Langage certifié en Coq pour la provenance des données issues d'analyses bioinformatiques

Rébecca Zucchini
Encadrée par V. Benzaken, S. Cohen-Boulakia, C. Keller

Laboratoire de Méthodes Formelles (LMF), Université Paris-Saclay

Introduction

Beaucoup de données ...

Nombreuses données générées

en bioinformatique, objets connectés, satellites sondeurs,..

Analyse de données

Explosion du nombre de données générées

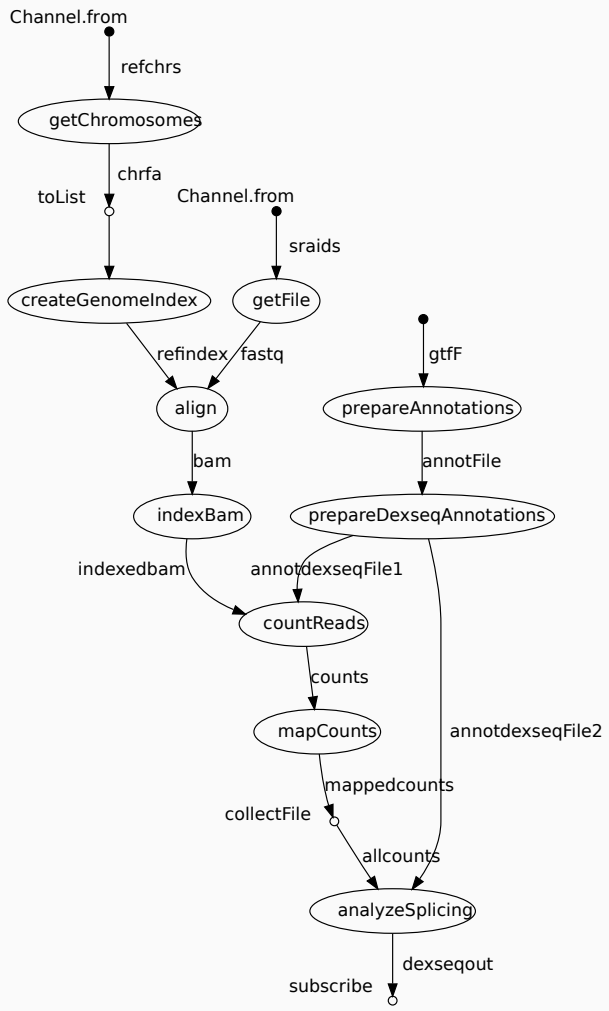


Figure 1: Exemple de workflow

.. à traiter

Nouveaux défis en traitement des données

⇒ quelles sont les données initiales qui produisent une donnée particulière ?

⇒ à quel point peut-on avoir confiance en une donnée particulière ?

⇒ peut-on reproduire un résultat à partir de données similaires ?

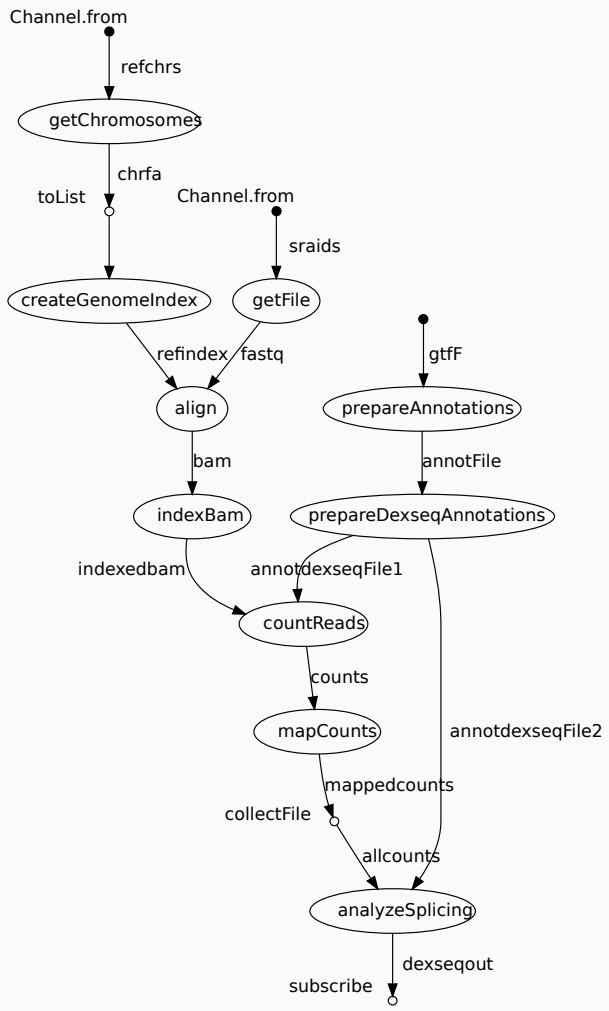


Figure 2: Exemple de workflow

Provenance

Principes données FAIR : trouvables, accessibles, interopérables et réutilisables

$\Rightarrow$ **PROVENANCE** :

garder la trace d'où proviennent les données (where-provenance) et de quelles manières elles ont été transformées (why-provenance)

