Sicore un outil au service des nomenclatures Georges Bourdallé





Georges BOURDALLÉ
INSEE

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Les différents points qui seront abordés

- Présentation de l'application Sicore
- Le processus de développement d'un environnement Sicore
- L'évolution et la maintenance d'un tel environnement
- Les moyens humains
- Les ressources statistiques
- Les enjeux d'une codification automatique

La codification par Sicore

La codification des principales variables d'enquêtes est faite:

- **soit automatiquement par le logiciel Insee Sicore**
- **soit manuellement par les sites de reprise**

Remarque: L'activité économique peut aussi être codée par la méthode MCA (nécessité de collecter les données employeurs et d'avoir des répertoires d'entreprise)

S*ystème* **I***nformatisé de* **CO***dage* des **R***éponses* aux **E***nquêtes*

Logiciel d'expertise développé par l'INSEE

Mis au point entre 1993 et 1996

Permet la codification automatique en s'appuyant sur des nomenclatures

Conçu pour:

- Mettre au point les bases de connaissances nécessaires au codage automatique d'une variable exprimée sous forme textuelle, le libellé pouvant être accompagné de variables annexes précisant son sens
- Chiffrer automatiquement des libellés
- Analyser les résultats de codage automatique: non-codés, mal-codés, codés multiples,……

La codification par Sicore

Suivi par différents acteurs :

- Un expert Sicore
- Un Responsable Informatique d'Application (R.I.A)
- Des experts variables (experts de leur domaine, bien souvent des experts nomenclatures)

La codification Sicore

Rôle des différents acteurs

- Expert Sicore: Assure le développement, la maintenance, le conseil, l'expertise, la veille technologique, la formation des utilisateurs, la diffusion de l'application, la gestion des sites de reprise (charge et planning), les plannings de codification automatique
- Le R.I.A: formalise les modifications, instruit les demandes d'évolutions et valide techniquement les nouveaux environnements
- L'expert variable: c'est l'expert de la variable à coder. Il doit y avoir un expert par variable. Par exemple pour l'activité, l'expert variable naturel est l'expert nomenclature
  - Il est aussi en charge de la construction, du suivi et de la qualité des bases de connaissances pour Sicore

La codification par Sicore

Les utilisateurs à l'INSEE

- Equipes d'enquêtes et de projets (statisticiens, informaticiens)
  - Pour les enquêtes ménages, le recensement, les DADS, l'état civil, Sirene, les répertoires d'entreprises….

Les utilisateurs hors INSEE

- Instituts de sondage privés, organismes publics de statistiques, INS étrangers,…..

Sicore, un projet INSEE pour la codification automatique

Un logiciel de codification universel

Peut tout coder a priori, il lui suffit d'apprendre…

Des bases de connaissances spécifiques

Pour chaque variable à coder, un ensemble d'informations pour apprendre à coder

Exemples de variables codées avec Sicore

Professions

Activités d ’entreprise

Communes, pays et nationalités

Diplômes, niveaux et spécialités de formation

Dépenses des ménages (achats et lieux d’achat des enquêtes Budget De Famille)

Occupations (enquêtes Emploi du temps)

……

Taux de réussite des codages

Entre 70 et 80% dans les enquêtes (Dépôt-retrait) INSEE

90% avec Sicore « embarqué »

Communes, pays et nationalités

environ 98%

Vitesse de codage

Professions

Environ 1000 libellés à la seconde

Communes, pays et nationalités

Env. 7000 pour les communes

Aperçu du processus

Dans l'enquête : mobiliser l'information

- **Recueillir le libellé et des informations qui complètent ce libellé**

Avec Sicore : coder avec cette information

- **Simplifier les libellés de l'enquête**
- **Comparer les libellés d'enquête à des libellés de référence déjà codés (contenus dans les fichiers d'apprentissage)**
- **Si besoin, utiliser des variables annexes**

La codification par Sicore

2 types de développement Sicore

- **Environnement sans règles**
- **Environnement avec règles**

INSEE

Environnement sans règles

Le libellé est suffisant pour coder. La reconnaissance et le codage se font en une même étape

