Service de diffusion des données de recherche: de l'astronomie à l'ethnographie.

A. Tricoche, F. Weber, C.M. Zwölf

Contexte de travail

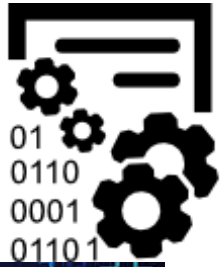
Collaboration entre PADC et l'ENS au sein de l'Initiative de Recherche Interdisciplinaires Stratégiques (IRIS) Science des données, données de la science (SDDS).

PADC a un rôle de conseil et transfert de compétences et technologies pour la gestion/diffusion des données de recherche.

Spécificité à prendre en compte dans les sciences humaines: caractère confidentiel des données personnelles.

Les données de recherche: plus des soucis que des bienfaits?

Données



Problèmes

Méthodologiques

Ethiques

Techniques
Organisationnels

Les données de recherche: plus des soucis que des bienfaits?

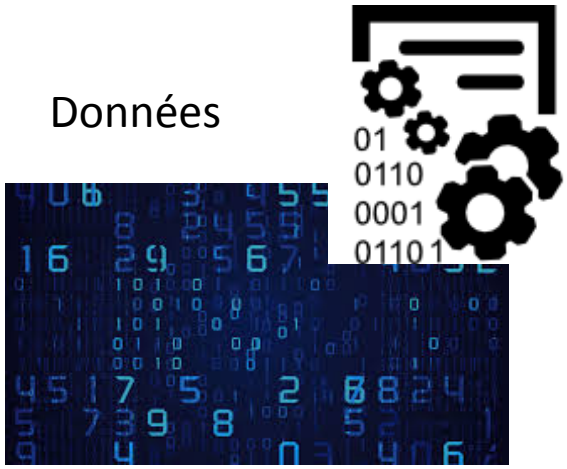
Données

Problèmes

- Méthodologiques
  - Définition univoque de la signification des données et métadonnées
  - Explicitation du processus de production des données dans une relation d'enquête
  - Possibilité de citer une donnée dans un papier scientifique
- Ethiques
- Techniques Organisationnels

Les données de recherche: plus des soucis que des bienfaits?

Données

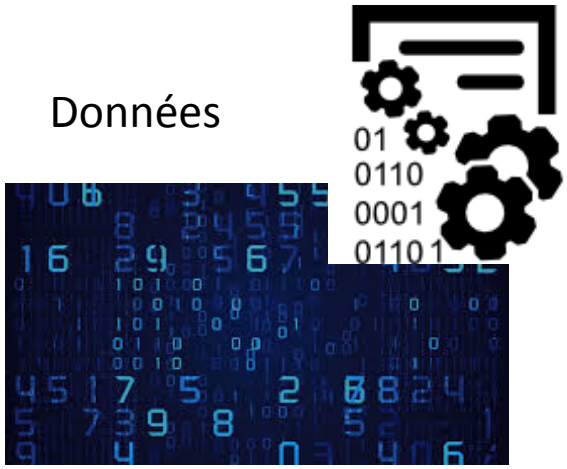


Problèmes

- Méthodologiques
  - Définition univoque de la signification des données et métadonnées
  - Explicitation du processus de production des données dans une relation d'enquête
  - Possibilité de citer une donnée dans un papier scientifique
- Ethiques
  - Protection des données personnelles sensibles
  - Droit de réutilisation des travaux d'autrui
  - Contrôle par les pairs et mutualisation
- Techniques Organisationnels

Les données de recherche: plus des soucis que des bienfaits?

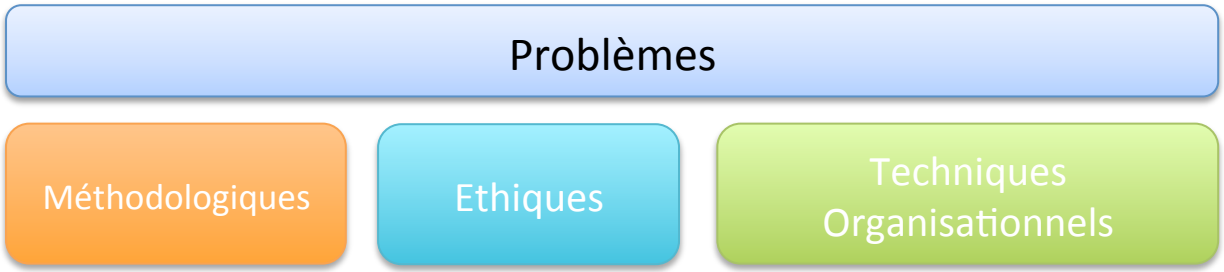
Données



Problèmes

- Méthodologiques
  - Définition univoque de la signification des données et métadonnées
  - Explicitation du processus de production des données dans une relation d'enquête
  - Possibilité de citer une donnée dans un papier scientifique
- Ethiques
  - Protection des données personnelles sensibles
  - Droit de réutilisation des travaux d'autrui
  - Contrôle par les pairs et mutualisation
- Techniques Organisationnels
  - Construction de l'infrastructure du centre de données (Physique / Logicielle)
  - Coûts de développement et maintien des éléments logiciels
  - Coûts de maintenance et d'exploitation du centre de données
  - Pérennisation du savoir-faire technique