Workflow

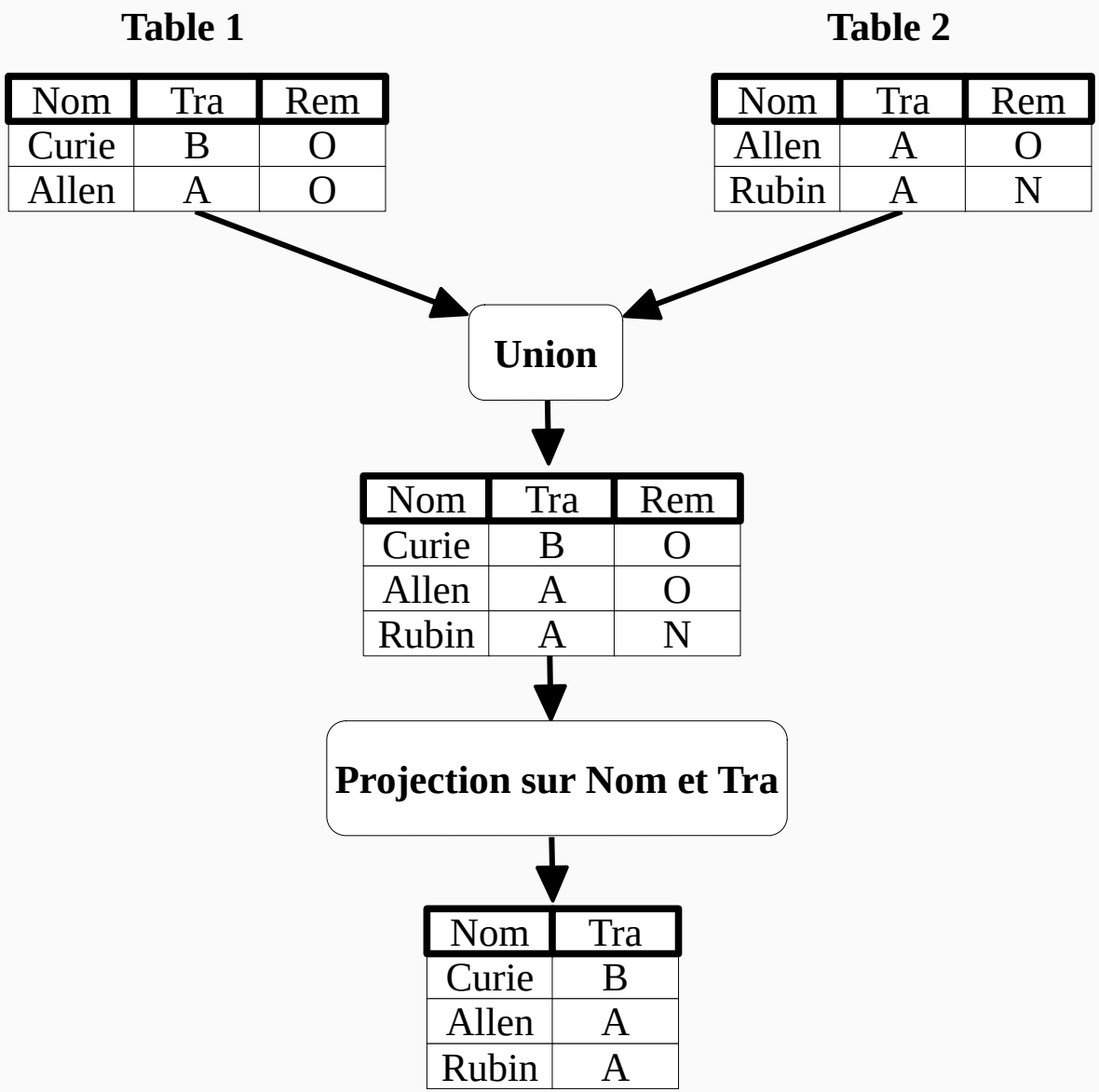
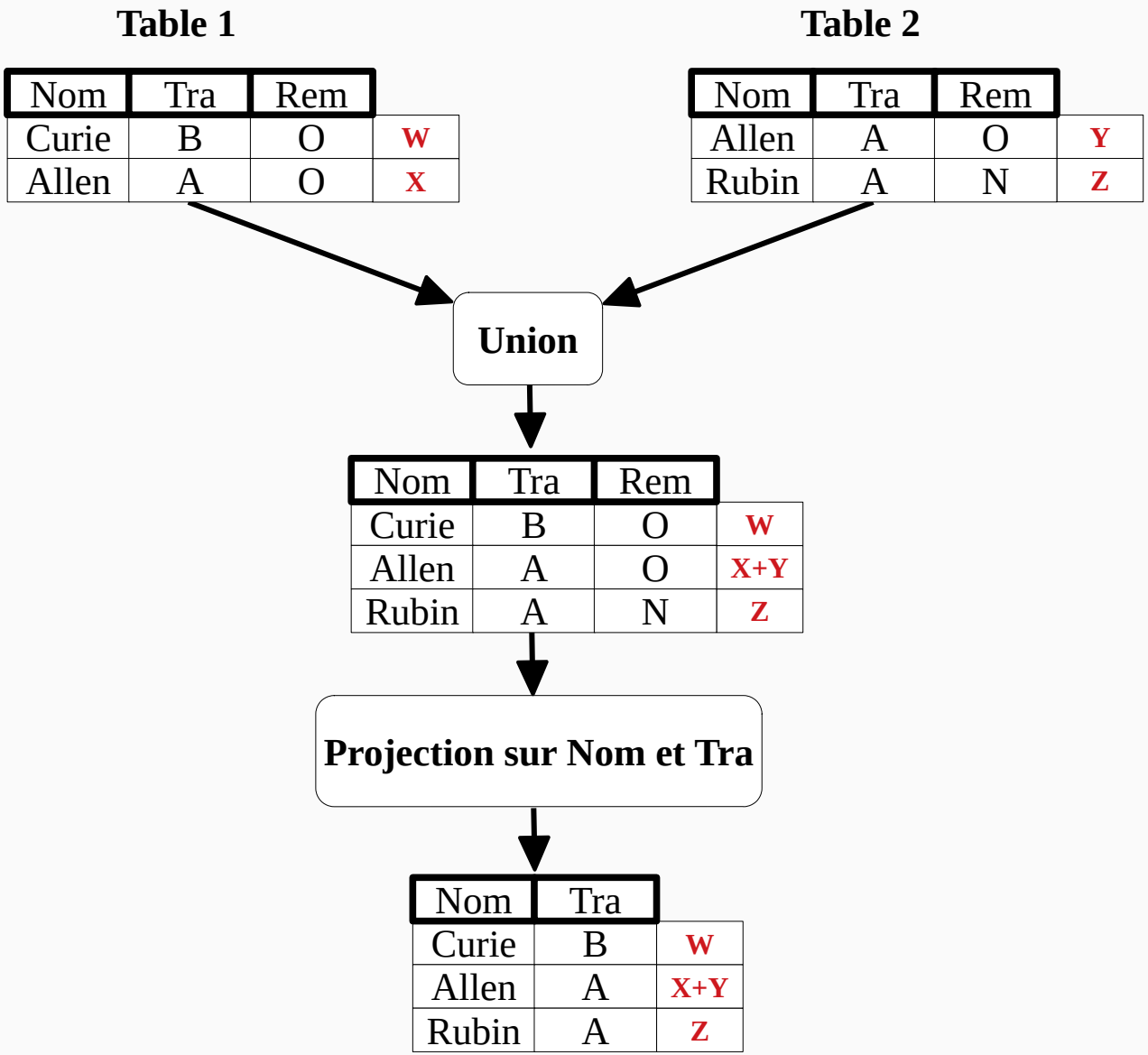


