BASES DE DONNÉES

Données manquantes, absentes, inconnues, nulles, etc.

BD010
v222b

2018-08-28

Christina KHNAISSER et Luc LAVOIE
Département d'informatique
Faculté des sciences

UNIVERSITÉ DE SHERBROOKE

Christina.Khnaisser@USherbrooke.ca
http://info.USherbrooke.ca/ckhnaisser
Luc.Lavoie@USherbrooke.ca
http://info.USherbrooke.ca/llavoie

Plan

- Préambule
- De quelques logiques non classiques
- *Approches structurelles*
- *Approches par modélisation*
- *Quelle approche choisir ?*
- De la théorie relationnelle aux modèles relationnels
- Vocabulaire
- Références

Note :
les sections en italiques peuvent être omises en première lecture.

Données absentes

- Pourquoi une donnée serait-elle absente?
  - Les réponses de SPARC!
- Un modèle simple
  - proposé par Codd
- Solutions AVEC annulabilité
- Solutions SANS annulabilité

Préambule
Données absentes… selon SPARC (1/4)

1. L'information est applicable, mais la valeur n'est pas encore connue (*date de décès d'une personne vivante*).
2. L'information est inapplicable (*nombre de sommets d'un cercle*).
3. L'information existe, mais il n'est pas permis (*légalement*) de l'enregistrer *(religion d'un employé*).
4. L'information existe, mais on n'a pas les moyens de trouver la valeur (*évaluation d'un employé alors qu'il travaillait pour une organisation concurrente*).
5. L'information existe, mais elle n'est pas encore enregistrée (*en raison de l'absence de l'employé préposé à la saisie*).

Préambule
Données absentes… selon SPARC (2/4)

6. L'information est enregistrée, mais pas encore disponible (*texte écrit, saisi, stocké, mais pas encore publié*).
7. L'information a été enregistrée puis supprimée (*un utilisateur ne veut plus que le nom de son conjoint soit conservé*)
8. L'information est disponible, mais en changement et donc potentiellement invalide (*solde d'un compte bancaire sur lequel une opération est en cours*).
9. L'information est disponible, mais on ne sait pas si elle est fiable (*la note d'examen non encore approuvée par le doyen*).
10. L'information est disponible, mais invalide (*si une erreur s'est produite lors du calcul de la valeur*)

Préambule
Données absentes… selon SPARC (3/4)

11. La classe d'information est sécurisée (*les informations personnelles des professeurs ne sont pas accessibles aux étudiants*).
12. L'objet représentant l'information est sécurisé (*un utilisateur bloque l'accès à ses infos personnelles sur un réseau social*).
13. Une information est sécurisée durant un certain laps de temps (*le budget préalablement à sa communication au parlement*).
14. L'information est calculée à partir d'au moins une information absente ou incertaine (*l'âge en fonction d'une date de naissance par ailleurs absente*).

DONNÉES ABSENTES… SELON SPARC (4/4)

- Incomplet.
  - En fait, les raisons de l'absence varie selon le contexte et leur interprétation selon la finalité de la requête.
- Trop complexe.
  - Tant pour la saisie que pour les requêtes, la complexité induite est trop élevée en regard de de l'effort requis et des moyens de vérification/validation.

Préambule
Données absentes hiérarchisées (1/2)

- [02] L'information est inapplicable.
- L'information est applicable, mais
  - [01] elle n'est pas encore connue.
  - [03] il n'est pas permis de l'enregistrer.
  - [04] il n'y a pas moyen d'en trouver la valeur.
  - [14] elle est calculée à partir d'au moins une information absente.
  - [05] elle n'est pas encore enregistrée.
  - bien qu'elle ait été enregistrée,
    - [07] elle a été supprimée.
    - [06] elle n'est pas encore disponible.
    - bien qu'elle soit disponible,
      - [08] elle est en cours de modification et donc potentiellement invalide.
      - [09] elle n'est pas fiable.
      - [10] elle est invalide.
      - bien qu'elle soit valide,
        - [11] sa classe d'information est sécurisée.
        - [12] elle est sécurisée.
        - [13] elle est temporairement sécurisée.

