16|17 NOVEMBRE 2016

CITÉ DES SCIENCES ET DE L'INDUSTRIE - PARIS

Les rencontres du NUMÉRIQUE de l'ANR

ANR 10 ANS

Les rencontres du NUMÉRIQUE de l'ANR

ACCORDYS
et les données de fœtopathologie

Jean Charlet[1,2] & Ferdinand Dhombres[1,3]
[1]INSERM U1142/LIMICS
[2]DRCD, Assistance Publique – Hôpitaux de Paris
[3]Hôpital Trousseau, Assistance Publique – Hôpitaux de Paris

16|17 NOVEMBRE 2016

Objectifs, moyens

- **Exploitation des données de fœtopathologie**
  - Textes brut et images
  - Pour proposer des orientations diagnostiques à des médecins en leur présentant des situations antérieures et similaires de malformations fœtales

- **Création d'une base de cas de fœtus porteurs de malformations**
  - A partir des dossiers non numériques (papier et diapos) sur une période de 23 ans
  - Permettant exploitation à travers une modélisation sémantique et des algorithmes de raisonnement à partir de cas

Les rencontres du NUMÉRIQUE de l'ANR

Acquisition des données

- **Source : service de fœtopathologie de Trousseau (CPDPN)**
  - Sélection parmi 7000 dossiers archivés des cas de malformations → 2476 dossiers
  - Sélection d'un échantillon des diapositives (<20) pour chacun de ces dossiers

- **Protocole**
  - Calibration des scanners sur échantillons
  - Numérisation des textes et des diapositives
  - Reconnaissance de caractères (OCR), grande disparité, fonction des années et des papiers
  - Anonymisation par (Medina)
  - Dédoublonnage
  - Correction orthographique

Démarches réglementaires

- **CPP**
- **CNIL** (DAJ de l'AP-HP)
  - Demande d'autorisation
  - Finalement, déclaration normale (Janvier 2015)

- Données sensibles (autopsies fœtales, consentement rétrospectif éthiquement non recevable)
- Arrêt du projet durant **12 mois** pour obtenir l'autorisation CNIL grâce à une révision du protocole de traitement avec la DAJ :
  - Numérisation et OCR sur site à Trousseau
  - Supervision par les médecins du CPDPN
  - Double validation exhaustive de l'anonymisation →165 000 pages
  - Dossiers anonymisés exploitables dans les locaux des partenaires.
- Serveur final accessible uniquement au sein du CPDNP

Difficultés liées aux données

- **Délais d'obtention**
  - Travail initial des membres du projet sur un **nano-échantillon**
  - Données réelles disponibles, maintenant en totalité, mais après le départ d'une partie des chercheurs non permanents

- **Modification des contrats/processus**

  Pour réintégration des OCR sur site, double validation de l'anonymisation, dédoublonnage des CRs, correction orthographique, mise en œuvre d'algorithmes de RàPC pour documents – arbres – « à trous »

Les rencontres du NUMÉRIQUE de l'ANR

Conclusion : si on avait quelque chose à changer

- Préparer la demande d'accord CNIL pour envoi le jour de l'acceptation du projet (une CNIL plus rapide ?)
- Préparer un échantillon de dossiers originellement numériques pour s'affranchir du traitement scan/OCR
- Pas d'alternative sur le matériel (dossiers à « sauvegarder » dans le cadre du programme Contint)
- Le projet se poursuit et on raccorde les différentes démarches/modules développés indépendamment