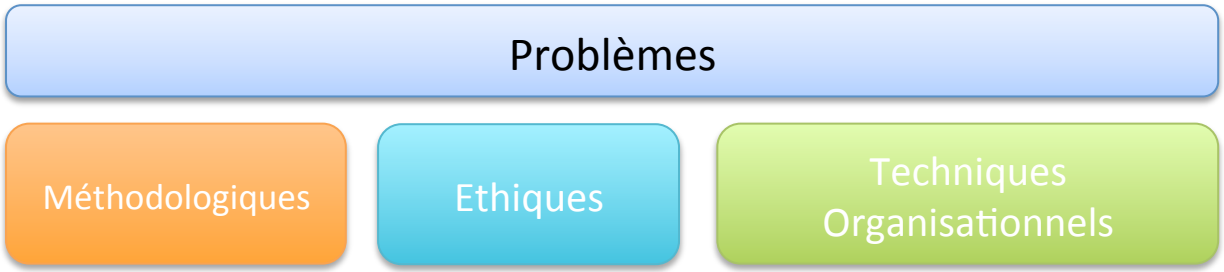
Les données de recherche: plus des soucis que des bienfaits?



Des solutions fragmentées ne sont pas adaptées:

- À affronter les défis technologiques posés par le '*data deluge*'
- À affronter le panorama des réductions d'effectifs et de support à la recherche.

Les données de recherche: plus des soucis que des bienfaits?



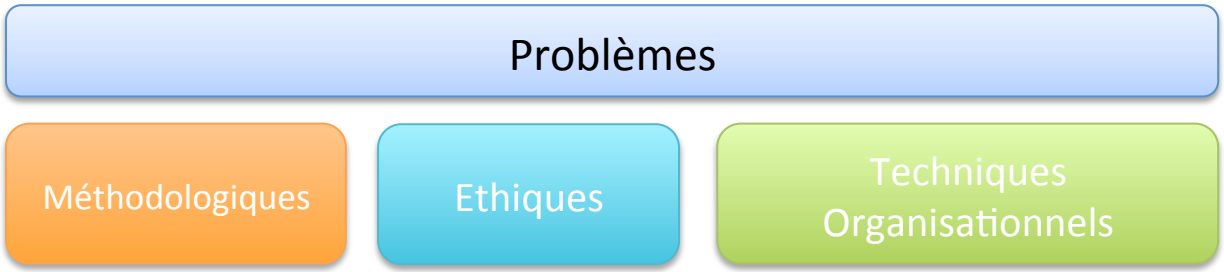
Des solutions fragmentées ne sont pas adaptées:

- À affronter les défis technologiques posés par le '*data deluge*'
- À affronter le panorama des réductions d'effectifs et de support à la recherche.

La solution que nous proposons (et que nous sommes en train de développer)

- S'attaque directement aux trois problématiques identifiées
- Est suffisamment générique pour être adaptée à différentes disciplines

Les données de recherche: plus des soucis que des bienfaits?



Des solutions fragmentées ne sont pas adaptées:

- À affronter les défis technologiques posés par le '*data deluge*'
- À affronter le panorama des réductions d'effectifs et de support à la recherche.

La solution que nous proposons (et que nous sommes en train de développer)

- S'attaque directement aux trois problématiques identifiées
- Est suffisamment générique pour être adaptée à différentes disciplines

Nous réalisons actuellement un démonstrateur sur la refonte de l'archive ArchEthno.

L’expérience d’ArchEthno

- Utile pour l’archivage scientifique des données de la sociologie qualitative

L'expérience d'ArchEthno

- Utile pour l'archivage scientifique des données de la sociologie qualitative
  - Suppose une question scientifique déjà stabilisée
  - Autorise trois niveaux de confidentialité: montrer la « cuisine de la recherche » à tous publics, autoriser le contrôle par les pairs (referees de revues, jurys de soutenance), interdire l'identification personnelle (anonymat des contextes et des cas)
  - En ethnographie réflexive réduite, permet une comparaison de l'entrée dans différents contextes et de la circulation dans différents cas

  ⇒ Archivage délégable à un tiers

L'expérience d'ArchEthno

- Utile pour l'archivage scientifique des données de la sociologie qualitative
  - Suppose une question scientifique déjà stabilisée
  - Autorise trois niveaux de confidentialité: montrer la « cuisine de la recherche » à tous publics, autoriser le contrôle par les pairs (referees de revues, jurys de soutenance), interdire l'identification personnelle (anonymat des contextes et des cas)
  - En ethnographie réflexive réduite, permet une comparaison de l'entrée dans différents contextes et de la circulation dans différents cas