Préambule
Données absentes hiérarchisées (2/2)

- Incomplet.
- Trop complexe.

DONNÉES ABSENTES… UN MODÈLE SIMPLE (PROPOSÉ PAR CODD)

- **N**
  - L'information n'est **pas applicable**.
  - Dans ce cas, l'utilisation de l'annulabilité est à remettre en question; une bonne modélisation permet généralement d'éviter d'y avoir recours.
- **I**
  - L'information est **inconnue**.
  - Dans ce cas, l'annulabilité pourrait être légitime; la question est de savoir comment la représenter pour que cela pose le moins de problèmes possible.
- **X**
  - L'information n'est **pas accessible**.
    - **À court terme** : le gestionnaire transactionnel permet d'éviter l'utilisation de l'annulabilité en différant la mise à disponibilité tout en conservant le contrôle des accès concurrents et en préservant lé cohérence de la BD.
    - **À long terme** : équivalent à **I**.

PRÉAMBULE
SOLUTIONS AVEC ANNULABILITÉ

- Que faire lorsqu'une donnée est absente?
- Trois solutions classiques
  - corriger cette lacune à la source (dans la réalité);
  - modifier le modèle pour en tenir compte;
  - introduire la notion d'*annulabilité* dans la théorie relationnelle, ce qui induit le recours
    - à une logique non classique
      - (afin de pouvoir définir l'égalité, essentielle aux opérations d'affectation, de restriction, de jointure, d'union...)
    - et à l'un des deux artifices suivants :
      - un marqueur NUL (une propriété des attributs) ou
      - une valeur NULLE (ajoutée à tous les domaines).

PRÉAMBULE
SOLUTIONS SANS ANNULABILITÉ

- Principes
  - Séparer les propositions complètes des incomplètes.
  - Conserver les causes d'absence séparément.

- Pour un inventaire des techniques de modélisation
  - http://www.dcs.warwick.ac.uk/~hugh/TTM/Missing-info-without-nulls.pdf

De quelques logiques non classiques

- Belnap
- Kleene
- SQL
- ...

Logique non classique
4V - Belnap

- B : sur-déterminé; N : sous-détrminé

$f_{\neg}$	
T	F
B	B
N	N
F	T

$f_{\wedge}$	T	B	N	F
T	T	B	N	F
B	B	B	F	F
N	N	F	N	F
F	F	F	F	F

$f_{\vee}$	T	B	N	F
T	T	T	T	T
B	T	B	T	B
N	T	T	N	N
F	T	B	N	F

- C'était la proposition révisée de Codd... et elle n'a pas été suivie par le comité de standardisation du langage SQL.

Logique non classique
3V – Kleene

P3 (de Priest : – I est sur-détrminée – T et I sont vraies – avec tautologie)

$\neg$	
T	F
I	I
F	T

$\wedge$	T	I	F
T	T	I	F
I	I	I	F
F	F	F	F

$\vee$	T	I	F
T	T	T	T
I	T	I	I
F	T	I	F

$\rightarrow_K$	T	I	F
T	T	I	F
I	T	I	I
F	T	T	T

$\leftrightarrow_K$	T	I	F
T	T	I	F
I	I	I	I
F	F	I	T

B3 (faible : I est réductrice – T seule valeur vraie – avec tautologie)

$\neg$	
T	F
I	I
F	T

$\wedge_+$	T	I	F
T	T	I	F
I	I	I	I
F	F	I	F

$\vee_+$	T	I	F
T	T	I	T
I	I	I	I
F	T	I	F

$\rightarrow_+$	T	I	F
T	T	I	F
I	I	I	I
F	T	I	T

$\leftrightarrow_K$	T	I	F
T	T	I	F
I	I	I	I
F	F	I	T

K3 (forte : I est sous-déterminée - T seule valeur vraie – pas de tautologie)

non retenue

CHECK
satisfait ssi
T ou U

WHERE
satisfait ssi
T

...

