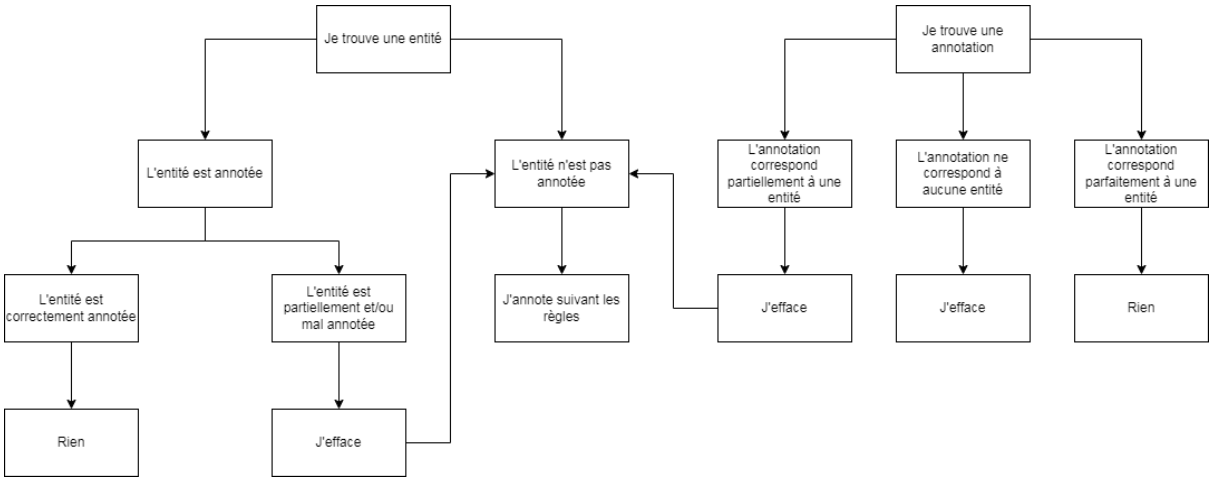
Guide d'annotation CU1
Désidentification et Pseudonymisation

Objectifs

Pour garantir la protection des données personnelles comme établi par la CNIL, il est nécessaire de masquer les données identifiantes dans les documents des établissements de santé (ES) pour pouvoir les traiter. Dans le cadre de PARTAGES, le CU1 prévoit la livraison d'un modèle de pseudonymisation construit à partir d'une version du LLM publiée préalablement. Le modèle sera finetuné et évalué sur les CR fictifs du WP6, puis évalué grâce à la collaboration de plusieurs ES partenaires du projet disposant préalablement d'un jeu de données de pseudonymisation.
Ce document a pour vocation de définir les différentes entités identifiantes que les experts seront amenés à annoter dans les CR fictifs. Le schéma qui en découle a pour but d'adopter une granularité fine au niveau des entités, pour simplifier le processus de conversion vers les schémas pré-établis par les ES avant l'évaluation. Les entités définies ci-dessous sont regroupées par catégories; au sein d'une même catégorie, les entités partagent des caractéristiques similaires pour l'annotation. Le choix s'est porté sur une définition exhaustive plutôt que hiérarchique (entités mères, entités filles) ou granulaire (entités avec attributs) dans un premier temps pour plus de flexibilité dans la conversion vers d'autres schémas d'annotation.

Lignes directrices pour l'annotation

Les attentes envers les annotateurs concernent principalement la qualité de l'annotation. Pour ce qui est de la méthodologie, on peut reprendre le schéma proposé par le CHU de Bordeaux lors du lancement de sa campagne d'annotation pour la pseudonymisation. La branche gauche du schéma correspond aux lignes directrices pour la phase d'annotation, la branche droite à celle de curation.

