Data Documentation Initiative (DDI), un standard de documentation des données

Alina Danciu, Alexandre Mairot

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

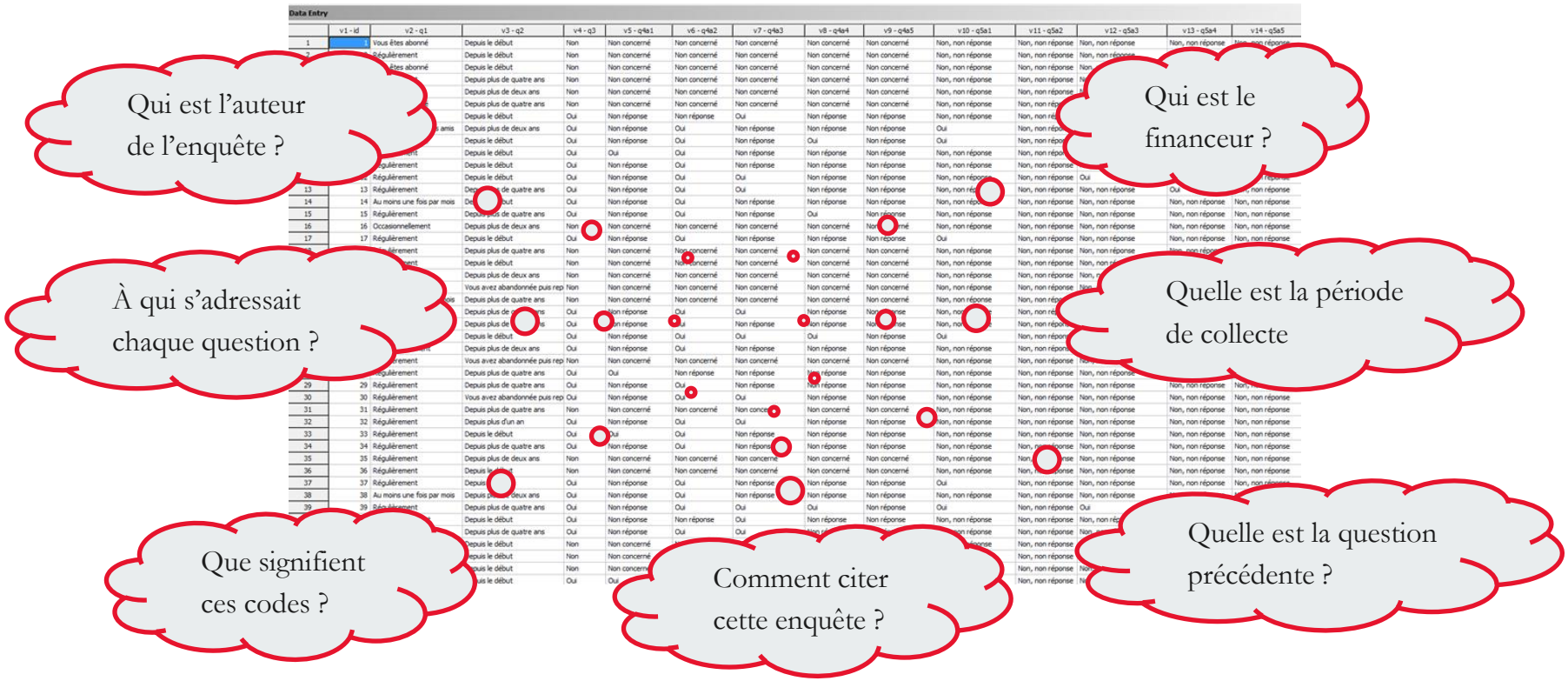
Data Doccumentation Initiative (DDI), un standard de documentation des données

Webinar Tuto@MATE
14 mars 2019

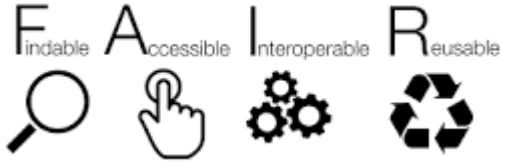
Alina Danciu, Alexandre Mairot

Quelles sont les informations qui sont indispensable pour l'utilisation d'un fichier de données ?