Chercher la logique!

OR	true	unknown	false
true	true	true	true
unknown	true	unknown	unknown
false	true	unknown	false

AND	true	unknown	false
true	true	unknown	false
unknown	unknown	unknown	false
false	false	false	false

P	*NOT P*
true	fase
unknown	unknown
false	true

IS	*TRUE*	*UNKNOWN*	*FALSE*
true	true	false	false
unknown	false	true	false
false	false	false	true

APPROCHES STRUCTURELLES

- Extension de domaines
- Marqueur d'attributs
- Ce qui est réglé
- Ce qui ne l'est pas

Approches structurelles
Extension de domaines

- Au tableau!

Approches structurelles
Marqueur d'attributs

- Au tableau!

Approches structurelles
Ce qui est réglé

- Au tableau!

Approches structurelles
Ce qui ne l'est pas

- Au tableau!

APPROCHES PAR MODÉLISATION

- Décompositions
  - PJ (McGovern)
  - RU (Darwen)
- Utilisations
  - Applicabilité : PJ (McGovern)
  - Causalité : RU (Darwen)
- Solutions mixtes

Approches par modélisation
Décompositions

- Au tableau!

Approches par modélisation
Utilisations

- Au tableau!

Approches par modélisation
Solutions mixtes

- Au tableau!

QUELLE APPROCHE CHOISIR ?

- Les dangers du NUL
- La lourdeur des décompositions
- L'attribut non applicable
- La valeur inconnue

De: oracle-acct_ww@oracle.com
Objet: Nom d'utilisateur de votre compte Oracle
Date: 2 octobre 2014 19:44
À: luc.lavoie@usherbrooke.ca

ORACLE

Cher/Chère NULL !,

Vous avez demandé à recevoir par email le nom d'utilisateur de votre compte Oracle.

Votre nom d'utilisateur est : **luc.lavoie@usherbrooke.ca**

Merci !

L'équipe de gestion des comptes Oracle

Mettez votre compte à jour :

- Abonnez-vous aux communications dédiées aux thèmes qui vous intéressent.
- Devenez membre des communautés Oracle.
- Pour modifier votre adresse email, votre mot de passe ou toute autre information de votre compte, cliquez sur le lien Compte en haut des pages Oracle.com.

Obtenir de l'aide

- Des questions ? Aide (page Account Help)
- Se connecter
  - Envoyer une demande d'aide
  - profilehelp_ww@oracle.com

Hardware and Software
ORACLE
Engineered to Work Together

Aide (page Account Help) | Ne pas envoyer d'email | Mentions légales | Conditions d'utilisation | Confidentialité

