GdR MADICS
Journées Sciences des Données
22 – 23 Juin 2017

Atelier QUALIMADOS :
Qualité des Masses de Données Scientifiques

Organisateurs :
Allel HADJALI (LIAS/ENSMA, Poitiers)
Laure BERI-EQUILLE (IRD, Montpellier)
Angela BONIFATI (LIRIS.CNRS, Lyon)

Qualité des Données

- **Problème majeur**
  - 10 à 30% des données sont estimées impures/imparfaites
  - Décisions et analyses non fiables
- **Problème multidimensionnel**
  - Multiples dimensions : Données dupliquées/redondantes, Données imparfaites, Valeurs manquantes, Données obsolètes, …
  - Indicateurs de qualité : Sémantique claire, Lien fort avec la réparation
- **Réparation / nettoyage**
  - Vision holistique (interactions entre les anomalies)
  - Réparation post-traitement (anomalies apparaissant après traitement)
  - Réparation semi-automatique (mettre l'expert au cœur du processus)

Données Massives et Scientifiques

- **Ère du Big data**
  - Qualité des données loin d'être parfaite
    - ✓ Due au volume des données générées, vitesse d'arrivée de ces données, large variété d 'hétérogénéité des données
  - Méthodes de réparation performante en termes de scalabilité et de temps
- **Données Scientifiques**
  - Issues de capteurs, de processus de simulation, d'observations par satellites, …
  - Forte présence d'erreurs, d'incertitude/bruit, de valeurs manquantes, …

La qualité des masses de données scientifiques constitue un réel challenge

Atelier QUALIMADOS

- **Genèse**
  - Suite aux discussions et échanges avec les responsables du GdR MADICS
    - ✓ Séminaires de restitution des projets Mastodons (Février 2017, Paris)
    - ✓ Importance d'une action autour de la qualité des données
  - Un atelier d'abord sur la qualité des données
    - ✓ Faire le point sur l'état des méthodes/approches/applications liées à la qualité
    - ✓ Lancer une discussion sur le projet de création de l'action

Programme

- **10h15 – 10h30 : Présentation de l'atelier et de ses objectifs** (Allel Hadjali, LIAS/ENSMA, Poitiers)
- **10h30 – 11h00 : Data Quality: where are we on the journey from theory to practice?** (Angela Bonifati, LIRIS, Lyon)
- **11h – 11h30 : Tour d'horizon des données scientifiques et des problématiques particulières liées à leur qualité** (Laure Berti-Equille, IRD, Montpellier)
- **11h30 – 12h15 : Gestion des annotations sémantiques en santé - Le projet ELISA** (Cédric Pruski, ITIS, Luxembourg)
- **12h15 – 14h00 : Déjeuner**
- **14h – 14h45 : Prise en compte des données manquantes dans les modèles de mélanges : Application aux séries temporelles d'images multispectrales** (Serge Iovleff, Université de Lille)
- **14h45 – 15h30 : Qualité dans l'entrepôt de données cliniques de l'HEGP** (Bastien Rance, HEGP, Paris)
- **15h30 – 16h : Discussions et clôture de l'atelier**

Merci au GdR Madics pour son Soutien

MERCI DE VOTRE PRESENCE

Enjoy the workshop