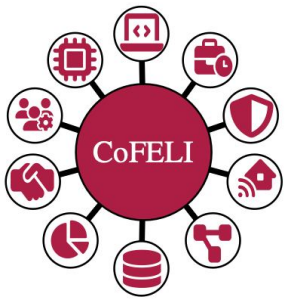


**Collectif francophone pour l'enseignement libre de l'informatique**

# Données manquantes

*Fondements*

**TMR_07**

Christina KHNAISSER *(christina.khnaisser@usherbrooke.ca)*
Luc LAVOIE *(luc.lavoie@usherbrooke.ca)*
*(les auteurs sont cités en ordre alphabétique nominal)*

—

*Scriptorum/Scriptorum/TMR_07-Donnees_manquantes, version 0.1.2.a, en date du 2025-09-02*

*— document de travail, ne pas citer —*

## *Sommaire*

Le module traite du problème des données manquantes et présente quelques-unes des solutions courantes en les caractérisant tant du point de vue théorique que pratique. Il met aussi l'accent sur la primaute de la modélisation qui doit guider la conception… et non l'inverse comme il arrive parfois.

Le présent module a été rédigé dans le cadre de l'exploration du thème « Modélisation, conception et exploitation de données » (MCED) par des membres du CoFELI. Il vise à présenter les connaissances usuellement couvertes au sein d'une formation universitaire en informatique (B. Sc. nord-américain, licence ou mastère européen), voire celles proposées par certaines écoles d'ingénieur. Ce module est destiné aux personnes étudiant les disciplines de l'informatique, de l'informatique appliquée et du génie logiciel. Nous espérons toutefois qu'il pourra être utile à toute personne curieuse d'apprivoiser ce champ de connaissance.

## *Mise en garde*

Ce document contient les noes et les réflexions colligées dans le but de préparer à la rédaction d'un essai sur la question des données manquantes.

L'intégration des plusieurs autres sources n'est pas en core réalisées, par exemple

- CoFELI::Scriptorum/doc/MCED/TMR/TMR_07-Logiques.xlsx
- Akademia::Domaines/Modules/Bases de donnees/MCED/SQL_08-Annulabilite/SQL_08-Operateurs-pour-les-nuls_PRE.pptx
- Akademia::Domaines/Modules/Bases de donnees/MCED/TMR_07-Donnees-absentes/TMR_07-Donnees-manquantes_PRE.pptx

Le présent document est donc en cours d'élaboration ; en conséquence, il est incomplet et peut contenir des erreurs.

## *Historique*

| diffusion | resp. | description |
|---|---|---|
| 2025-01-23 | LL | Synthèse et réécriture de différents modules proposés par les groupes *Akademia*, Μητις et GRIIS. |

# Table des matières

## Introduction

Le présent document a pour but de présenter une synthèse du problème posé par la modélisation des données manquantes et de diverses solutions proposées par Edgar F. Codd et développée grâce, notamment, aux contributions de Christopher J. Date, Hugh Darwen, Nikos A. Lorentzos et Jeffrey D. Ullman.

# 1. Étude de cas

## 1.1. Entité de référence

*Étudiant*

- matricule
- ddn
- nom
- prénom
- adresse postale
- adresse de résidence
- téléphone
- citoyenneté
- référent
- sexe
- genre
- id_UdeS (CIP)
- id_QC
- id_CA
- passeport
- visa

## 1.2. Analyse

*matricule*

- représentation interne et externe
- droit d'accès
- modifiabilité
- deux matricules plutôt qu'un ?
  - m1 interne, automatisé, non modifiable, jamais externalisé
  - m2 externe, automatisé, modifiable, externalisable (accès restreint)

*ddn*

- granularité
- calendrier
- fuseaux horaires
- modifiabilité

*nom, prénom*

- us et coutumes
  - Le nom complet se divise-t-il ? Si oui, en combien de parties ? Le nombre de parties détermine-t-il le rôle de chacune ?
  - L'ordre dans lequel sont énumérées les parties varie-t-il ?
  - Le nom varie-t-il dans en fonction des étapes de vie ? Dans ce cas, faut-il conserver les anciennes valeurs ?
- représentations, alphabets, etc.
  - translittération

- modifiabilité
- multiplicité
    - Jean-Baptise Poquelin, dit Molière
    - celui du passeport, de l'état civil, du permis de conduire, du diplôme…
- séparation du nom à l'état civil et des ses «formes» selon divers usages
    - un nom complet dans un alphabet prescrit
    - diverses formes utiles dans le contexte d'application