Table 1
Nom Tra Rem
Curie B O
Allen A O
Table 2
Nom Tra Rem
Allen A O
Rubin A N
Union
Nom Tra Rem
Curie B O
Allen A O
Rubin A N
Projection sur Nom et Tra
Nom Tra
Curie B
Allen A
Rubin A

Annotation d'un workflow



Objectif

Besoin de connaître la provenance des données mais aussi de garantir la correction de celle-ci

Notre approche : utiliser des méthodes formelles

Certifier en Coq la provenance des données issues d'analyses bioinformatiques

Contexte

Provenance

Nom	Rem	Tra
Rubin	O	A
Allen	N	B
Allen	N	B
Curie	O	B

Provenance

Nom	Rem	Tra
Rubin	O	A
Allen	N	B
Allen	N	B
Curie	O	B

Nom	Rem	Tra	Nb_pat
Rubin	O	A	1
Allen	N	B	$2 = 1 + 1$
Curie	O	B	1

n-uplet $\rightarrow \mathbb{N}$

relation $\equiv \mathbb{N}$-relation

Provenance

Nom	Rem	Tra
Rubin	O	A
Allen	N	B
Allen	N	B
Curie	O	B

Nom	Rem	Tra	Nb_pat
Rubin	O	A	1
Allen	N	B	$2 = 1 + 1$
Curie	O	B	1

Nom	Rem	Tra	Ann
Rubin	O	A	a
Allen	N	B	b + c
Curie	O	B	d

n-uplet $\rightarrow \mathbb{K}$

$\mathbb{K}$-relation

Union

	Nom	Rem	Tra	Ann
n_1	Rubin	O	A	a
n_2	Allen	N	B	**b**

Table 1: relation r_1 : APHP Paul Brousse

	Nom	Rem	Tra	Ann
n_3	Allen	N	B	**c**
n_4	Johnson	O	A	d

Table 2: relation r_2 : APHP G. Pompidou

	Nom	Rem	Tra	Ann
n_5	Rubin	O	A	a
n_6	Allen	N	B	**b + c**
n_7	Johnson	O	A	d

Table 3: union des relations r_1 et r_2

Semi-anneaux [Green et al., 2007, Amsterdamer et al., 2011]

$\mathbb{K}$ muni de 0, 1, $+$ et $*$: Semi-anneau commutatif

Algèbre étendue	Annotations
$\cup$	$+$
$\bowtie$	$*$
π	somme
σ_P	si P(.) alors . sinon 0
Γ	δ(somme)

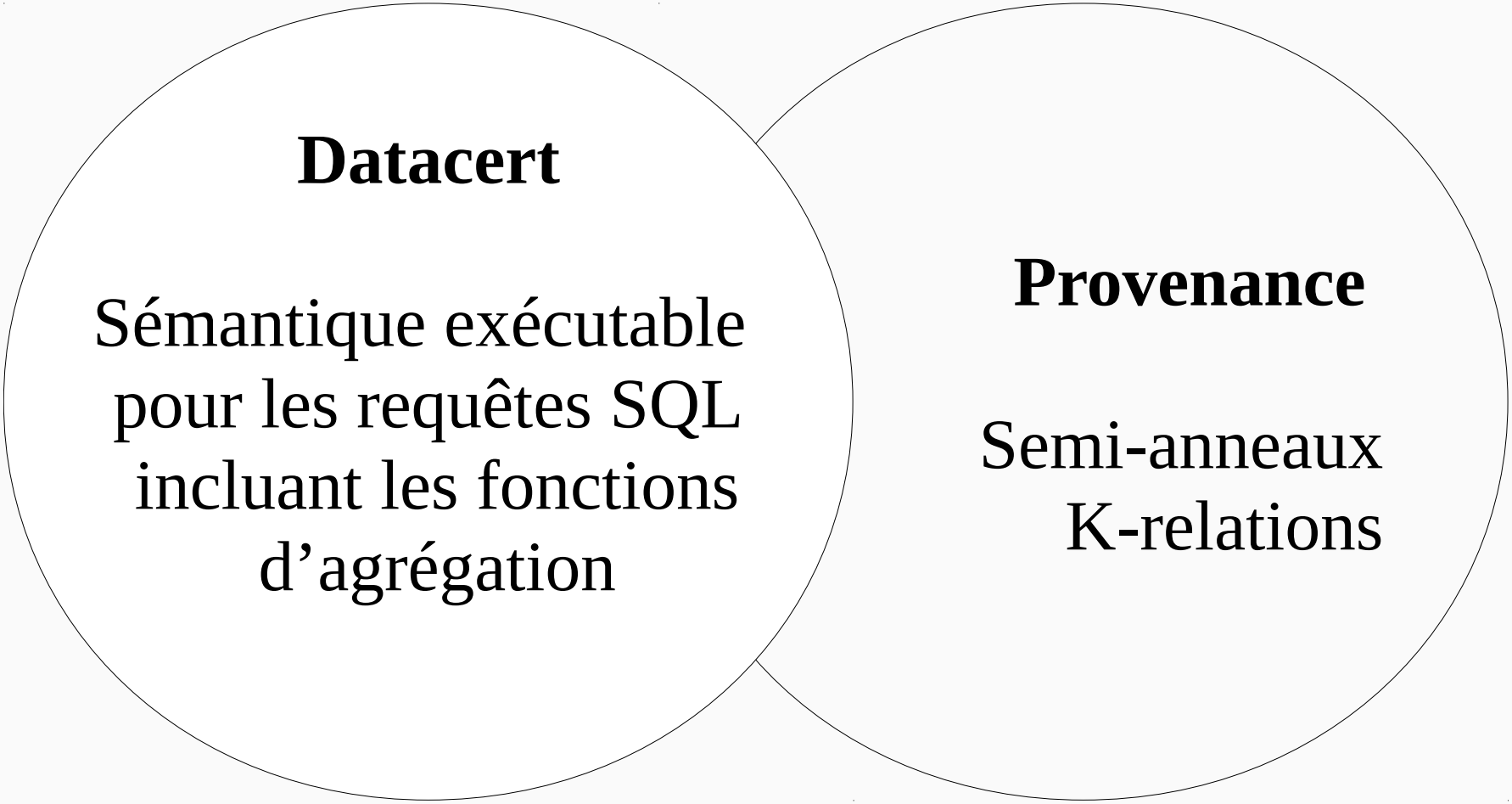
Semi-anneaux : entiers naturels ou tropicaux, sécurité, booléens, polynômes ...