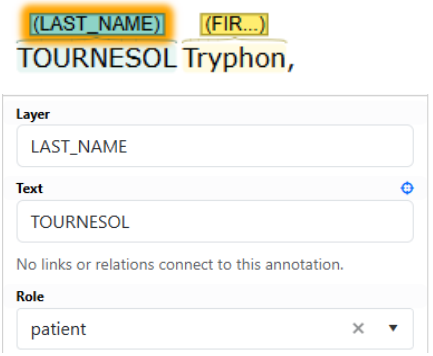


On peut aussi ajouter qu'il est important de bien délimiter les entités annotées, notamment faire ne pas laisser d'espaces ou de ponctuations qui pourraient ensuite bruiter les résultats de l'évaluation.

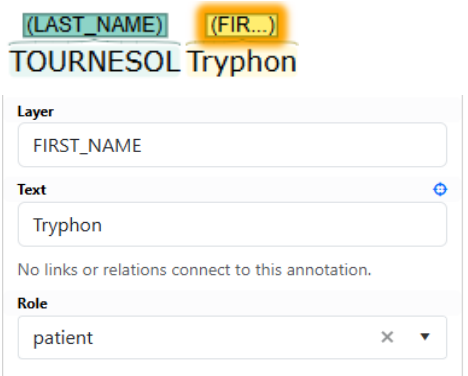
Description des entités

Entités de la personne

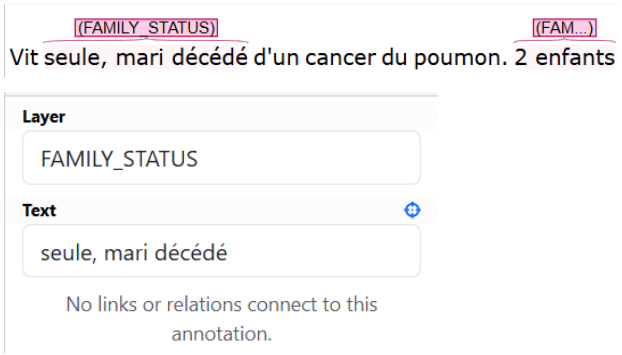
Nom de famille

- Définition : nom de famille ou initiale du nom d'une personne physique, nom de jeune fille.
- Annotation : surligner le nom et lui donner le label LAST_NAME, ainsi qu'un attribut permettant de préciser la nature du nom mentionné (cf. ci-dessous)
- Attribut role : distingue le rôle de la personne dans l'institution hospitalière, i.e. le patient des médecins/infirmiers/etc.
    - Valeurs possibles : [patient, carer, other]
- Exemples
    - Georges DANTON → Georges FIRST_NAME[patient]
    - Georges Jacques DANTON → Georges Jacques FIRST_NAME[patient]
    - 

Prénom

- Définition : prénom(s) ou initiale du prénom d'une personne physique
- Annotation : surligner le prénom et lui donner le label FIRST_NAME, ainsi qu'un attribut permettant de préciser la nature du nom mentionné (cf. ci-dessous)
- Attribut role : distingue le rôle de la personne dans l'institution hospitalière, i.e. le patient des médecins/infirmiers/etc.
    - Valeurs possibles : [patient, carer, other]
- Exemples
    - Georges DANTON → DANTON LAST_NAME[patient]
    - Madeleine Alexandrine GEBELIN BRES → GEBELIN BRES LAST_NAME[carer]
    - 

Statut familial

- Définition : statut familial et/ou marital du patient
- Annotation : surligner la mention du statut et lui attribuer le label FAMILY_STATUS
- Exemple
  - 

Identité sociale du patient

- Définition : métier et/ou statut social du patient (actif, retraité, étudiant, etc.)
- Annotation : surligner le statut et lui attribuer le label PATIENT_SOCIAL_IDENTITY
- Exemple
  - 