*adresse postale, adresse de résidence*

- standard de référence
- modifiabilité
- multiplicité

*téléphone*

- standard de référence
- modifiabilité
- multiplicité

*citoyenneté*

- standard de référence
- modifiabilité
- multiplicité

*référent*

- modifiabilité
- multiplicité
- → référence à une autre entité

*sexe*

- standard de référence
- modifiabilité

*genre*

- standard de référence
- modifiabilité
- multiplicité

*id_UdeS (CIP)*

- redondance avec matricule ?
    - différenciation entre identification d'information et identification de service
- standard de référence
- modifiabilité

*id_CA*

- standard de référence
- modifiabilité

*passeport, visa*

- standard de référence
- modifiabilité
- multiplicité

- → référence à une autre entité

# 2. Synthèse de la démarche

## 2.1. Préambule

Attentes → Besoins → Exigences → (F, NF)

Politique → Directive → Règlement → Procédure

## 2.2. Vers une méthode

- Inventaire des attributs
- Association du domaine de valeurs (le type), la référence (standard), les représentations
- Exigences applicables
    - Identification
    - Applicabilité
        - Est-elle applicable en tout temps ? Attention aux dépendances fonctionnelles ! Y compris les dépendances fonctionnelles multi-relations !
        - Si non applicable dans tous les cas, il faut pouvoir distinguer ces cas !
    - Participation et cardinalité (0, 1, n)
        - Si le zéro fait partie de l'intervalle de participation, l'information peut être (légitimement absente)
        - Cette participation varie-t-elle dans le temps ?
        - attribut non atomique (non scalaire) ou référence ?
    - Variation dans le temps et selon les procédés
        - À partir de quand l'information est-elle requise, nécessaire ? En pratique, cela s'exprime le plus souvent entre d'une étape d'un procédé que d'une échéance temporelle proprement dite. Ce qui est plus simple à modéliser… pourquoi ?
- Dépendances fonctionnelles
- Analyse des clés

## 3. Pistes de solutions

### 3.1. Solution par la logique 4V et le typage (4VT ;  Scott, Strachey et Stoy)

- Inapplicabilité prise en compte
- Causalité non prise en compte
- Logique 4V
- Typage avec suprémum et infimum

### 3.2. Solution par la logique 4V et le marquage (4VM ; Codd II)

- Inapplicabilité prise en compte
- Causalité non prise en compte
- Logique 4V
- Typage avec suprémum et infimum
- Marquage des attributs
- Convention d’usage unique

### 3.3. Solution par la modélisation simple (MODS ; McGovern)

- Inapplicabilité prise en compte
- Causalité non prise en compte
- Logique 2V
- Modélisation par décomposition PJ

### 3.4. Solution par la modélisation causale (MODC ; Darwen)

- Inapplicabilité prise en compte
- Causalité prise en compte
- Logique 2V
- Modélisation par décomposition RU

*Variante déclarative*

RU uniquement

*Variante contrôlée*

PJ+RU combinés

### 3.5. Solution par la logique 3V, le marquage et une convention uniforme (3VU, Codd I)

- Inapplicabilité non prise en compte
- Causalité non prise en compte
- Logique 3V de Priest
- Marquage des attributs
- Convention d’usage unique

### 3.6. Solution par la logique 3V, le marquage et des conventions différentiées (3VD, SQL)

- Inapplicabilité non prise en compte
- Logique 3V de Priest
- Marquage des attributs
- Conventions d’usage différentiées
- Causalité non prise en compte

# 4. Exemples

## 4.1. Ce qu'il faut considérer

- obligatoire
- non obligatoire
  - inapplicable, donc absent
  - applicable, mais absent, donc la cause
  - applicable et présent, donc la valeur

## 4.2. Cas d'étude

*Patient*

- matricule
- nom
- ddn
- ddd
- sexe
- genre
- enceinte
- nba

## 5. De la logique

À rédiger

## Conclusion

À rédiger

Merci aux membres de CoLOED, CoFELI et Μητις :-))

## Définitions

```
Sources consultées de juin 2023 à juillet 2024

* Antidote : Antidote 11 v4.2 (2023), voir https://www.antidote.info
* Le Larousse : https://www.larousse.fr/dictionnaires/francais
* Le Robert : https://dictionnaire.lerobert.com
* Wikipédia : https://fr.wikipedia.org/wiki
```

**algèbre**

Branche des mathématiques qui étudie les structures abstraites en employant les lois de composition.