La codification par Sicore

Environnement avec règles

- **Le libellé n'est pas suffisant pour coder. La reconnaissance et le codage se font en deux étapes**
  - **Etape1: le libellé est reconnu ou pas**
  - **Etape 2: le codage s'effectue en fonction de règles s'appuyant sur des informations**

INSEE

Aperçu du processus

Simplifier les libellés (normalisation)

Enlever ce qui ne donne pas d'information

(les caractères et mots « vides » : articles, ponctuation…)

Résumer l'information

(remplacer des expressions ayant la même signification par une seule expression synonyme)

Calibrer l'information

(ne garder d'un certain nombre de mots ayant une longueur maximum fixée)

Aperçu du processus

Comparer des libellés d'enquête à des libellés de référence déjà codés

Si le libellé ressemble à un libellé de référence

↳on lui attribue le pré-code du libellé de référence

Si besoin, utiliser des variables annexes

La seule comparaison des libellés ne suffit pas toujours pour aboutir à un code précis de la nomenclature.

↳ **Utilisation d'information complémentaire :**

Les variables annexes mises en oeuvre dans des tables contenant des règles de décision (règles logiques)

Si le libellé ne ressemble pas à un libellé de référence

↳on est en échec de codification automatique

↳**Reprise manuelle**

La codification par Sicore

Les échecs de codification:

- Libellés mal orthographiés
- Libellés figurant dans une phrase (ex: « je suis boulanger à Paris »)
- Libellés avec sigle (ex: agent SNCF)
- Libellés non reconnu dans le fichier d'apprentissage

Les équipes de reprise

Reprise manuelle des libellés en échec de codification automatique

L'INSEE dispose de deux pôles d'expertise et de reprise

- un pôle d'expertise et de reprise spécialisé dans la reprise des professions et activités
- un pôle d'expertise et de reprise spécialisé dans la reprise des diplômes et niveau de formation

Un exemple du processus : enquête Emploi du temps 1998

- *Libellé déclaré* => « je prépare une pizza pour les enfants »
- *1ère étape : simplification* => « prépare pizza enfants »
- *2ème étape : synonymisation* => « prépare **aliments** enfants »
- *3ème étape : comparaison du libellé d'enquête avec les libellés de référence de Sicore* => différence (échec) / ressemblance (=> Affectation d'un pré-code: 4ème étape)
- *4ème étape : utilisation des variables annexes (ici, le but) en vue de l'affectation d'un code*

si but personnel	=> code 311
si but professionnel	=> code 211
si but associatif	=> code 542
si but non renseigné	=> code 311 (arbitrairement choisi car le plus fréquent)

Les différentes formes de Sicore

Sicore batch

Sicore poste Expert

Sicore embarqué

Sicore interactif

Les différentes formes de Sicore

Sicore batch

Permet la codification en batch (en différé) et l'intégration dans les chaînes de traitement des enquêtes

Les différentes formes de Sicore

Sicore poste Expert

- ✓ Utilisation sous Windows, pour l'expert Sicore, les experts variables et certains utilisateurs
- ✓ Permet non seulement de coder **mais surtout de construire ou modifier les bases de connaissances**

Les différentes formes de Sicore

Sicore embarqué

- ✓ **Utilisation de Sicore sous Capi en cours d'entretien**
  - ➢ **Permet d'améliorer la collecte de l'information**
    - ➢ **Divise par 2 les échecs de codification**
    - ➢ **Signale les libellés non reconnus ou trop vagues**
      - ➢ L'enquêteur doit corriger les fautes de frappe ou d'orthographe