  ⇒ Archivage délégable à un tiers

- Utile pour l'analyse « à chaud » des données ethnographiques

L'expérience d'ArchEthno

- Utile pour l'archivage scientifique des données de la sociologie qualitative
  - Suppose une question scientifique déjà stabilisée
  - Autorise trois niveaux de confidentialité: montrer la « cuisine de la recherche » à tous publics, autoriser le contrôle par les pairs (referees de revues, jurys de soutenance), interdire l'identification personnelle (anonymat des contextes et des cas)
  - En ethnographie réflexive réduite, permet une comparaison de l'entrée dans différents contextes et de la circulation dans différents cas

  ⇒ Archivage délégable à un tiers

- Utile pour l'analyse « à chaud » des données ethnographiques
  - Privilégier la notation des « surprises » de l'observateur
  - Construire la question scientifique et les variables pertinentes
  - Séparer le « cas » à analyser de son « contexte » à décrire
  - Protéger ce que le chercheur considère comme la « cuisine privée » de la recherche

  => Autosaisie non délégable à un tiers

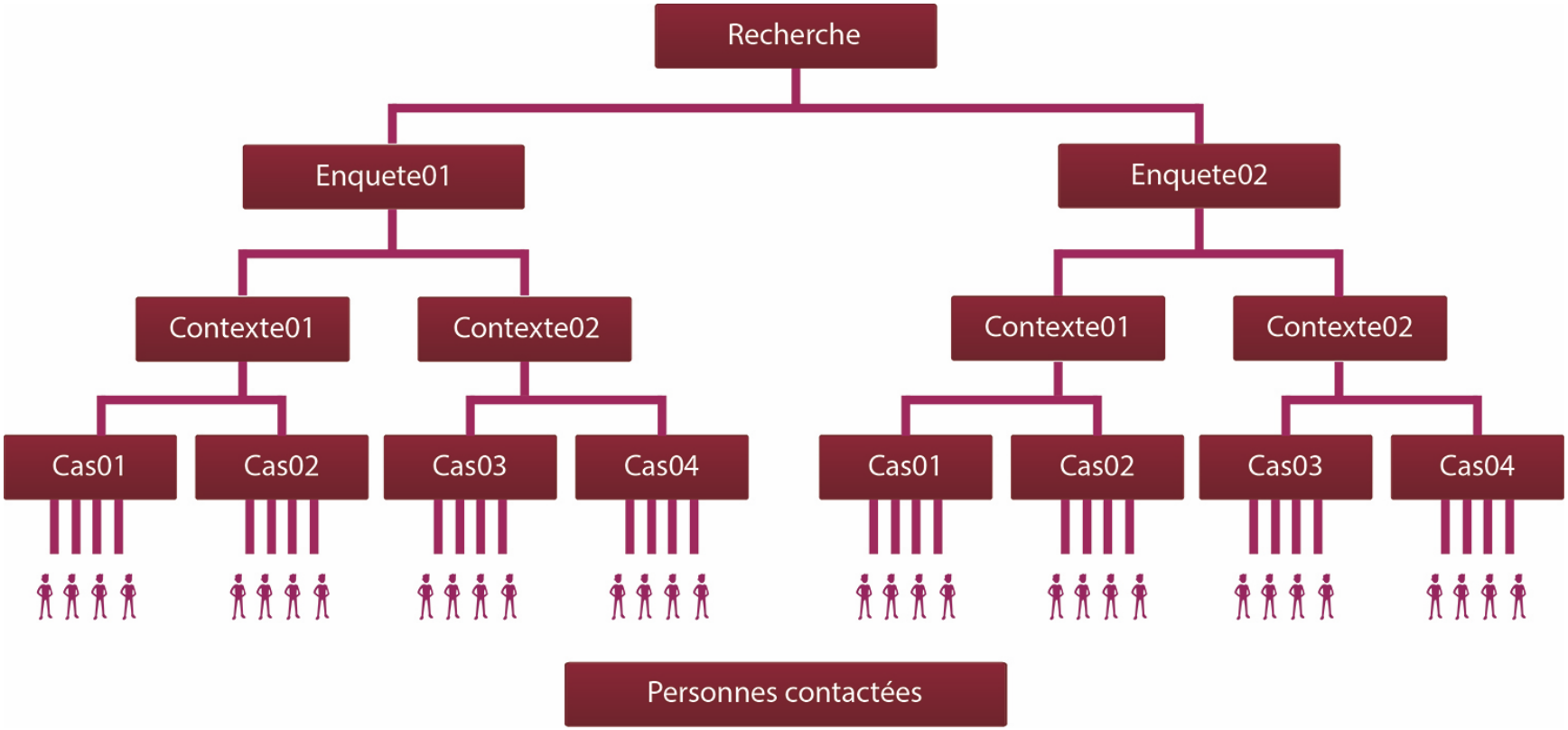
L'archivage scientifique des données de l'enquête MEDIPS-Alzheimer