Quelles sont les informations supplémentaires que vous aimeriez avoir, au cas où elles étaient disponibles ?


Data Entry
Qui est l'auteur de l'enquête ?
Qui est le financeur ?
À qui s'adressait chaque question ?
Quelle est la période de collecte
Que signifient ces codes ?
Comment citer cette enquête ?
Quelle est la question précédente ?

Dans quel but ?



- Documentation ;
- Exploration ;
- Interopérabilité ;
- Réutilisation.

DDI n’est pas un logiciel, mais un standard
Sélectionnez les éléments DDI qui vous correspondent

Définition

La Data Documentation Initiative (DDI) est une norme internationale permettant de décrire les données issues d'enquêtes et d'autres méthodes d'observation en sciences sociales, comportementales, économiques et de la santé. DDI peut documenter et gérer différentes étapes du cycle de vie des données de recherche, telles que la conception, la collecte, le traitement, la diffusion, la découverte et l'archivage.

Source : DDI Alliance

Un fichier documenté de données

Qui utilise DDI ? Quelques exemples

Produits DDI (1)

DDI 2 

- Version initiale publiée en 2000 (version courante 2.5)
- Le but initial est la **documentation** des données pour permettre leur **exploration** et leur **préservation**

DDI 3 

- Version initiale publiée en 2008 (version courante 3.2, version 3.3 à venir)
- Le but principal est d'ajouter du contenu pour permettre :
  - Traitement des fichiers de **données complexes** (ex : données longitudinales…)
  - **Réutilisation** des métadonnées
  - Rendre **compte du cycle de vie** complet des données et des métadonnées dès la conception, au traitement et à la documentation des données, jusqu'à leur diffusion, leur préservation et leur réutilisation

Produits DDI (2)

DDI 4

- **Prototype** en cours d'élaboration. **Implémentation** pas encore possible ;
- Écrit en UML (Langage de Modélisation Unifié) ;
- Structure standard pour les métadonnées qui englobera une grande partie de ce qui est disponible dans les structures XML DDI aujourd'hui. Elle inclura également les fonctionnalités des vocabulaires DDI RDF

Vocabulaires contrôlés

Les spécifications DDI 2

- Se concentre sur la création d'un livre des codes statistiques (*codebook*) ;
- Conçu pour des utilisations limités ;
- Permet la découverte des données par les utilisateurs en identifiant la variable ou l'étude recherchée ;
- La documentation selon ce modèle se concentre sur une seule enquête, un seul fichier de données ;
- La documentation des variables contient la majorité des informations (question, modalités de réponses, nature des données…)
- La documentation est conçue comme un complément à la collecte des données :
    - La variable doit exister avant que les métadonnées (documentation) puissent être créées ;
    - La documentation entraîne un coût supplémentaire car elle ne soutient pas le processus de développement ou de collecte des données.

DDI 2 : deux niveaux de documentation

Niveau “étude”

Niveau “variable”

DDI 2 : spécifications et métadonnées

Source : *Workshop DDI – Train the Trainers* (2018), Schloss Dagstuhl, 24-28 septembre