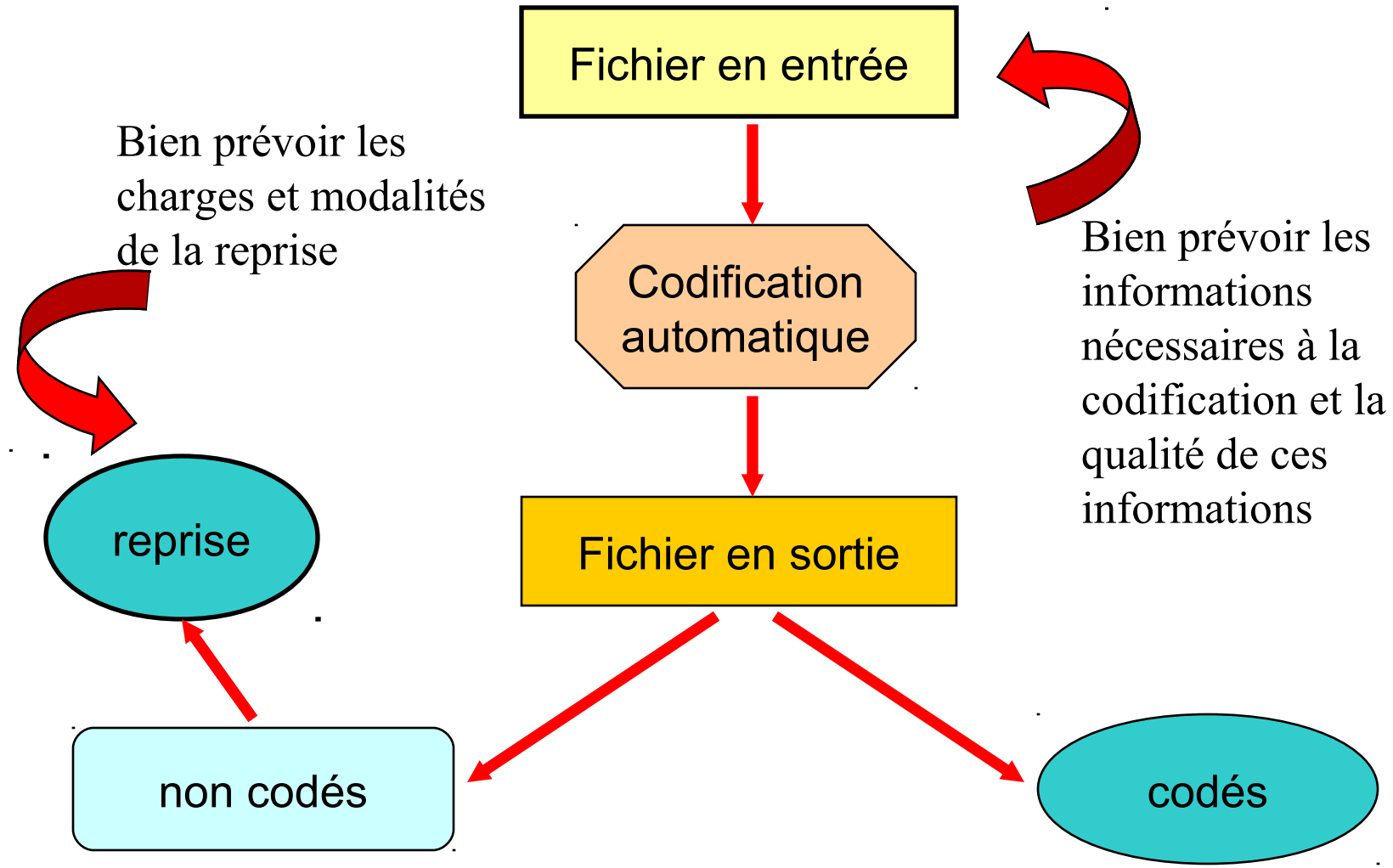
Les différentes formes de Sicore

Sicore interactif

- Utiliser pour la reprise des non codés

  - **Interrogation d'un serveur Sicore directement à partir des postes de reprise**

La codification par Sicore



Le suivi de la qualité

En amont

La qualité des bases de connaissances, tests

La qualité des libellés à coder (formation des enquêteurs)

Les retours aux Directions régionales fait par les Pôles d'expertise et de reprise