- Une base adaptée à l'ethnographie réflexive réduite

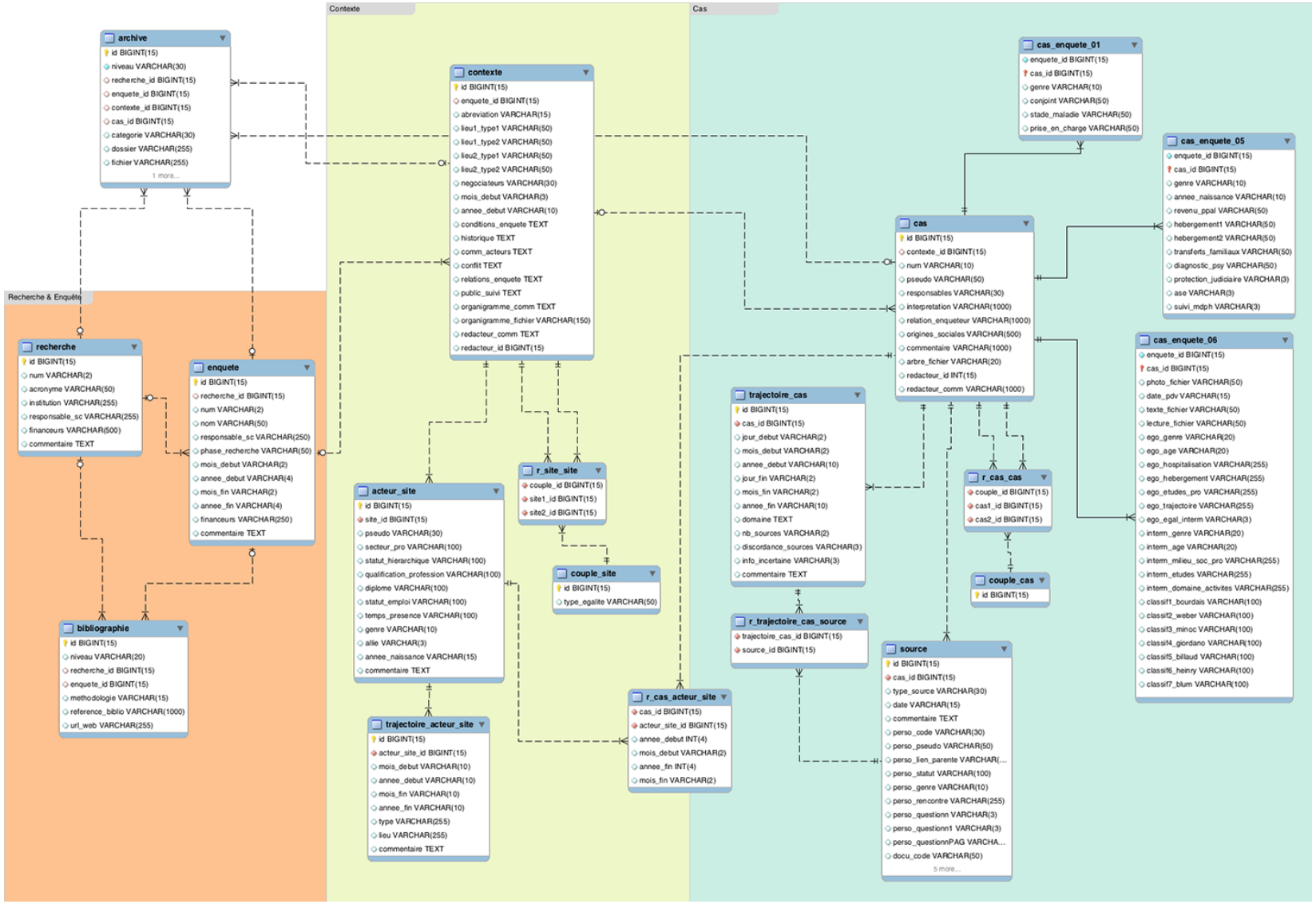
- Une question de recherche: comment s'organisent les « familles » face à la « dépendance » d'un des leurs?