Sommaire

Datacert [Benzaken and Contejean, 2012]

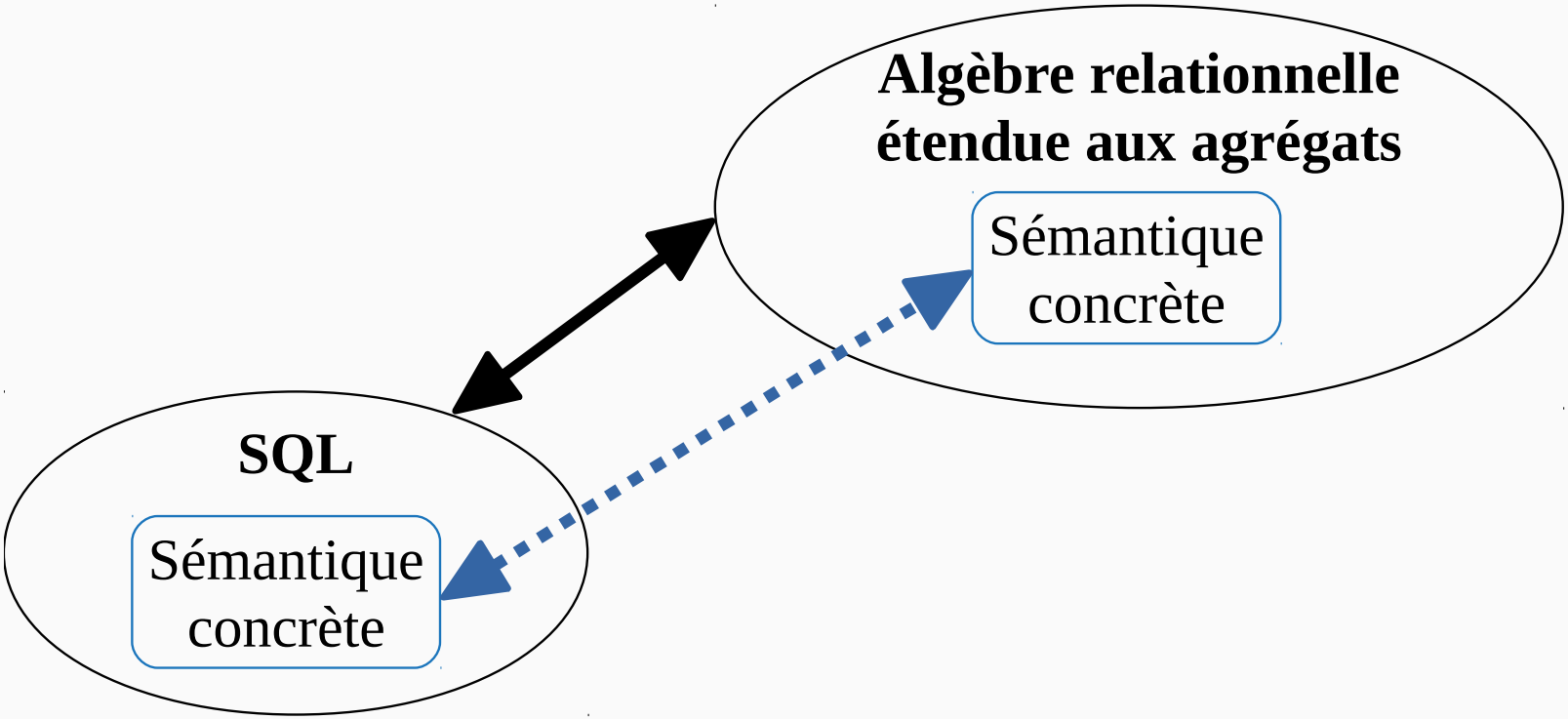
Datacert

Sémantique exécutable pour les requêtes SQL incluant les fonctions d'agrégation

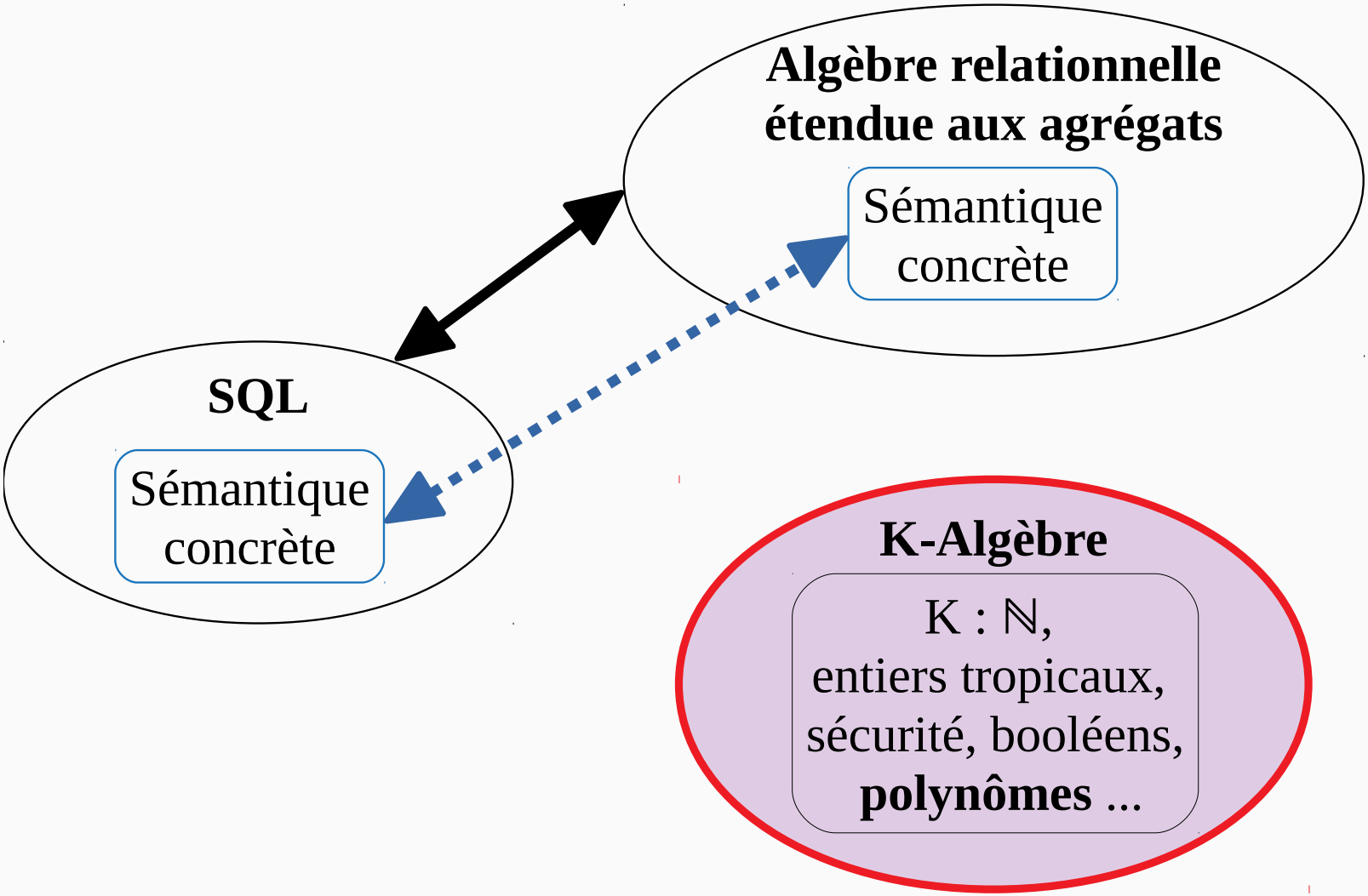

Datacert
Sémantique exécutable pour les requêtes SQL incluant les fonctions d'agrégation
Provenance
Semi-anneaux K-relations