**informatique**

1. Science
    1. Science du traitement de l'information.
    2. Science du traitement automatique et rationnel de l'information.
    3. Science fondamentale du traitement automatique de l'information. (CoFELI).
2. Technique
    - Ensemble des techniques de la collecte, du tri, de la mise en mémoire, du stockage, de la transmission, et de l'utilisation des informations traitées automatiquement à l'aide de logiciels mis en oeuvre sur des ordinateurs.
3. Spécialisation
    - Informatique théorique : concerne la définition de concepts et modèles.
    - Informatique pratique : s'intéresse aux techniques concrètes de mise en oeuvre.
    - Informatique appliquée : s'intéresse à l'utilisation de l'informatique pour la résolution de problèmes formulés en regard d'autres domaines (que l'informatique).

> *Hypothèse*
>
> L'information est le fondement (la base commune et indispensable) aux connaissances et aux communications.
>
> *Corolaires*
>
> L'information ne peut être traitée automatiquement que si elle est représentée par des données, elles-mêmes réalisées sous la forme d'un phénomène (représentation) tangible (physique).
>
> Par ailleurs, seul le traitement rationnel de l'information peut être considéré en soutien à une science. L'informatique étant une science fondamentale, elle ne peut donc s'intéresser qu'au traitement rationnel de l'information. D'autres sciences, non fondamentales, pourraient s'intéresser au traitement non rationnel.
>
> Par ailleurs, l'information est notamment relative à l'espace-temps et à l'agent, ce qui, entre autres, distingue l'informatique de la mathématique.
>
> D'autre part, le traitement est décrit par des procédés et rendu effectif par des processus dont les actions atomiques sont définies par un modèle d'automates (plusieurs tels modèles ont été définis, proposés et utilisés, automate à états finis, machine de Turing, etc.). Si, en pratique, il est utile de pouvoir associer une durée à chaque action atomique, la mesure du temps n'est pas pour autant requise, bien que fréquente.
>
> En conséquence,
>
> - Le coeur de métier s'intéresse à la modélisation (de l'information et du traitement) ;

dans le cas du traitement, il est suffisant d'étudier (de modéliser) les seuls procédés de traitement de l'information ; par contre, par définition, la modélisation de l'information doit s'appliquer à tout type d'information, donc à toute information relative à l'univers physique.

- (Toutes) les grandeurs physiques doivent donc être prises en compte même si, pour l'information non connotée physiquement, il pourrait être jugé suffisant de se limiter à l'espace et au temps — ce qui est le cas en informatique théorique.
- Une grandeur propre à l'informatique doit être ajoutée aux grandeurs physiques afin de mesurer l'information : le bit. Une autre unité, le qubit, est proposée pour l'information quantique.

*Débats*

- Le bit mesure-t-il l'information, la donnée, une représentation spécifique de la donnée ?
- Les unités de mesure du temps et de l'espace utilisées en informatique doivent-elles être les mêmes qu'en physique ?
- Plusieurs propriétés du temps sont encore très débattues, en physique comme en informatique :
  - Le temps est-il discret ou continu ?
  - Est-il borné (dans le passé, dans le futur) ?
  - Est-il cyclique ?
  - Le moment de référence du 1er janvier 1970, souvent retenu en informatique et en physique, doit-il être à midi ou à minuit ?
  - Ce moment de référence ne devrait-il pas être le 1er janvier 1958 ? Pourquoi ?
  - Le temps doit-il être coordonné ou non ?
  - Faut-il adopter une notation qui s'affranchit complètement des calendriers, comme la mesure de la longueur s'est affranchie de la morphologie humaine (pouce, paume, pied, coudée, pas, double pas) et celle de la masse, d'un objet étalon (le mètre étalon) ?

**information**

Élément de connaissance susceptible d'être transmis au moyen d'une suite de signes.
Élément de connaissance représentable par une donnée.

**logique**

- Antidote : Science qui a pour objet l'étude des méthodes de raisonnement, de pensée, par lesquelles on peut atteindre la vérité.
- Larousse : Science du raisonnement en lui-même, abstraction faite de la matière à laquelle il s'applique et de tout processus psychologique.
- Le Robert : Étude scientifique, surtout formelle, des normes de la vérité.
- Wikipedia : Étude des règles formelles que doit respecter toute argumentation correcte.
- CoFELI : Science fondamentale du raisonnement en lui-même (faisant donc abstraction des connaissances auxquelles il s'applique et de l'agent qui l'exerce).