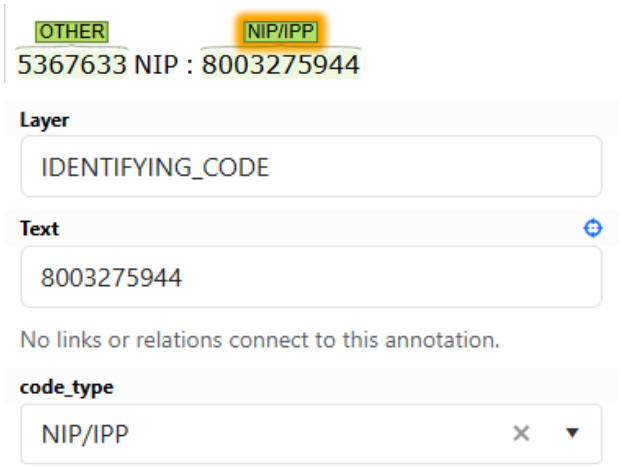
Nationalité du patient

- Définition : nationalité du patient si mentionnée
- Annotation : surligner la mention de la nationalité et lui attribuer le label PATIENT_NATIONALITY

Codes identifiants

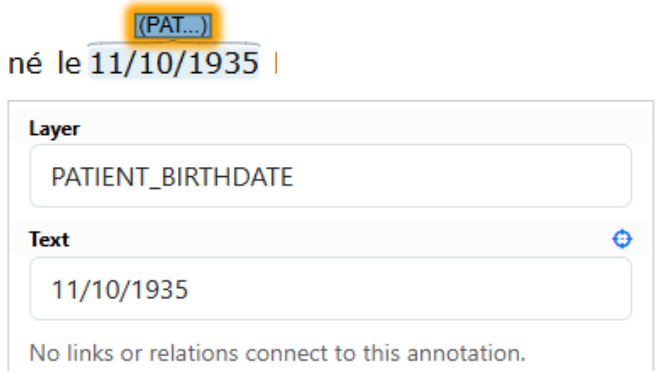
Cette entité a pour objectif de standardiser l'identification et l'étiquetage des numéros ou codes personnels présents dans les CR. Ces codes appartiennent à des systèmes d'identification nationaux ou hospitaliers et nécessitent une reconnaissance précise pour garantir le maintien de la confidentialité des données de santé. Les annotateurs devront repérer et marquer toutes les occurrences des identifiants suivants : le NIR (Numéro d'Inscription au Répertoire), le NSS (Numéro de Sécurité Sociale), le NIA (Numéro d'Inscription en Attente), l'INS (Identité Nationale de Santé), l'IPP (Identifiant Patients Permanents, aussi NIP ou NIPP), l'IEP (Identifiant d'Épisode Patient), ainsi que tout autre code ou numéro identifiable. La redondance de ces identifiants au sein des CR nous étant inconnue,plutôt que de tout aplatir en créant une entité pour chaque code, ils seront regroupés au sein d'une unique entité et un attribut permettra de différencier leur nature.

Code identifiant

- Définition : cf. ci-dessus
- Annotation : surligner la mention d'un code identifiant et lui attribuer le label IDENTIFYING_CODE, ainsi qu'un attribut spécifiant sa typologie (cf. ci-dessous)
- Attribut type : type de code identifiants parmi les codes possibles liés à la pratique médicale et hospitalière française
  - Valeurs possibles : [NIR, NSS, NIA, INS, IEP, IPP/NIP/NIPP, other]
- Exemple
  - 

Dates

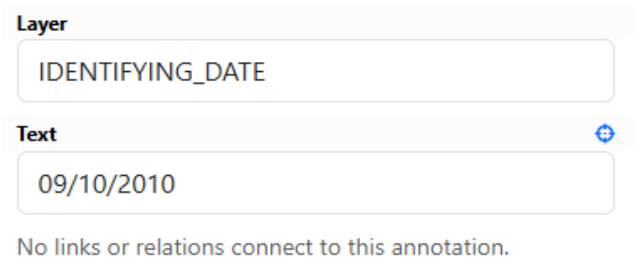
Date de naissance du patient

- Définition : date de naissance du patient (si mentionnée)
- Annotation : surligner la mention de la date de naissance sans les déterminants et lui attribuer le label PATIENT_BIRTHDATE
- Exemple
  - 