Contributions

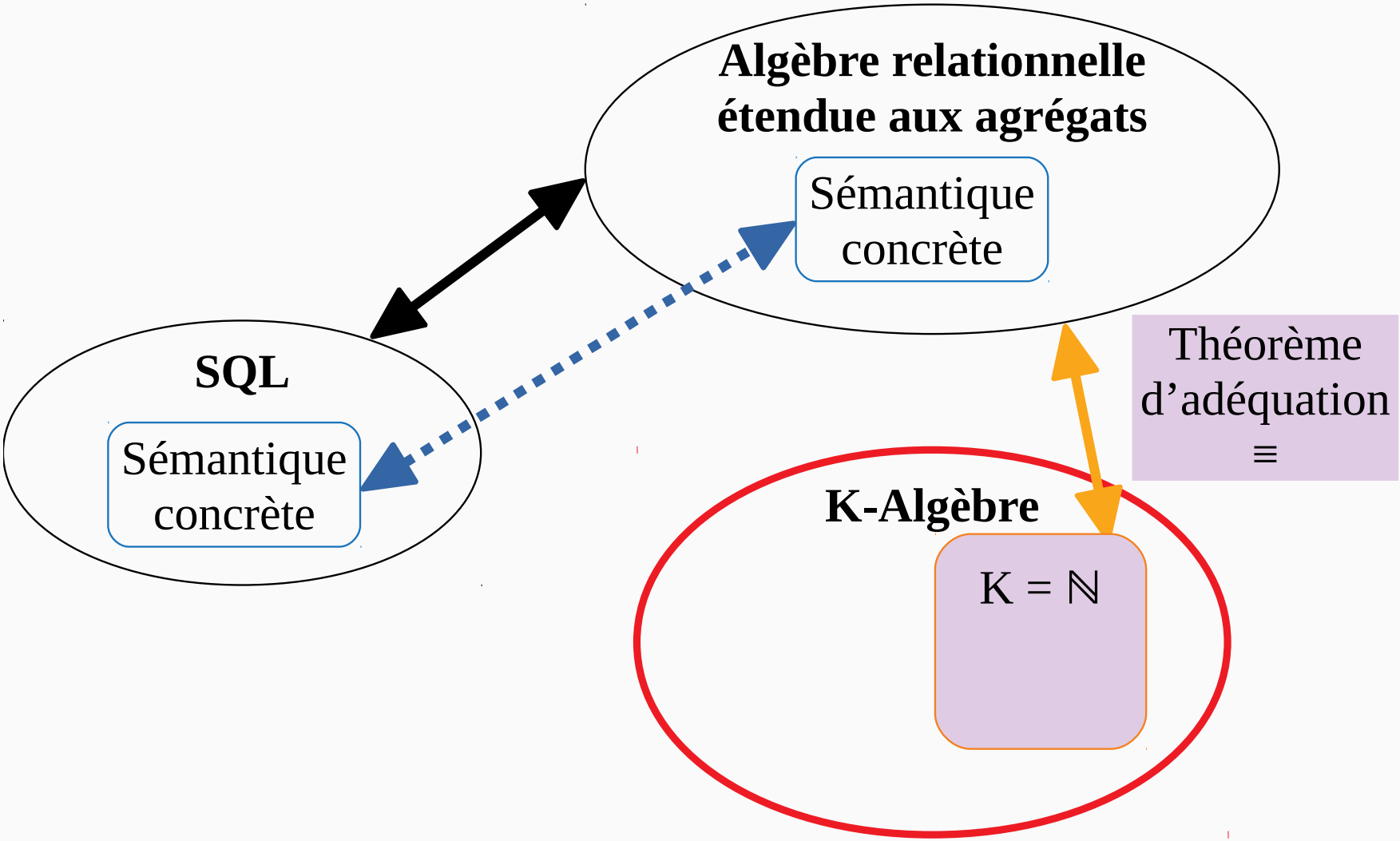
Why-provenance en Coq



Why-provenance en Coq



Why-provenance en Coq



Why-provenance en Coq

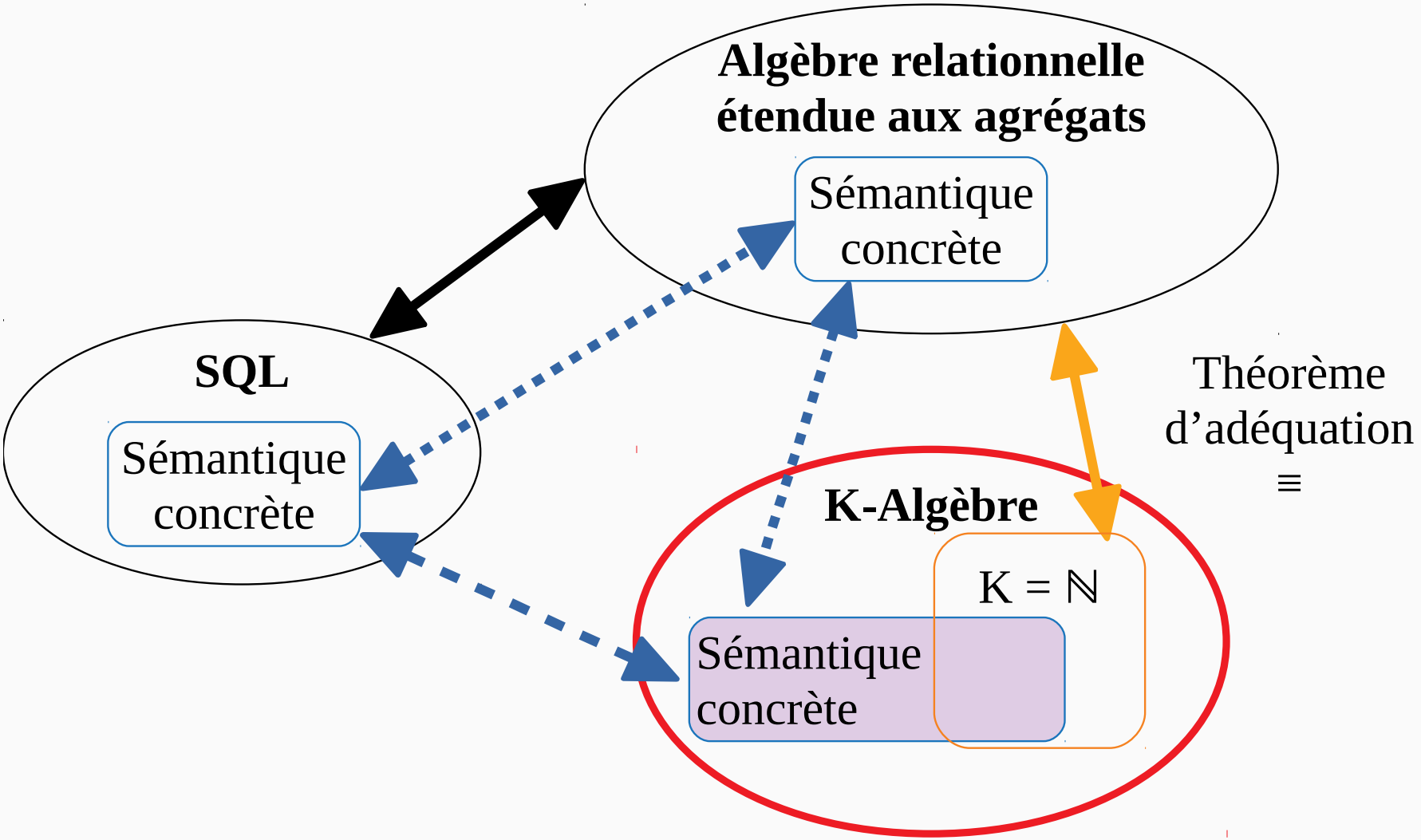
Théorème (Adéquation)

```
Theorem adequacy : ∀ q : query,
  well_formed q = true →
  ∀ (t:tuple),
    f (eval_query_prov q) t =
    nb_occurences t (eval_query_rel q).
```

eval_query : renvoie la table résultante par rapport à une requête

f : associe la table et le n-uplet à une annotation

Why-provenance en Coq



Caractérisation de propriétés

Caractériser des propriétés sur les données à partir de la provenance

Caractérisation de propriétés

A	B	C	Annot
1	3	2	x

Table 4: relation r_5

A	B	C	Annot
6	3	7	y

Table 5: relation r_6

A	B	C	Annot
1	3	7	x∗y

Table 6: $\pi_{AB}(r_5) \bowtie \pi_{BC}(r_6)$

A	B	C	Annot
6	3	2	x∗y

Table 7: $\pi_{AB}(r_6) \bowtie \pi_{BC}(r_5)$

Propriété

Existe-t-il un semi-anneau tel que si deux n-uplets ont la même annotation cela implique qu'ils sont égaux ?

