Levenshtein, Jaro-Winkler)
  - Tf-Idf et similarité cosinus
  - **SVM linéaire**
  - Word2Vec et SVM linéaire
  - Combinaison de ces méthodes

DARES

3. Résultats

- Sur les données de test Pôle Emploi au niveau :

Appellations

	f1-score	precision	recall	support
micro avg	0.92	0.92	0.92	303585
macro avg	0.83	0.86	0.81	303585
weighted avg	0.92	0.92	0.92	303585

Domaines professionnels

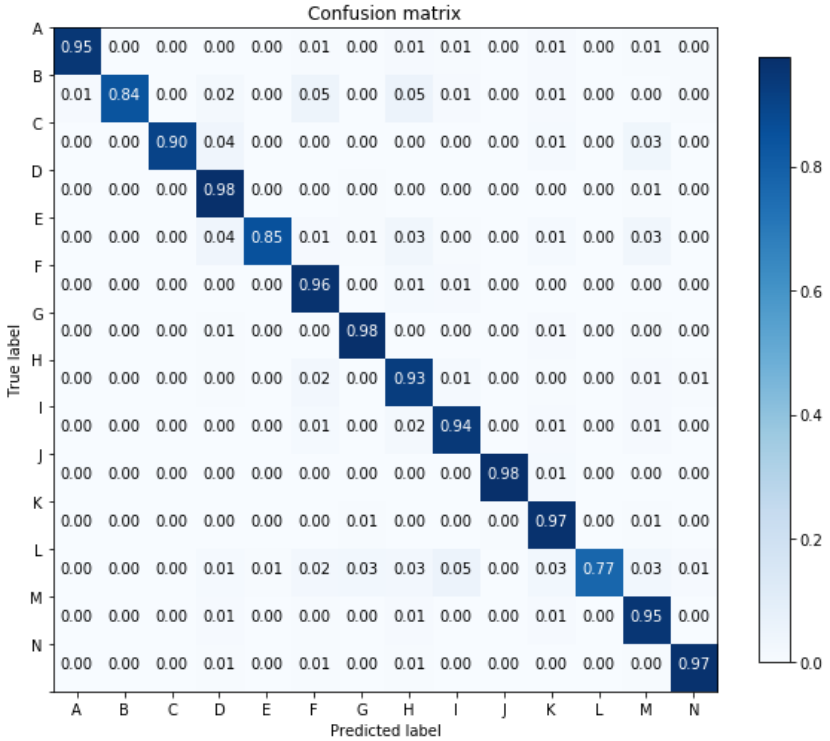
	f1-score	precision	recall	support
micro avg	0.94	0.94	0.95	303585
macro avg	0.88	0.89	0.86	303585
weighted avg	0.94	0.94	0.94	303585

DARES

3. Résultats

- Sur les données de test Pôle Emploi au niveau **<u>familles de métiers</u>**

	f1-score	precision	recall	support
micro avg	0.96	0.96	0.96	303585
macro avg	0.94	0.94	0.93	303585
weighted avg	0.96	0.96	0.96	303585



DARES

3. Résultats

- Sur les données partenaires au niveau **<u>familles de métiers</u>**

	f1-score	precision	recall	support
micro avg	0.89	0.89	0.89	187986
macro avg	0.81	0.83	0.80	187986
weighted avg	0.89	0.89	0.89	187986

Les données de Pôle emploi ne sont pas représentatives

DARES

3. Résultats

- Entraînement d'un SVM sur les données partenaires
- Résultats sur les données au niveau **<u>familles de métiers</u>**

Partenaires

	f1-score	precision	recall	support
micro avg	0.93	0.93	0.93	187986
macro avg	0.88	0.91	0.86	187986
weighted avg	0.93	0.93	0.93	187986

Pôle emploi

	f1-score	precision	recall	support
micro avg	0.92	0.92	0.92	303585
macro avg	0.85	0.86	0.84	303585
weighted avg	0.92	0.92	0.92	303585

DARES

3. Résultats

- Entraînement d'un SVM sur les données partenaires et Pôle emploi
- Résultats sur les données au niveau **<u>familles de métiers</u>**

Partenaires

	f1-score	precision	recall	support
micro avg	0.92	0.92	0.92	187986
macro avg	0.87	0.91	0.85	187986
weighted avg	0.92	0.92	0.92	187986

Pôle emploi

	f1-score	precision	recall	support
micro avg	0.96	0.96	0.96	303585
macro avg	0.93	0.94	0.92	303585
weighted avg	0.96	0.96	0.96	303585

DARES

3. Résultats

- Entraînement d'un SVM sur les données partenaires et Pôle emploi
- Résultats sur les données au niveau **appellations**

Partenaires

	f1-score	precision	recall	support
micro avg	0.85	0.85	0.85	187986
macro avg	0.72	0.78	0.72	187986
weighted avg	0.84	0.85	0.85	187986

Pôle emploi

	f1-score	precision	recall	support
micro avg	0.91	0.91	0.91	303585
macro avg	0.81	0.84	0.80	303585
weighted avg	0.91	0.92	0.91	303585

DARES

Merci de votre attention !

Avez-vous de questions ?

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

MINISTÈRE DU TRAVAIL

DARES
direction de l'animation de la recherche, des études et des statistiques

STATISTIQUE PUBLIQUE

dares.travail-emploi.gouv.fr

A l’avenir

Mise à jour de la nomenclature des Familles Professionnelles : match des libellés avec les appellations des PCS / avec les intitulés de poste des codes Rome.

DARES