- Une enquête à deux niveaux: le « cas » (monographie de famille) permet de répondre à la question de recherche, le « contexte » (l'institution médicale ou sociale qui nous donne accès à des listes d'adresses) permet d'analyser le domaine socio-historique de validité de nos résultats statistiques sur les cas

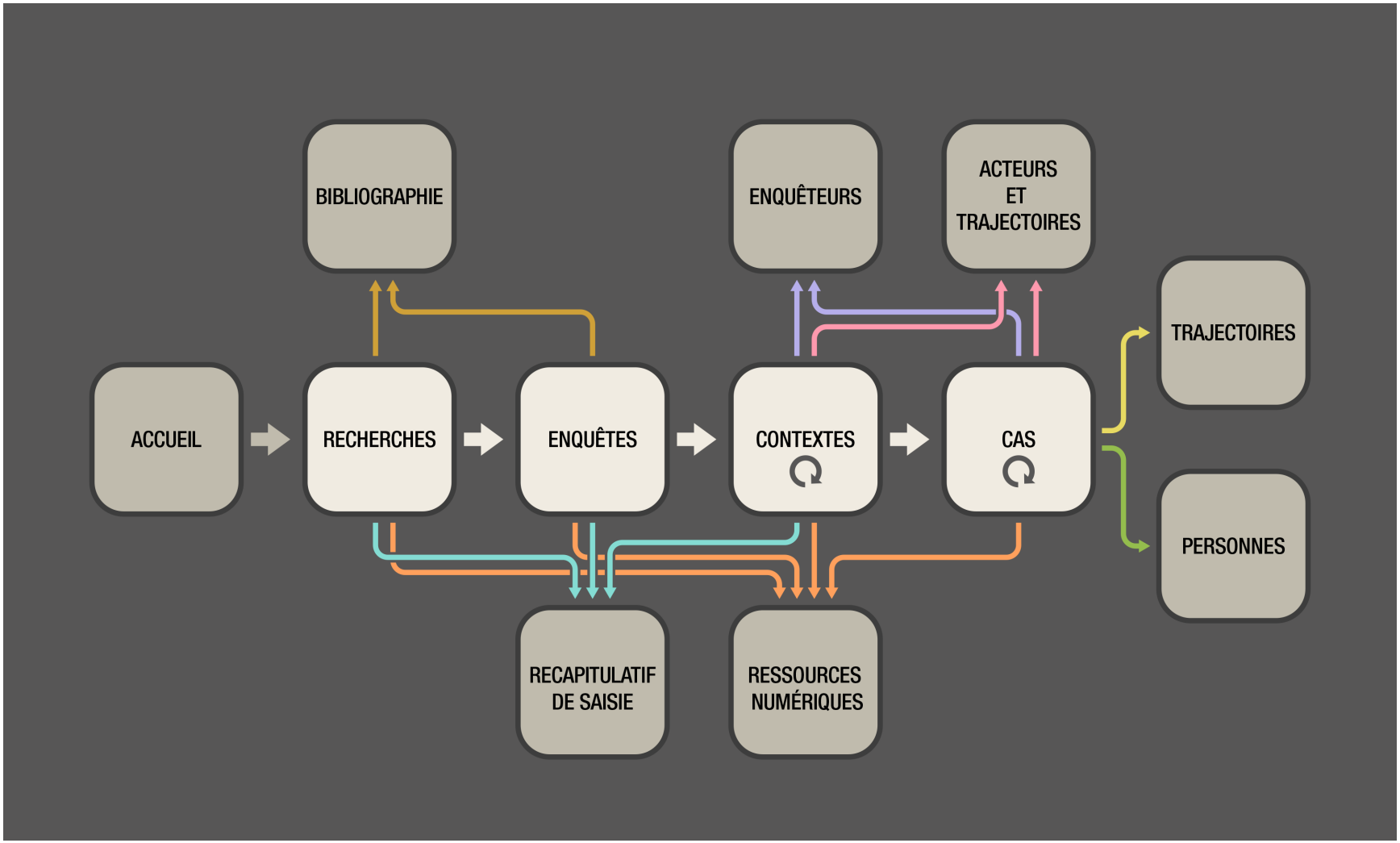
Les principales entités d'ArchEthno



Modèle de données du prototype

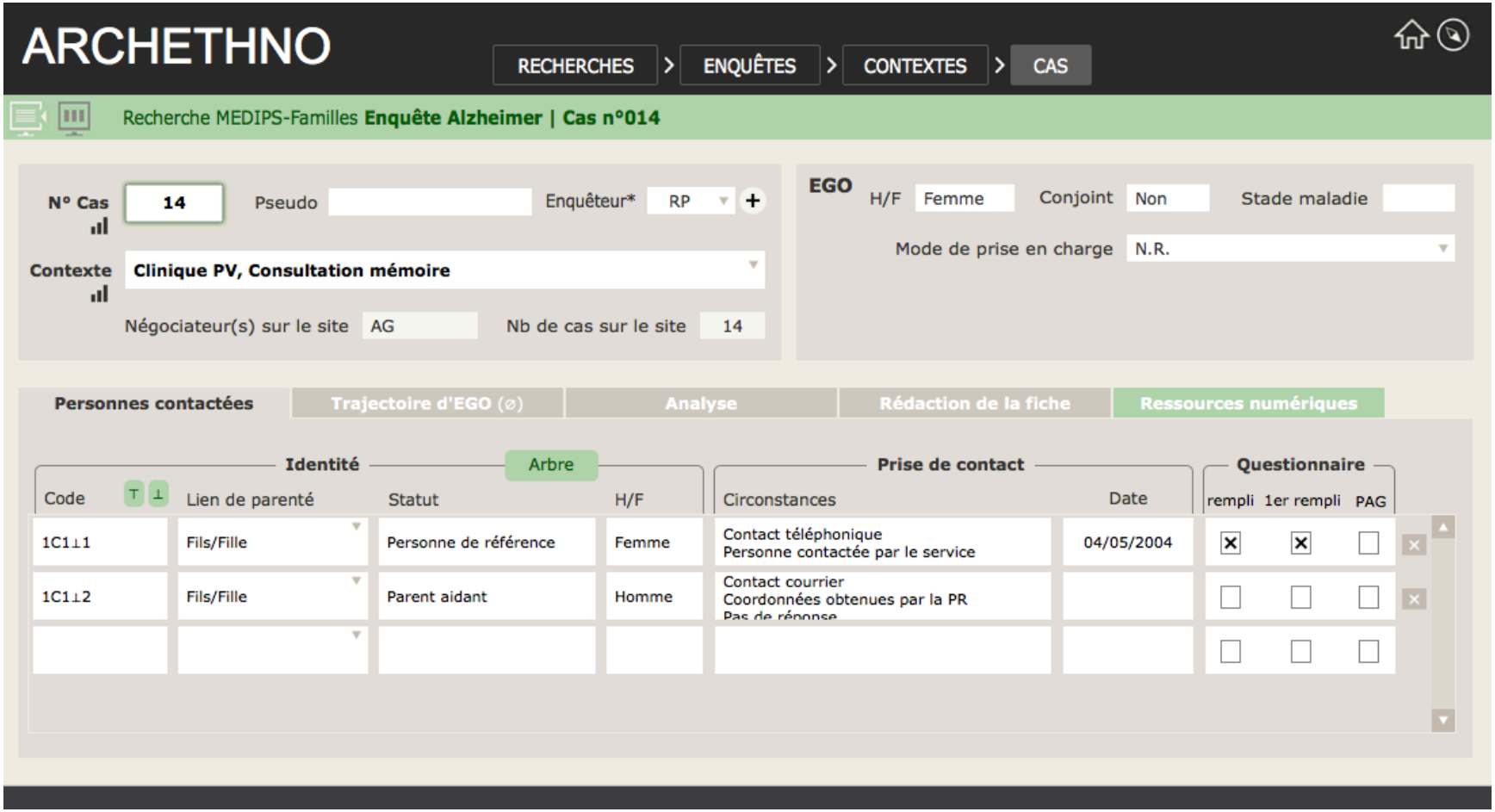


Plan de navigation du prototype