Caractérisation de propriétés

A	B	C	Annot
1	3	2	{({A,B,C},x)}

Table 4: relation r_5

A	B	C	Annot
6	3	7	{({A,B,C},y)}

Table 5: relation r_6

A	B	C	Annot
1	3	7	{({A,B},x),({B,C},y)}

Table 6: $\pi_{AB}(r_5) \bowtie \pi_{BC}(r_6)$

A	B	C	Annot
6	3	2	{({B,C},x),({A,B},y)}

Table 7: $\pi_{AB}(r_6) \bowtie \pi_{BC}(r_5)$

Propriété

Existe-t-il un semi-anneau tel que si deux n-uplets ont la même annotation cela implique qu'ils sont égaux ?

Caractérisation de propriétés

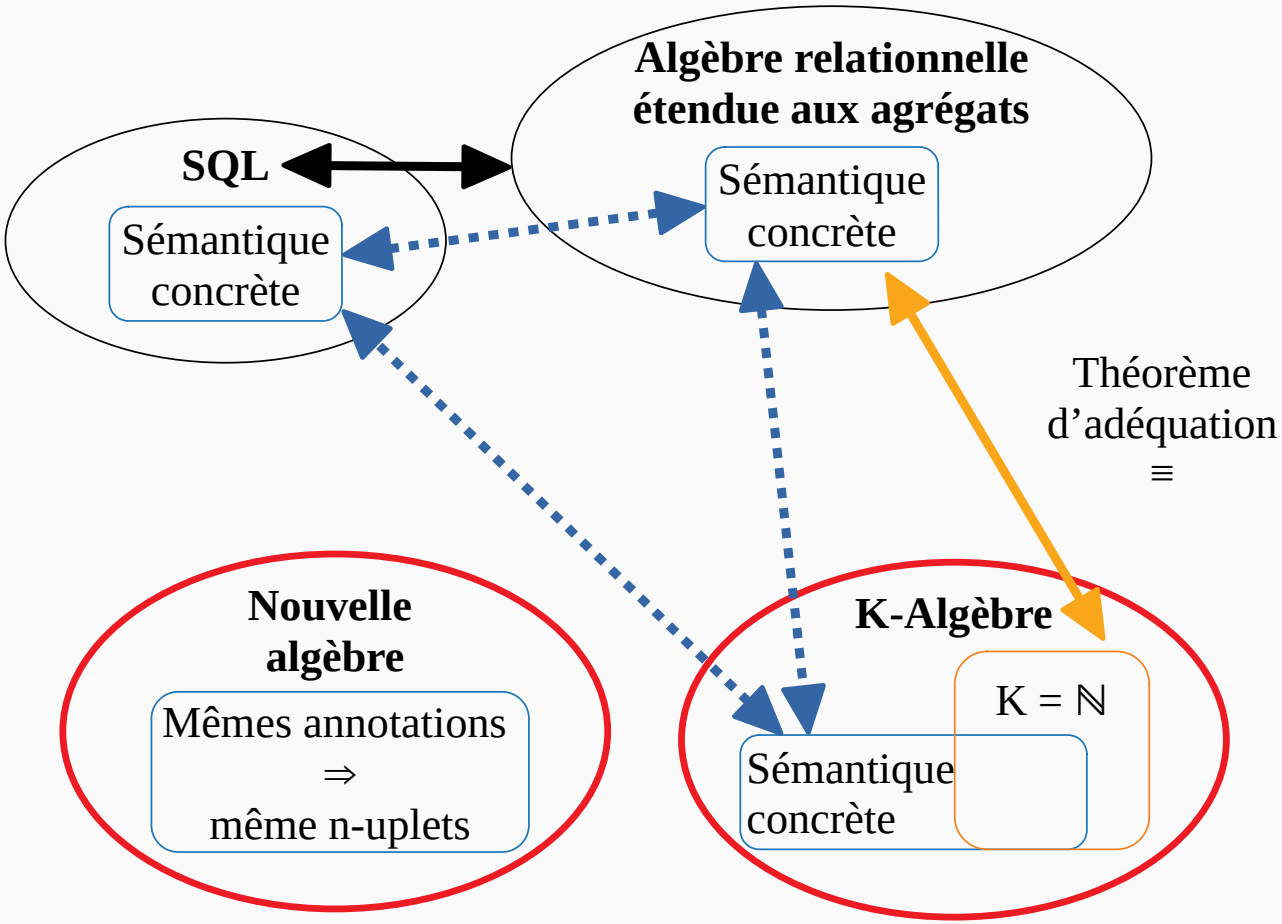
Algèbre	**Why-provenance**	**Annotations ayant cette caractéristique**
$\cup$	$+$	$. \cup .$
$\bowtie$	$*$	$. \cup .$
π_A	somme	$(A \cap ., .)$
σ_P	si P(.) alors . sinon 0	si P(.) alors . sinon $\emptyset$
Γ	δ(somme)	