**représentation**

Au sens général, « Action de rendre sensible quelque chose au moyen d'une figure, d'un symbole, d'un signe. Cette figure, ce symbole, ce signe. » [Antidote, 2025-08-25]

Dans un contexte informatique, on distingue une représenttion interne (une suite de signaux) d'une représentation externe (une suite de signes).

**type**

Définition variable selon les auteurs.

*LL*

Selon la théorie des types (Russell) :

- defA : dénotation d'un ensemble de valeurs propres (selon le modèle, l'ensemble peut être infini, ou pas).
- defB : dénotation d'un sous-ensemble de defA déterminé par une contrainte (selon le modèle, le sous-ensemble peut être impropre, ou pas).

Suivant Russell, la plupart des logiciens, des mathématiciens et des informaticiens algorithmiques ont adopté les dénominations suivantes :

- Type : defA
- Sous-type : defB

Toutefois, la plupart des ontologistes et des informaticiens en modélisation de données (dont Codd et Date première manière) utilisent les dénominations suivantes :

- Domaine : defA
- Type : defB

En SQL, on augmente encore la confusion :

- Type : defA
- Domaine : defB

À noter que Date s'est rallié à la dénomination de Russell depuis « Type Inheritance and the Relational Theory » (2016).

Dans les documents du CoFELI, nous utiliserons la dénomination de Russell et consorts.

# Sigles

**ACID**

Acronyme désignant conjointement les propriétés d'atomicité, de cohérence, d'isolation et rémanence (pérennité ou *durability* en anglais) relativement au traitement transactionnel.

**EA**

Acronyme désignant les modèles conceptuels de données fondés sur la théorie entité-association.

**SGBD (Système de gestion de bases de données)**

Service informatique permettant de stocker, manipuler, gérer et partager des données, à l'aide d'un langage fondé sur un modèle permettant d'abstraire la complexité des opérations internes requises tout en garantissant la qualité, la pérennité et la confidentialité des données.

**SGBDR (Système de gestion de bases de données relationnelles)**

SGBD utilisant un modèle d'abstraction fondée sur la théorie relationnelle de Codd et intégrant comme critère de qualité le maintien des propriétés ACID.

**SQL (*Structure Query Language*)**

Langage de programmation axiomatique fondé sur un modèle inspiré de la théorie relationnelle proposée par E. F. Codd.
[Normes applicables : ISO 9075:2016, ISO 9075:2023]

# Références

**[Codd1970a]**
Edgard F. CODD ;
*A Relational Model of Data for Large Shared Data Banks* ;
Communications of the ACM, 13(6), pp. 377–387, 1970 ;
doi:10.1145/362384.362685.

**[Codd1990a]**
Edgard F. CODD ;
*The Relational Model for Database Management: Version 2* ;
Addison-Wesley Longman Publishing, Boston (MA, USA), 1990 ;
ISBN 0-201-14192-2.

**[Date1998a]**
Chris J. DATE, Hugh DARWEN ;
**First Edition** : *Foundation for Object/Relational Database Systems: The Third Manifesto* ;
Addison-Wesley Redwood City (CA, US), 1998 ; ISBN 0-201-30978-5.
**Second Edition** : *Foundation for Future Database Systems: The Third Manifesto* ;
Addison-Wesley, Redwood City (CA, US), 2000 ; ISBN 0-201-70928-7.
**Third edition** : *Databases, types, and the relational model: The third manifesto* ;
Addison-Wesley (Pearson Education), 2007 ; ISBN 0-321-39942-0.
**Third *revised* edition** : *Databases, types, and the relational model: The third manifesto* ;
2014, https://www.dcs.warwick.ac.uk/~hugh/TTM/DTATRM.pdf (consulté le 2024-05-30).

**[Date2015a]**
Chris J. DATE ;
*SQL and relational theory: how to write accurate SQL code* ;
O'Reilly Media, 2015 ;
ISBN 978-1-4919-4117-1.

**[Date2020a]**
Chris J. DATE ;
*Logic and Relational Theory* ;
Technics Publications, Basking Ridge (NJ, US), 2020 ;
ISBN 978-1634628754.

**[Russell1908a]**
Bertrand RUSSELL ;
*Mathematical Logic as Based on the Theory of Types* ;
American Journal of Mathematics, vol. 30, no 3, 1908, p. 222–262 ;
ISSN 0002-9327, DOI 10.2307/2369948.

Produit le 2025-09-04 08:10:29 -0400

**Collectif francophone pour l'enseignement libre de l'informatique**