Les spécifications DDI 3

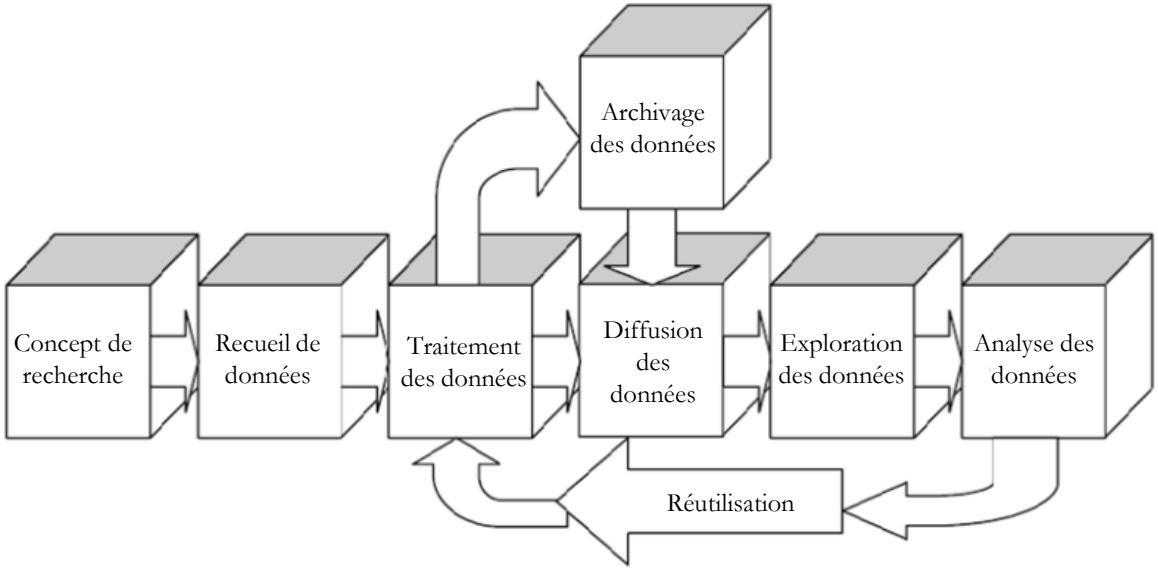
Conçu :

- Pour répondre à un large éventail d'exigences typiques de la gestion et de l'utilisation des métadonnées ;
- Pour prendre en charge tous les types de réutilisation et pour fonctionner avec des approches par registre et par référentiel.

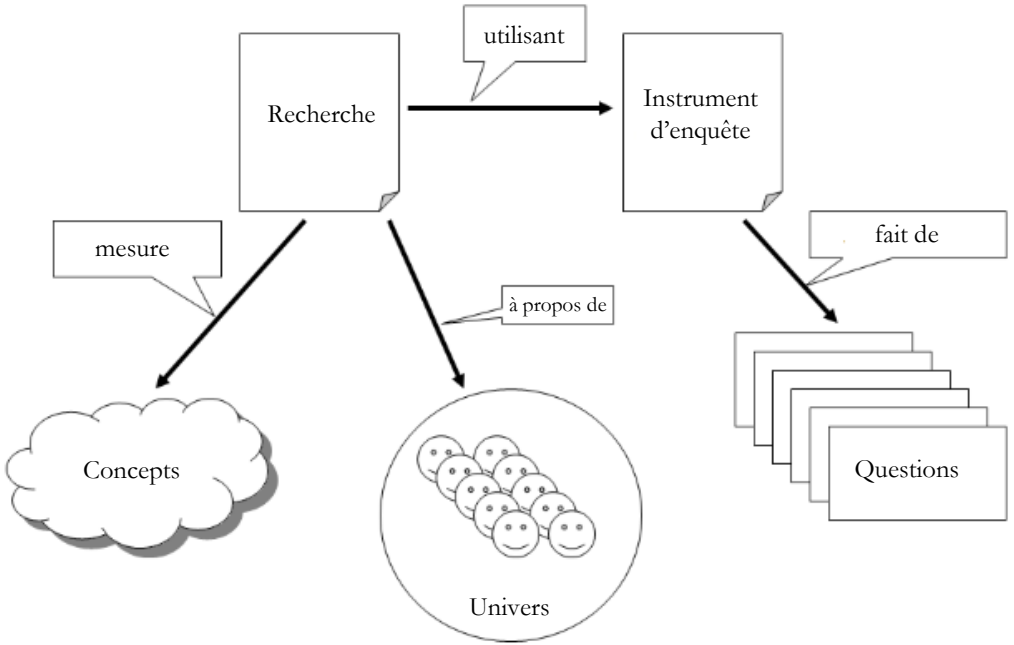
Implique :

- Une centralisation des systèmes de métadonnées ;
- Une réutilisation des métadonnées à des fins d'uniformité et de qualité ;
- L'exploration des ressources à l'aide de systèmes et de processus axés sur les métadonnées ;
- Beaucoup de métadonnées incluses par référence ;
- L'indentification et la résolution unique pour récupérer les métadonnées à partir de sources distribuées ;
- Des politiques claires sur la gestion des versions et des métadonnées.