Caractérisation de propriétés

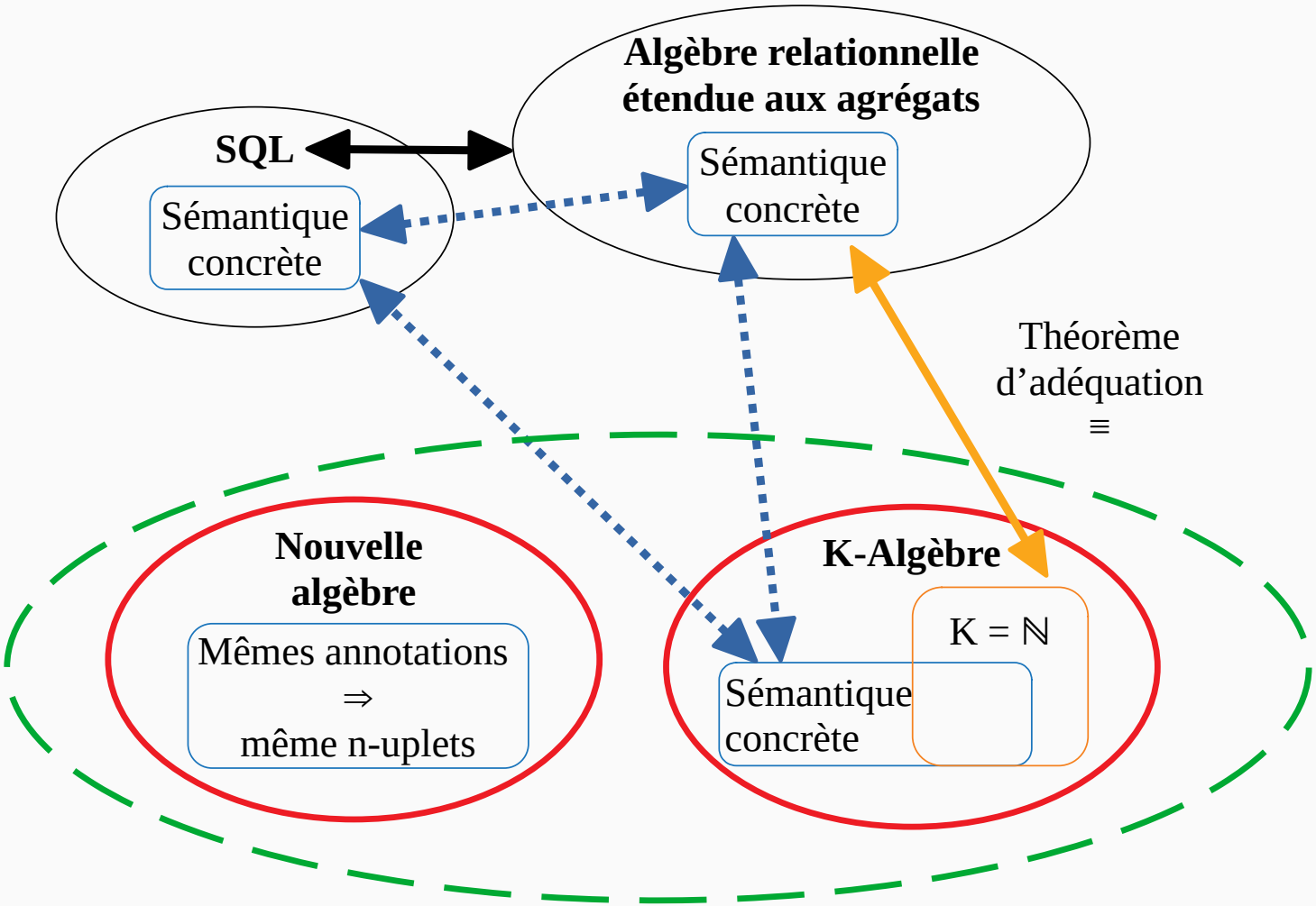
Théorème (Egalité Anotations-nuplets)

```
Lemma equal_tuple : ∀ (q1 q2 : query) (t1 t2 : tuples),
    equal (annot q1 t1) (annot q2 t2) = true →
    is_zero (annot q1 t1) = false →
    eq t1 t2.
```

Contributions



Perspectives



Pistes envisagées

Pistes envisagées

Trouver d'autres caractéristiques

Application aux workflows : caractériser des propriétés pour des "boites noires"

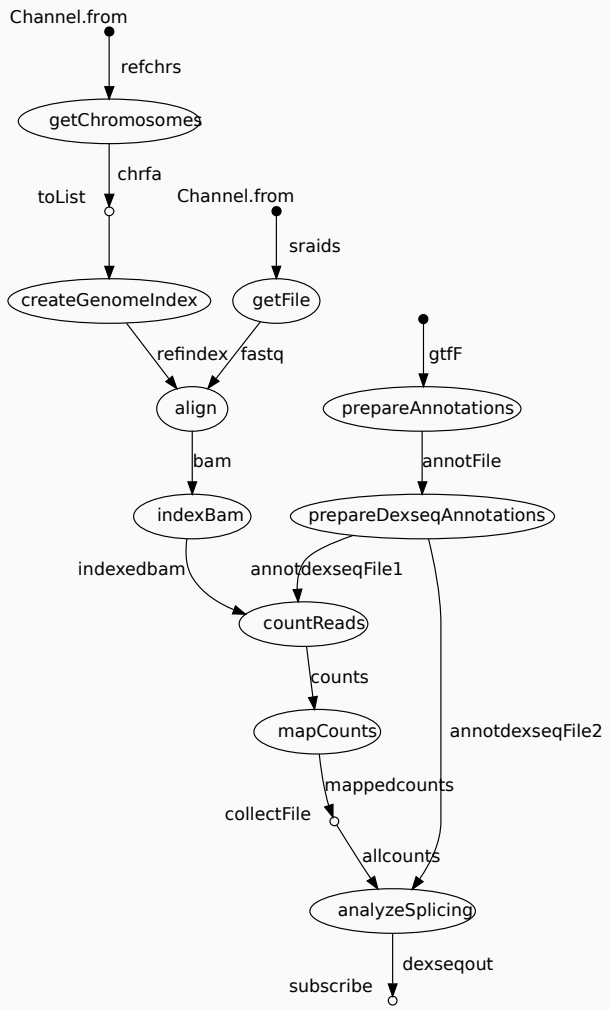


Figure 3: Exemple de workflow

Pistes envisagées

- Extraction du code en OCaml, benchmark et comparaison avec d'autres implémentations de la provenance


Amsterdamer, Y., Deutch, D., and Tannen, V. (2011).
Provenance for aggregate queries.
In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 153–164. ACM.


Benzaken, V. and Contejean, E. (2012).
The datacert library (`http://datacert.lri.fr/`).


Green, T. J., Karvounarakis, G., and Tannen, V. (2007).
Provenance semirings.
In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40. ACM.

Why-provenance en Coq

Nombre de lignes de code Coq: 15300

Publications

- **A Coq formalization of data provenance**, *Véronique Benzaken, Sarah Cohen-Boulakia, Évelyne Contejean, Chantal Keller et Rébecca Zucchini*, CPP - 11th International Conference on Certified Programs and Proofs - 2021
- **Vers une formalisation en Coq de la provenance de données**, *Véronique Benzaken, Sarah Cohen-Boulakia, Évelyne Contejean, Chantal Keller et Rébecca Zucchini*, JFLA - 31 ème Journées Francophones des Langages Applicatifs - 2020