Date identifiante

- Définition : date décrivant des interventions médicales ou des actes médicaux pouvant permettre d'identifier un patient, un membre du corps médical (ex: opération, hospitalisation). Il peut s'agir d'une date précise ou d'une périodicité (ex: "Le patient a décrit des douleurs thoraciques depuis mai 2025 [...].")
- Annotation : surligner sa mention sans les déterminants et lui attribuer le label IDENTIFYING_DATE
- Exemples

- Le patient a décrit des douleurs thoraciques depuis mai 2025 → mai 2025 IDENTIFYING_DATE
- Hospitalisé du 09/10/2010 au 11/10/2010



- Patient hospitalisé du 12 au 15 juin → 12 au 15 juin IDENTIFYING_DATE

Date non identifiante

- Définition : date décrivant des évènements non identifiants et/ou sans lien avec le personnel ou le corps médical
- Annotation : surligner sa mention sans les déterminants et lui attribuer le label UNIDENTIFYING_DATE

Remarque: **dates périodiques**

- On peut distinguer 2 type de mentions de dates périodiques: périodes définies par 2 dates complètes (ex: du 09/10/2010 au 11/10/2010) ou par 2 dates partielles n'ayant de sens que l'une avec l'autre (ex: du 12 au 15 juin)
- **Si l'une des deux dates de l'intervalle apporte des précisions sur l'autre** (ex: "du 12 mai au 15 mai 2025" : on sait alors en lisant la deuxième que la première est en 2025) alors il faut annoter les deux dans la même entité DateIdentifiante "du [12 mai au 15 mai 2025]"). Par exemple
  - du [12 au 15 mai]
  - du [12 au 15 mai 2025]
  - de [mai à juillet 2025]
  - entre le [14 mai 2025 et le 18 mai]
- Si aucune des deux n'apporte de précision à l'autre, alors on les laisse séparées:
  - du [12 mai] au [15 juin]
  - du [12 mai 2025] au [15 juin 2025]
  - de [mai] à [juillet]

Remarque 2 : Ne pas annoter les âges (exp: "patiente X, 84 ans" et les dates relatives (exp : "il y a 6 mois").

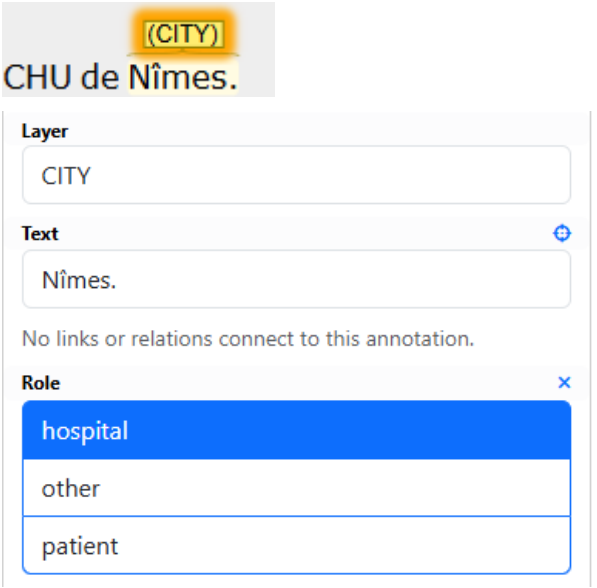
Entités postales

Adresse

- Définition : numéro et nom de rue d'une adresse, ainsi qu'un complément d'adresse tel qu'un lieu-dit, le nom d'une résidence ou un nom d'hôpital.
- Annotation : surligner la mention de numéro et nom de rue et lui attribuer le label ADDRESS, ainsi qu'un attribut précisant la nature de l'adresse (cf. ci-dessous)
- Attribut role : permet de distinguer les adresses des patients des adresses liées aux unités hospitalières