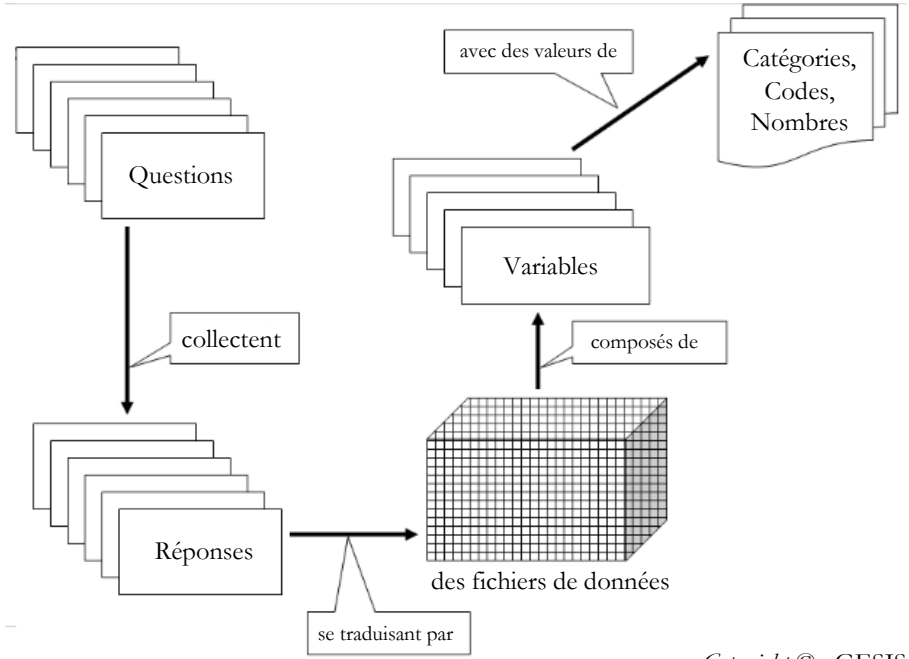
DDI 3 : le cycle de vie des données



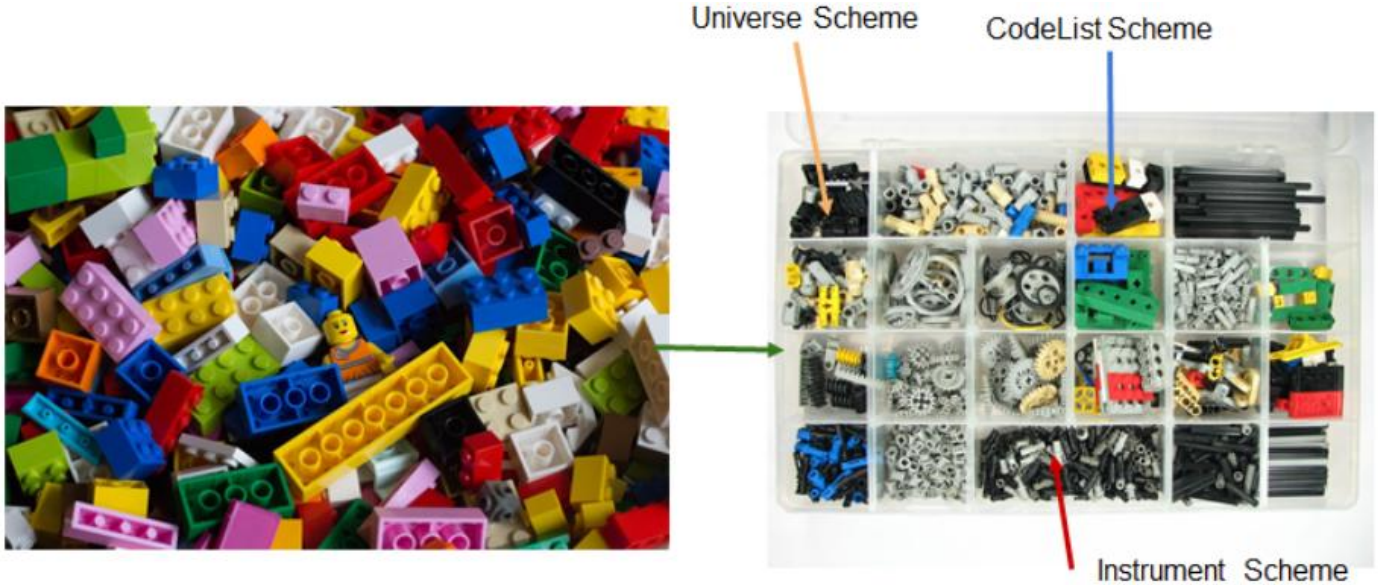
DDI 3 en 60 secondes



DDI 3 en 60 secondes