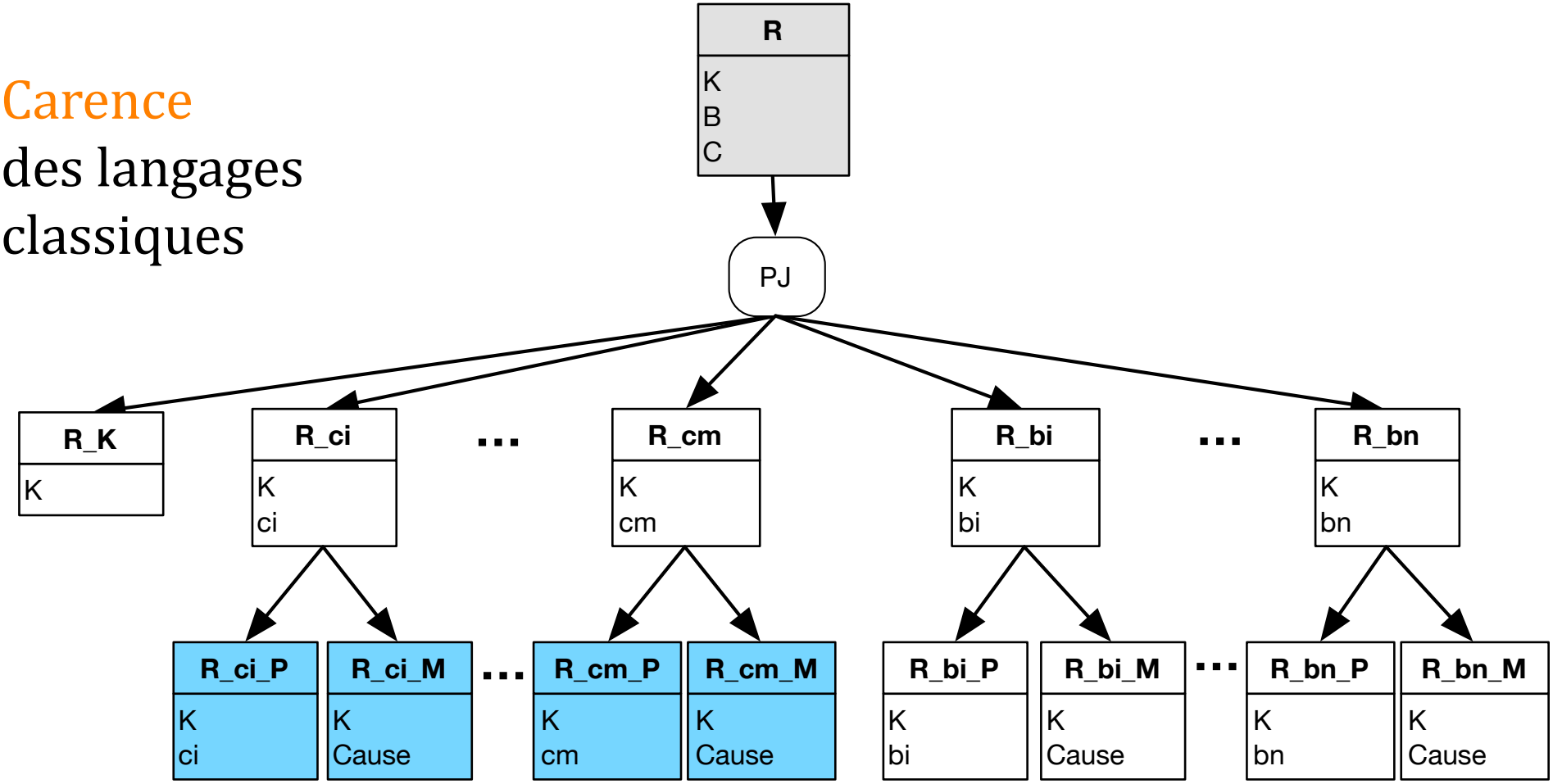
Explosion
des cas

Impraticabilité
des logiques
non classiques

Complexité
de la
vérification/
validation

Explosion
des relations

Carence
des langages classiques



QUE CHOISIR -1- ?

- Donnée absente ::=
  - donnée (attribut) non applicable
  - donnée (valeur) inconnue

- La différence entre « non applicable » et « inconnue » est irréconciliable.

- Règle A :
  - La non-applicabilité doit se refléter dans le modèle (le schéma).

QUE CHOISIR -2- ?

- L'interprétation d'une « valeur inconnue » est dépendante de plusieurs facteurs dont
  - la cause de l'absence et
  - le prédicat associé au résultat.

- Règle B :
  - Aucune valeur ne peut être substituée ou inférée par le schéma lui-même (pas de valeur « par défaut »).
- Règle C :
  - Il peut être souhaitable que le schéma conserve la cause de l'absence.
  - Il est nécessaire que la requête détermine explicitement l'interprétation devant être donnée à l'absence.

De la théorie aux modèles

- Pourquoi?
- Modèle de Codd I
- Modèle de Codd II
- Modèle de Date
- Modèle d'Ullman
- Modèles SQL
- Au final...

De la théorie aux modèles
Pourquoi n'y a-t-il pas un seul modèle?

- Parce qu'il n'y a (avait?) pas consensus sur la bonne façon de traiter les données absentes.
- Parce que certains sont prêts à sacrifier l'intégrité de leurs données et des résultats de leurs requêtes au profit des gains de performance (généralement éphémères et illusoires).
- Pour permettre d'intégrer de nouveaux résultats théoriques facilitant la modélisation et l'exploitation de données.