    - Valeurs possibles : [patient, hospital, other]
- Exemples
    - on envoie à Rothschild, 5 rue Santerre, Paris → Rothschild, 5 rue Santerre ADDRESS[hospital]

Ville

- Définition : mention du nom de la ville dans une adresse, et mention d'un arrondissement pour des métropoles
- Annotation : surligner la mention de la ville et lui attribuer le label CITY, ainsi qu'un attribut précisant la nature de l'adresse (cf. ci-dessous). Dans le cas de la mention d'un arrondissement, si ce dernier est contextualisé par le nom d'une ville auquel il est attaché (ex: Paris 15e, 15e arrondissement de Paris) le tout est annoté sous une unique entité. Par contre, si l'arrondissement est mentionnée de manière isolée (ex: Le patient habite à Paris, il a déménagé du 10e au 12e arrondissement en 2020), il doit être annoté avec sa propre entité CITY.
- Attribut role : permet de distinguer les adresses des patients des adresses liées aux unités hospitalières
    - Valeurs possibles : [patient, hospital, other]
- Exemples
    - on envoie à Rothschild, 5 rue Santerre, Paris → Paris CITY[hospital]
    - on envoie à Rothschild, 5 rue Santerre, Paris 12e → Paris 12e CITY[hospital]
    - on envoie à Rothschild, 5 rue Santerre, dans le 12e → 12e CITY[hospital]
    - on envoie à Rothschild, 5 rue Santerre, dans le 12e arrondissement → 12e arrondissement CITY[hospital]
    - 

Code postale

- Définition : mention du code postal dans une adresse
- Annotation : surligner la mention du code postal et lui attribuer le label ZIPCODE, ainsi qu'un attribut précisant la nature de l'adresse (cf. ci-dessous)
- Attribut role : permet de distinguer les adresses des patients des adresses liées aux unités hospitalières
    - Valeurs possibles : [patient, hospital, other]
- Exemple

- on envoie à Rothschild, 5 rue Santerre, 75012 → 75012 ZIPCODE[hospital]

Pays

- Définition : mention d'un pays dans une adresse
- Annotation : surligner la mention du pays et lui attribuer le label COUNTRY, ainsi qu'un attribut précisant la nature de l'adresse (cf. ci-dessous)
- Attribut role : permet de distinguer les adresses des patients des adresses liées aux unités hospitalières
  - Valeurs possibles : [patient, hospital, other]
- Exemple
  - on envoie à Rothschild, 5 rue Santerre, 75012 FRANCE → FRANCE COUNTRY

Entités numériques

Adresse email

- Définition : adresse email d'un patient, d'un membre d'un service hospitalier ou d'une unité hospitalière
- Annotation : surligner l'adresse email et lui attribuer le label EMAIL_ADDRESS, ainsi qu'un attribut précisant sa nature (cf. ci-dessous)
- Attribut role : permet de distinguer les adresses emails des patients de celles liées aux unités ou personnels hospitaliers
  - Valeurs possibles : [patient, carer, hospital, other]

Numéro de téléphone ou fax

- Définition : numéro de téléphone ou de fax d'un patient, d'un membre d'un service hospitalier ou d'une unité hospitalière
- Annotation : surligner le numéro de téléphone et lui attribuer le label PHONE_NUMBER, ainsi qu'un attribut précisant sa nature (cf. ci-dessous)
- Attribut role : permet de distinguer les adresses emails des patients de celles liées aux unités ou personnels hospitaliers
  - Valeurs possibles : [patient, carer, hospital, other]

URL

- Définition : lien vers un site internet, des ressources, etc.
- Annotation : surligner la mention du lien et lui attribuer le label URL