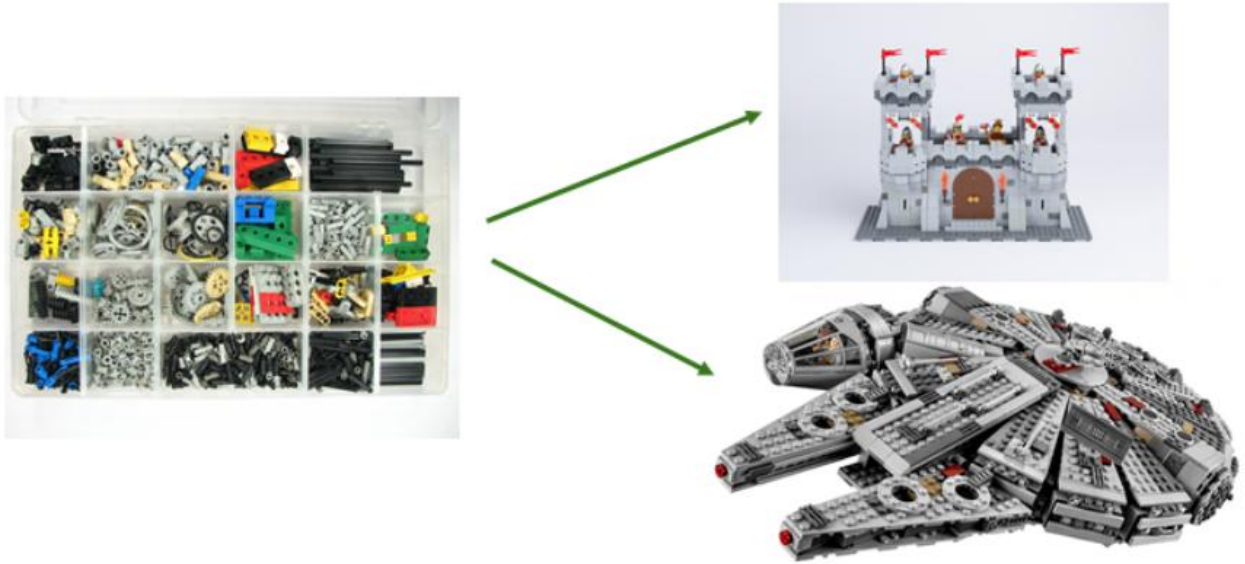


DDI 3 : spécifications et métadonnées



Source : *Workshop DDI – Train the Trainers* (2018), Schloss Dagstuhl, 24-28 septembre