Il faut cependant être très prudent avant d'introduire un nouveau modèle, car tout mauvais modèle dès lors qu'il est utilisé acquiert une latence ÉNORME.

De la théorie aux modèles
Modèle de Codd I

- Transposition directe de la théorie avec les exceptions suivantes
  - marqueur «nul»
  - logique trivaluée
  - pas de relations dans les relations

De la théorie aux modèles
Modèle de Codd II

- Transposition directe de la théorie avec les exceptions suivantes
  - marqueurs «non applicable» et «nul»
  - logique quadrivaluée
  - pas de relations dans les relations

De la théorie aux modèles
Modèle de Date

- Transposition directe de la théorie, conséquemment
  - pas de marqueur nul ni de valeur nulle
  - logique bivaluée
  - intégration des relations dans le système de typage (donc ajout des opérateurs *tclose*, *wrap* et *unwrap*)

De la théorie aux modèles
Modèle d'Ullman

- Transposition de la théorie relationnelle à l'aide de collections
  - marqueur «nul»
  - logique trivaluée
  - possibilité de doublons dans les relations

Modèle SQL ISO

- Transposition de la théorie relationnelle à l'aide de collections et de listes
  - marqueur «nul»
  - logique trivaluée
  - possibilité de doublons dans les relations
  - les attributs d'un tuple sont ordonnés et peuvent être anonymes, voire synonymes (!!!)
  - possibilité d'attributs à valeurs multiples
  - …

ET LES AUTRES MODÈLES?

- TSQL2, BCDM, DDLM, AV, noSQL, coRel...
- Certains de ceux-ci seront couverts par les activités IFT 287, IGE 487 et IFT 723

Au final

- Nous maintenons la position adoptée en BD010 :
  - Pour l'exposé des principes relationnels, nous utiliserons toujours le modèle de Date.
  - Pour la programmation SQL, nous présenterons des techniques permettant d'être aussi proche que possible du modèle de Date, en indiquant les écarts possibles en fonction du modèle SQL ISO 2011.

Les colles du prof (2/2)

- Reclasser les 14 cas recensés par SPARC selon les catégories N, I et X.
- Faire le lien entre les catégories N, I, X et les trois solutions permettant de traiter les valeurs absentes (corriger la source, modifier le modèle, introduire le concept d'annulabilité).
- Conclure en statuant sur la nécessité (ou non) du concept d'annulabilité.
- Quels sont les modèles relationnels utilisés en cours?