Exemple de notice dans le prototype

(Enquête MEDIPS-Alzheimer : Cas n°14)



Besoins d'évolutions

Différentes enquêtes peuvent avoir des métadonnées différentes. Dans la configuration actuelle cela oblige à

- Modifier le modèle de données (schéma de la base)
- Modifier les interfaces de saisies
- Modifier les interfaces de consultations

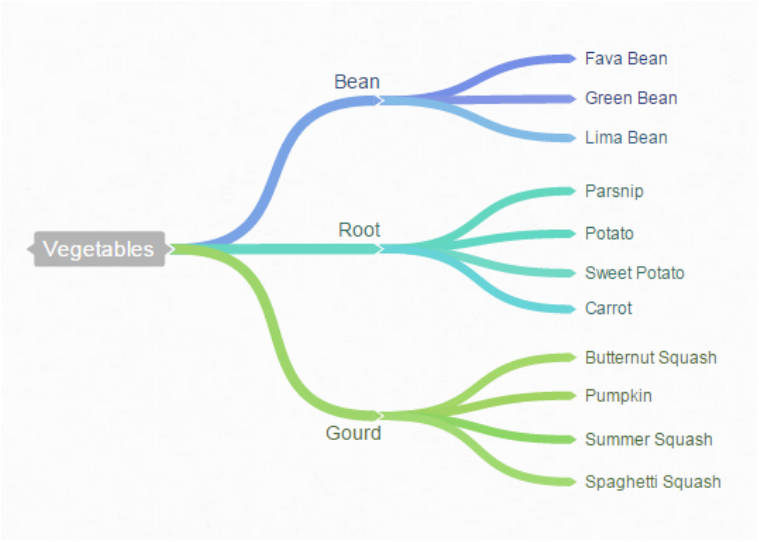
La réflexion et la refonte en cours a pour but de dépasser ces difficultés en fournissant plus de flexibilité.