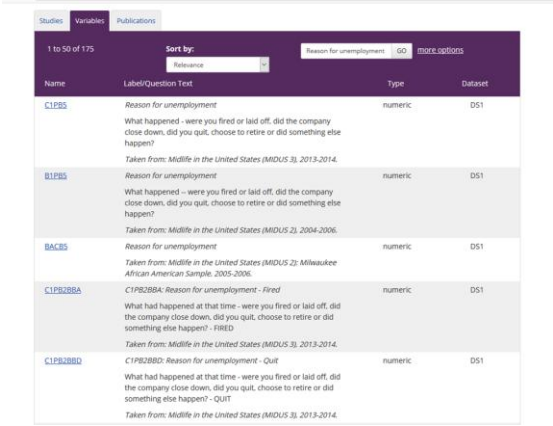
DDI 3 : la réutilisation des métadonnées

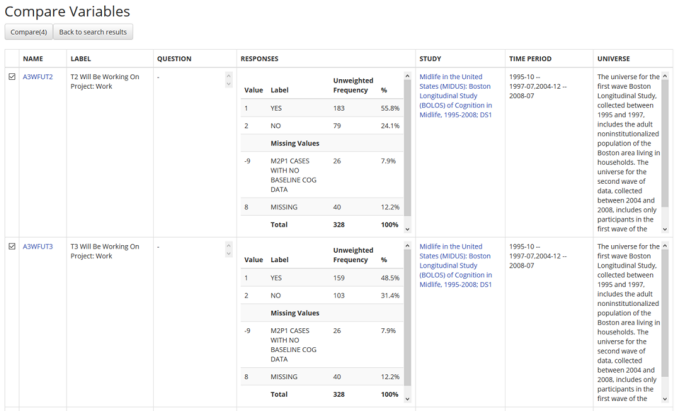


Source : *Workshop DDI – Train the Trainers* (2018), Schloss Dagstuhl, 24-28 septembre

DDI 3 en action

MIDUS : une étude longitudinale sur la santé et le bien-être, réalisée aux États-Unis depuis les années 1960, 3 vagues, plus de 500 publications par an à partir des données





Les points communs entre DDI 2 et DDI 3

- L'utilisation de vocabulaires contrôlés ; 
- La structuration des métadonnées ;
- L'échange d'informations ;
- La documentation de variables dans un ensemble de données ;
- La création d'un dictionnaire de codes (*codebook*) pour les utilisateurs finaux ;
- Documenter l'univers des variables (à qui les questions ont été posées) ;
- Créer des groupes de variables ;
- Documenter des jeux de données produits dans différents formats ;
- Distinguer les valeurs manquantes.

Les spécificités de DDI 2 et DDI 3

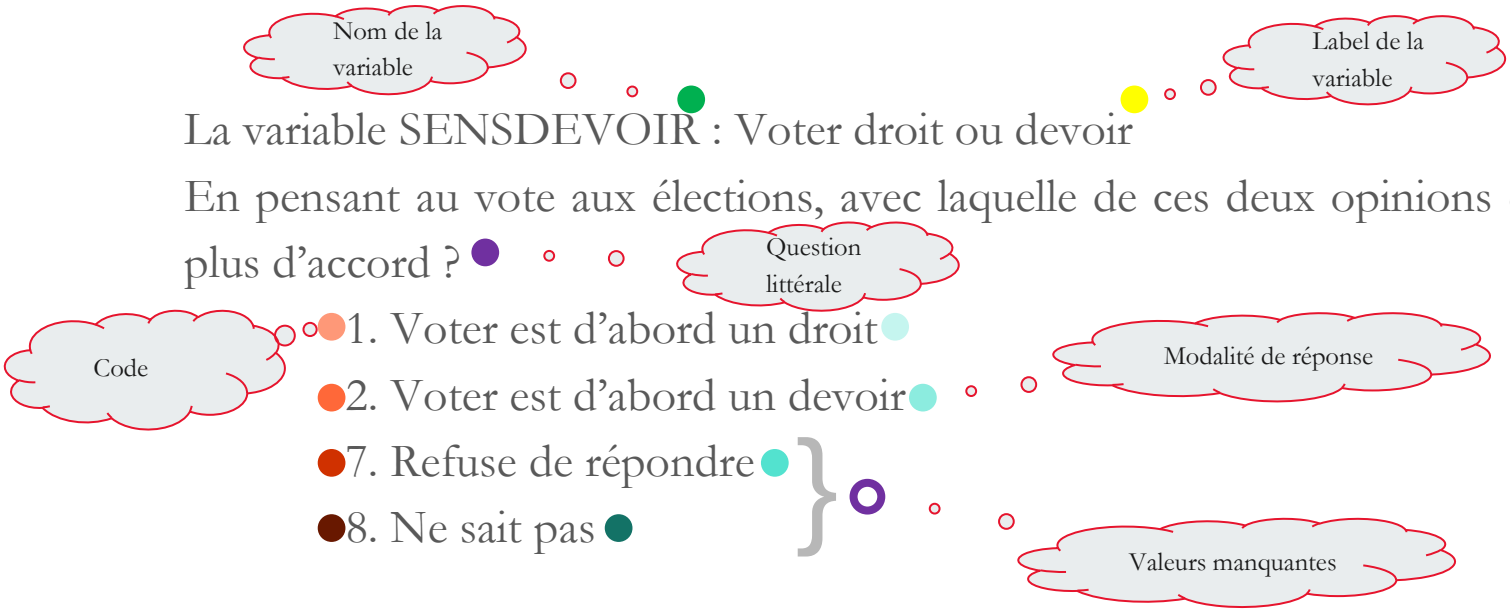
Les spécifications DDI 2 sont plus adaptées pour :