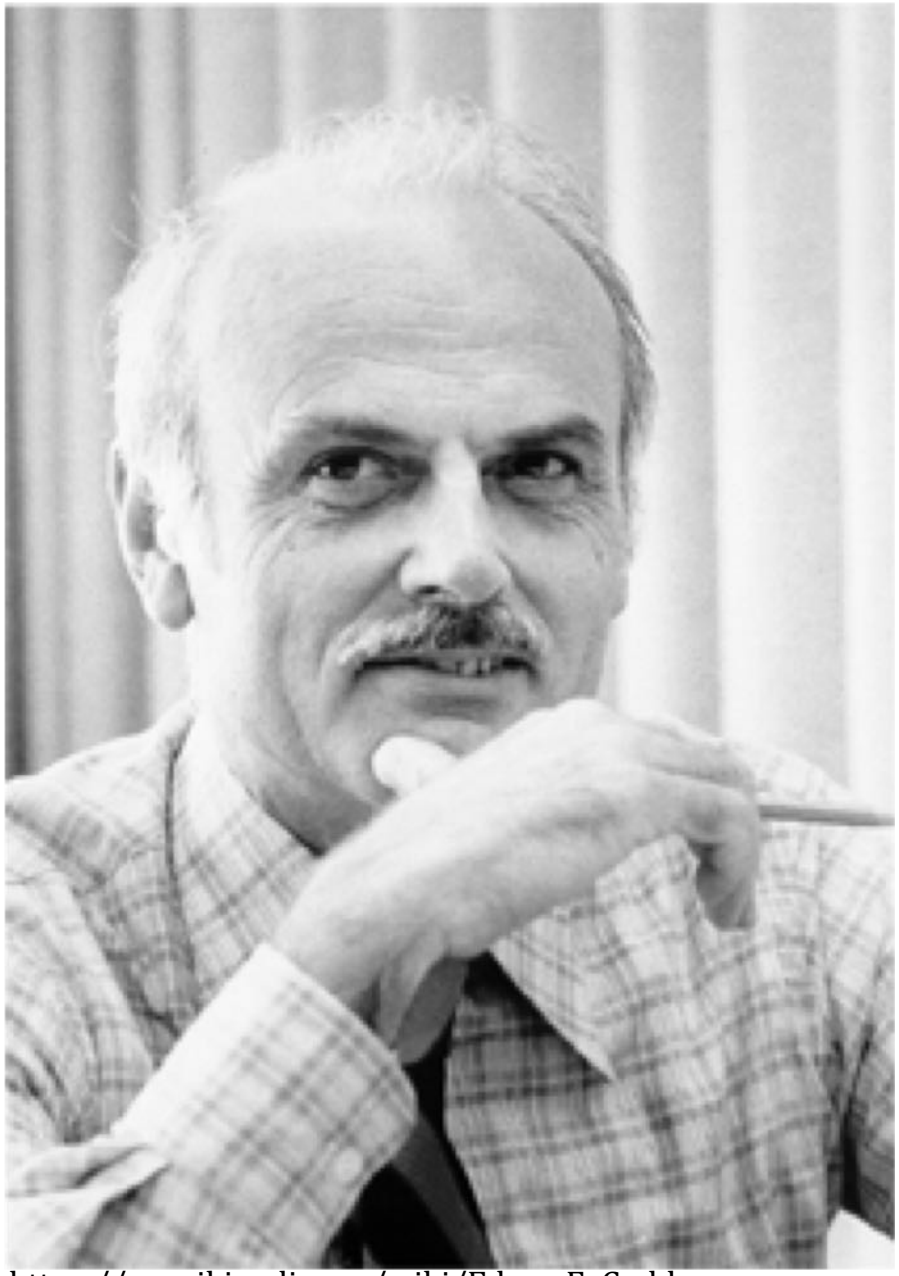
RÉFÉRENCES

- Théorie relationnelle
  - E.F. Codd. 1990.
    The Relational Model for Database Management: Version 2.
    Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
  - C.J. Date, H. Darwen. 2007.
    Databases, types and the relational model: the third manifesto.
    Reading, Mass.: Addison-Wesley.
  - F. de Sainte Marie. 2013.
    Bases de données relationnelles et normalisation : de la première à la sixième forme normale.
    ftp://ftp-developpez.com/fsmrel/basesrelationnelles/normalisation/normalisation.pdf
  - H. Darwen. 2006.
    How To Handle Missing Information Without Using NULL.
    http://www.dcs.warwick.ac.uk/~hugh/TTM/Missing-info-without-nulls.pdf
- Manuels classiques
  - [C. J. Date 2004], chapitre 3.
  - [Elmasri and Navathe 2004], chapitre 4.
  - [Elmasri and Navathe 2011], chapitre 3.
  - [Elmasri and Navathe 2016], chapitre 8.
  - [Ullman and Widom 2008], chapitre 3.

AUTRES SOURCES

- Une synthèse des conséquences du NULL en SQL
  - https://en.wikipedia.org/wiki/Null_(SQL)
- Codd et Date débattent du sujet
  - http://web.archive.org/web/20100531071357/http://www.dbdebunk.com/page/page/1706814.htm

Edgar Frank Codd et Christopher J. Date



https://en.wikipedia.org/wiki/Edgar_F._Codd

Photo of Chris Date by Douglas Robertson, Edinburgh

https://en.wikipedia.org/wiki/Christopher_J._Date