Migration vers une structure SKOS

Concept	Top-concept	Père	Type	Requis	Nb Occurrences	Liste de valeurs	Définition
Recherche	true			false			Ensemble d'enquêtes liées entre elles par une question
Recherche_Acronyme	false	Recherche	texte	true	1		Nom sous lequel se réfèrent à cette recherche le ou les
Recherche_Institution	false	Recherche	texte	true	-1		Employeur du ou des responsables scientifiques
Recherche_ResponsableSc	false	Recherche	texte	true	-1		Nom du ou des chercheurs ayant formulé la question commune
Recherche_Question	false	Recherche	texte	true	1		Formulation de la question et présentation des liens entre les
FinanceurRecherche	true	Recherche		false			Contrat obtenu pour cet ensemble d'enquêtes
FinanceurRecherche_NomF	false	FinanceurRecherche	texte	true	1		Organisme ayant versé une subvention à l'employeur du ou des
FinanceurRecherche_Statut	false	FinanceurRecherche	texte	true	1	Privé,Public sans précision,Public	Fonds privés ou publics et à quelle échelle territoriale
FinanceurRecherche_Mont	false	FinanceurRecherche	nombre réel	false	1		Somme reçue en euros par l'employeur du responsable
Enquete	true	Recherche		false			Dispositif empirique permettant de répondre à la question de
Enquete_Nom	false	Enquete	texte	true	1		Nom sous lequel se réfèrent à l'enquête le ou les responsables
Enquete_ResponsableScie	false	Enquete	texte	true	-1		Nom du chercheur responsable de l'enquête
Enquete_PhaseRecherche	false	Enquete	texte	true	1	réduite,ethnographique	Place de l'enquête dans la recherche: ethnographie exploratoire
Enquete_MoisDebut	false	Enquete	entier	true	1	1,2,3,4,5,6,7,8,9,10,11,12	Date de commencement de l'enquête (mois)
Enquete_AnneeDebut	false	Enquete	entier	true	1		Date de commencement de l'enquête (année)
Enquete_MoisFin	false	Enquete	entier	false	1	1,2,3,4,5,6,7,8,9,10,11,12	Date de fin de l'enquête (mois)
Enquete_AnneeFin	false	Enquete	entier	false	1		Dates de fin de l'enquête (année)
Enquete_Objet	false	Enquete	texte	true	1		Présentation du lien entre l'objet de l'enquête et la question de
FinanceurEnquete	true	Enquete		false			Contrat obtenu pour cette enquête, y compris contrat doctoral
FinanceurEnquete_NomFin	false	FinanceurEnquete	texte	true	1		Organisme ayant permis l'embauche éventuelle du responsable
FinanceurEnquete_StatutFi	false	FinanceurEnquete	texte	true	1	Privé,Public, sans précision,Public,	Fonds privés ou publics et à quelle échelle territoriale
FinanceurEnquete_Montant	false	FinanceurEnquete	nombre réel	false	1		Somme reçue en euros pour financer l'enquête
Contexte	true	Enquete		false			Ensemble de données communes à une partie des cas (y
Contexte_Abreviation	false	Contexte	texte	true	1		Nom sous lequel le responsable de l'enquête se réfère au
Contexte_Negociateur	false	Contexte	ID	true	-1		Initiales du responsable de la circulation du ou des enquêteurs
Contexte_LieuN1TypeN1	false	Contexte	texte	true	1		Catégorie scientifique à laquelle appartient le contexte: institution
Contexte_LieuN1TypeN2	false	Contexte	texte	false	1		Catégorie administrative dont relève éventuellement le contexte:
Contexte_LieuN2TypeN1	false	Contexte	texte	false	1		Catégorie indigène du niveau hiérarchique le plus bas où
Contexte_LieuN2TypeN2	false	Contexte	texte	false	1		Nom indigène du niveau où le ou les enquêteurs sont présents

Description de la méthode suivie

Inspiré du web sémantique
(W3C **S**tandard **S**imple **K**noweledge
Organisation **S**ystem, SKOS)



Description de la méthode suivie

Inspiré du web sémantique
(W3C **S**tandard **S**imple **K**noweledge **O**rganisation **S**ystem, SKOS)

On définit tous les concepts liés aux données et aux métadonnées de recherche

Ces concepts sont consigné dans un *thesaurus.*

Le *thesaurus* est représenté informatiquement sous forme d'un fichier **SKOS/RDF** (XML)

Description de la méthode suivie

Fichier SKOS

- ✔ Pérennité : on cristallise le savoir-faire scientifique.
- ✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

Description de la méthode suivie

Fichier SKOS

- ✔ Pérennité : on cristallise le savoir-faire scientifique.
- ✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

- ✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

Description de la méthode suivie

Fichier SKOS

✔ Pérennité : on cristallise le savoir-faire scientifique.
✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

Avec les SKOS concepts, une simple table permet de remplacer tout modèle de données

SKOS	Valeur
Concept « Prénom de l'auteur»	Gustave
Concept « Nom de l'auteur »	Flaubert

Des évolutions du modèle conceptuel (ex. Nouveaux concepts) n'affectent pas la structure de la table, uniquement la modélisation SKOS.