- La documentation des enquêtes simples et uniques ;
- Où il s'agit de fournir uniquement des informations concernant l'étude.

Les spécifications DDI 3 est plus adaptées pour :

- La réutilisation, l'harmonisation des questions, des variables ;
- La gestion cohérente des métadonnées, du versionning ;
- La documentation du déroulement du questionnaire ;
- L'échange d'informations avec d'autres normes ;
- La gestion et la documentation des données selon les producteurs et à travers le temps ;
- La documentation des données en différentes langues.

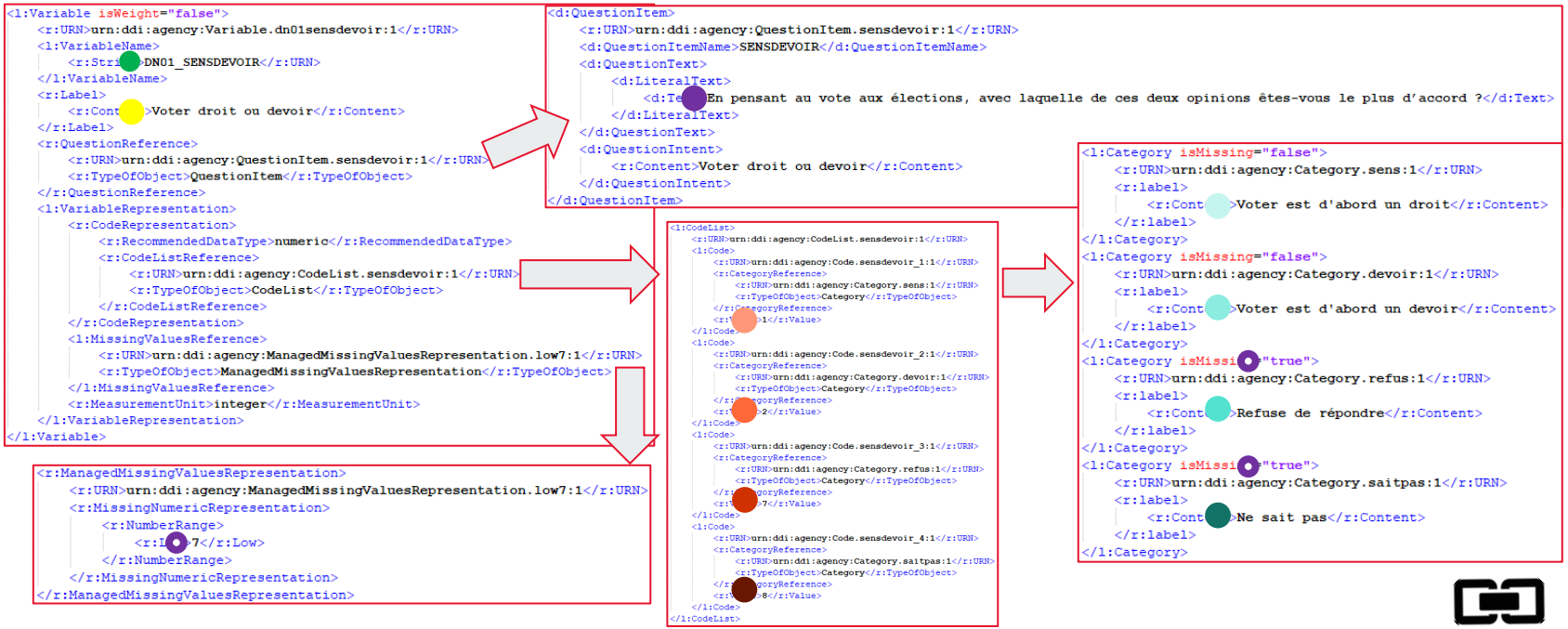
La documentation des variables et DDI



Documentation d’après DDI 2

```
<var ID="V15" name="DN01_SENSDEVOIR" wgt="not-wgt">
    <labl>Voter droit ou devoir</labl>
    <qstn>
        <qstnLit>En pensant au vote aux élections, avec laquelle de ces deux opinions êtes-vous le plus d'accord ?</qstnLit>
    <qstn>
    <valrng>
        <range UNITS="INT" maxExclusive="7">
    </valrng>
    <invalrng>
        <range UNITS="INT" min="7">
    </invalrng>
    <catgry>
        <catValu>1</catValu>
        <labl>Voter est d'abord un droit</labl>
    </catgry>
    <catgry>
        <catValu>2</catValu>
        <labl>Voter est d'abord un devoir</labl>
    </catgry>
    <catgry missing="">
        <catValu>7</catValu>
        <labl>Refuse de répondre</labl>
    </catgry>
    <catgry missing="">
        <catValu>8</catValu>
        <labl>Ne sait pas</labl>
    </catgry>
</var>
```

Documentation d'après DDI 3



Les systèmes de métadonnées

Le scénario (**A**) :

- Création de métadonnées standard, importation de métadonnées à partir de fichiers de données, prise en charge des systèmes d'exploration de données (portails, catalogues).