En aval

Analyse sur un échantillon de codés

Analyse sur les non codés et repris manuellement

Les mises à jour

L'expert variable

enrichit et corrige si nécessaire les bases de connaissances

L'équipe Sicore (Expert Sicore + R.I.A)

rend disponible les nouvelles bases de connaissances pour les utilisateurs

Démonstration de codage

Codification de l'activité (avec Sicore APE)

Activité par activité

À partir d'un fichier LIBELLE_ACTIVITE

L’information en sortie de Sicore

Suivant la variable à coder, 2 ou 3 grandes informations

- Un indicateur de codage
- Le code
- Les informations nécessaires absentes lors du codage
- Exemple de fichier en sortie de Sicore FICHIER CODE

Questions, remarques…

2. Développement d'un environnement Sicore





Georges BOURDALLÉ
INSEE

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Processus de développement

- Est-il nécessaire d'automatiser la codification?
  - L'opération de codification est-elle pérenne?
- Quelle est la variable à coder? Activité?
  - Quelle nomenclature utiliser?
  - Quelle est l'information nécessaire pour coder? (informations complémentaires)?
  - Quel type d'environnement Sicore?
    - Environnement avec ou sans règles?
- Quelles sont les sources disponibles?
  - Quelle méthodologie pour les fichiers d'apprentissage?
    - Comment établir les égalités LIBELLE=CODE
  - A partir de quelles bases de référence établir la base d'apprentissage?
- Quelles sont les expertises passées?
  - Expertise sur la qualité du codage manuel?
- Quels sont les moyens humains?
  - Expert variable? Informaticien? Codeurs?

INSEE

Environnement sans règles

Définir la base d'apprentissage

- **Fichier Apprentissage Brut (FAB)**
  - **Contient la connaissance pour coder c'est à dire un ensemble de couples LIBELLE = CODE**
  - **Exemple de FAB**

Environnement avec règles

Définir la base d'apprentissage

- **Fichier Apprentissage Brut (FAB)**
  - **Contient la connaissance pour coder c'est à dire un ensemble de couples LIBELLE= PRECODE**
- **Fichier de règles**
  - **Autant de règles que de PRECODE**
  - **Exemple de FAB avec PRECODE**
  - **Exemple de fichier de REGLES**

Méthodologie de définition des fichiers d'apprentissage:

A partir de quelles sources?

Que retenir?

Exemple: l'environnement Sicore Ape pour le codage de l'activité

3. Enjeux et méthodologie de mise en place d'une codification automatique





Georges BOURDALLÉ
INSEE

Codification automatique versus codification manuelle

- Codification automatique
  - Volumétrie
  - Moyens humains
  - Homogénéité du traitement
  - Rapidité du traitement
  - Partie intégrante de la chaîne de traitement
- Codification manuelle
  - Coûteux en moyen humain et en temps
  - Hétérogénéité du codage
  - Limité en volumétrie

Codification automatique

Temps de développement + ou - important

Mise à jour de l'environnement

Expertise

Méthodologie de mise en place d'un environnement Sicore

- **Identification de la source de référence**
  - **Le référentiel peut être construit à partir de la codification passée**
- **Analyse de la qualité de cette source**
- **Choix du type d'environnement (avec ou sans règles)**
- **Formation Sicore en vue la création de l'environnement Sicore**
- **Sources à coder, plusieurs exercices**
- **Mise en place et test qualité de l'environnement**
  - **Comparer codage sicore et codage manuel et robustesse sur plusieurs exercices**
    - **Par exemple: Si *n* exercices, compiler les *n-2* exercices pour fichier d'apprentissage et coder exercices *n-1* et *n***
- **Discussion autour d'un calendrier de mise en production**

Proposition de scénario de développement

Définir le référentiel et les règles de codage

Source existante: Activité codée

Analyse de la source afin de formaliser la codification

Établir le référentiel: un fichier du type : à un libellé suivi d'informations annexes (environnement avec règles) ou pas (environnement sans règles) correspond un code

Établir l'environnement Sicore (fichier d'apprentissage, fichier de règles,…..)

Tester l'environnement Sicore

Taux et qualité de codification et robustesse (définir des indicateurs de qualité)

Mise en production

INSEE