Description de la méthode suivie

Fichier SKOS

✔ Pérennité : on cristallise le savoir-faire scientifique.
✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

Avec les SKOS concepts, une simple table permet de remplacer tout modèle de données

Identifiant unique persistant	SKOS	Valeur
7a8a34f4-a7aa-11e7-abc4	Concept « Prénom de l'auteur»	Gustave
a975361a-a7aa-11e7-abc4	Concept « Nom de l'auteur »	Flaubert

Description de la méthode suivie

Fichier SKOS

✔ Pérennité : on cristallise le savoir-faire scientifique.
✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

Avec les SKOS concepts, une simple table permet de remplacer tout modèle de données

Identifiant unique persistant	SKOS	Valeur
7a8a34f4-a7aa-11e7-abc4	Concept « Prénom de l'auteur»	Gustave
a975361a-a7aa-11e7-abc4	Concept « Nom de l'auteur »	Flaubert

Grâce aux identifiants uniques persistants, il est possible de citer une donnée avec une granularité très fine.

Description de la méthode suivie

Fichier SKOS

✔ Pérennité : on cristallise le savoir-faire scientifique.
✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

Avec les SKOS concepts, une simple table permet de remplacer tout modèle de données

Identifiant unique persistant	SKOS	Valeur	Niveau visibilité
7a8a34f4-a7aa-11e7-abc4	Concept « Prénom de l'auteur»	Gustave	Public
a975361a-a7aa-11e7-abc4	Concept « Nom de l'auteur »	Flaubert	Visible par X, pas par Y

Description de la méthode suivie

Fichier SKOS

✔ Pérennité : on cristallise le savoir-faire scientifique.
✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

Avec les SKOS concepts, une simple table permet de remplacer tout modèle de données

Identifiant unique persistant	SKOS	Valeur	Niveau visibilité
7a8a34f4-a7aa-11e7-abc4	Concept « Prénom de l'auteur»	Gustave	Public
a975361a-a7aa-11e7-abc4	Concept « Nom de l'auteur »	Flaubert	Visible par X, pas par Y

Les données sensibles sont stockés sous forme cryptée et visibles uniquement pour les utilisateurs habilités:

✓ Protection des données sensibles
✓ Les scientifiques peuvent verrouiller leur travail limitant les droits de réutilisation
✓ Les aspects juridiques sont pris en compte.

Description de la méthode suivie

Fichier SKOS

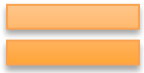
✔ Pérennité : on cristallise le savoir-faire scientifique.
✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

- Avec l'approche SKOS on cristallise dans un fichier XML toute la connaissance «métier» d'un domaine spécifique.
- Ce fichier peut être utilisé pour construire dynamiquement les interfaces de saisie et consultation des données, à partir de briques logicielles génériques.

spécifique

Couche logicielle générique



Interface ad hoc de consultation et saisie

Description de la méthode suivie

Fichier SKOS

✔ Pérennité : on cristallise le savoir-faire scientifique.
✔ Inéquivocabilité : les définitions sont claires et fixées *una tantum*.

✔ Flexibilité : le modèle conceptuel est inscrit dans la modélisation SKOS, pas dans un modèle de base de données

- Avec l'approche SKOS on cristallise dans un fichier XML toute la connaissance «métier» d'un domaine spécifique.
- Ce fichier peut être utilisé pour construire dynamiquement les interfaces de saisie et consultation des données, à partir de briques logicielles génériques.
  - Un seul développement pour nombreux déploiements (disciplines et centres de données)
    - La mutualisation permet d'optimiser les coûts de développement et maintenance.
    - Mutualisation du savoir faire technique
    - L'architecture technique est intégrée dans la solution logicielle générique.

Conclusions et remarques

L'approche décrite permet d'apporter des éléments des réponses aux problématiques

- Méthodologiques
- Ethiques
- Organisationnelles/Techniques

Qui sont couramment rencontrées par les équipes travaillant sur les données de recherche.

Il s'agit d'une méthode basée sur des concepts de web sémantique (SKOS)

- Qui peut être facilement appliquée à plusieurs disciplines des SHS
- Qui permet de limiter les coûts de développement/configuration/exploitation de l'infrastructure technique nécessaire à la publication et diffusion des données.

Le niveau intermédiaire de confidentialité (accès autorisé à personnes habilitées) suppose une instance de décision qui consulte le producteur mais peut s'y substituer si nécessaire. On propose de n'habiliter que des membres de la communauté scientifique tenus au secret professionnel, membres du jury, referees de revues, chercheurs proposant un projet soumis à validation.

Merci à José Sastre, à Jean-Robert Dantou, à Maxime Tissier et à la MSH de Dijon qui ont coopéré à ce travail.