Le scénario (**B**) :

- Centralisation de la gestion des métadonnées : *Single Source of Truth* (entrepôt de données) ;
- Usage optimal des outils et des processus de documentation existants dans le 1er niveau pour la création, l'édition des métadonnées ;
- Gestion des versions possibles.

Le scénario (**C**) :

- La collecte de métadonnées tout au long du cycle de vie des données dès la conception de l'enquête ;
- La gestion et la documentation des données selon les producteurs et à travers le temps.

Sledgehammer

- Convertir les données au format texte ASCII ;
- Générer des configurations pour de nombreux packages d'analyse et de bases de données ;
- Extraire les métadonnées standard (DDI) des fichiers de données ;
- Calculer des statistiques descriptives au niveau des variables et des catégories à inclure dans le fichier au format DDI ou à d'autres fins.

Colectica pour Excel

- Importer des fichiers de données à partir de SPSS, STATA ou SAS pour les documenter dans Excel ;
- Exporter les métadonnées au format DDI.

Nesstar

- Créer et éditer des métadonnées au format DDI ;
- Extraire les métadonnées des logiciels statistiques ;
- Valider les métadonnées et les variables ;
- Publier les données sur différentes plateformes de diffusion (portails, bases de questions…).

Colectica Designer

- Éditer des métadonnées au format DDI ;
- Concevoir des questionnaires d'enquête ;
- Intéragir avec Colectica Repository pour une gestion avancée des données ;
- Générer divers formats de données, de documentations et de codes sources.

Une communauté internationale très active

- Workshop DDI annuels organisés par GESIS ;
- Conférence annuelle européenne des utilisateurs de DDI (EDDI) ;
- Groupes de travail au sein de la DDI Alliance.