Workshop Open & Big Data 30/03/2015
Organisation FLI-IAM : **C. Barillot & M. Dojat**
Participants (52) : en majorité chercheurs, institutionnels et quelques industriels.

Les présentations et la liste des participants est disponible sur le site FLI-IAM

10:00-10.45 - Welcome and Introduction of the workshop - M. Dojat & C. Barillot / (FLI)

- **M. Dojat** (Grenoble)
  - M. Dojat rappelle l'intérêt du partage de données et d'outils autour de trois axes : scientifique, notamment pour dépasser la réalisation actuelle d'études sur des populations trop réduites qui génèrent des résultats statistiquement valides, économique, trouver un modèle permettant la mise à disposition de la communauté de plateformes de partage, et éthique pour respecter les droits à la confidentialité des individus mais aussi de ne pas freiner les travaux français face à la concurrence mondiale.
- **C. Barillot** (Rennes)
  - C. Barillot fait une synthèse de l'action FLI-IAM en mettant en exergue les 3 scénarios à l'étude pour un démonstrateur sur base des dialogues réalisés avec différentes communautés (radiologues, cliniciens, RMNistes)

10H45-11H20 Legal and ethics questions: What can be shared? To whom data own? Status of data, metadata and derived data? What are the requirements for sharing?

- **Paul Oliver Gibert** (Digital Ethics)
- PO Gibert a détaillé la position de digitalEthics qui se donne pour but d'aider à la gestion des problèmes posés par le numérique en terme d'éthique. Le problème de la confidentialité des données et du droit du sujet à l'accès à ses données est abordé via les articles de loi en cours. Du fait de la gestion des données en larges réseaux la traçabilité est difficile. Les aspects abordés concernaient notamment:
  - Digital Ethics & IDV
    - équipe de 5 personnes pluridisciplinaire
    - de l'éthique jusqu'au code (audit d'algorithme, enjeux éthiques d'un projet
    - recherche:

        - Projet IDV (projet de Sorbonne Paris Cité pour un atlas du vivant)
        - open data and sharing, data driven innovation
    - Legal framework today and tomorrow
        - loi 2002 : « *The patient at the heart of his health system* »
        - loi de 2011 sur la recherche biomédicale (consentement du patient et objectif de la recherche)
        - droit d'accès aux données et de revision de leurs usages
        - 3 key aspects: Patient's consent, Patient data must be stored securely, authorization of the CNIL
        - Art 47 of health law (2015): protection of personal data
        - evolution of the information systems: toward linked open data
        - Who governs the data?
          production is complex (sensors, human activity, processing, enhancement), different contributors can claim the value of the data
        - different retributions: money is not the best solution, reuse, integration is more a solution.
        - This is possible at the individual level, less at the institutional level
        - The legal model does not exist, there is some similarities with the "money" legal status
        - incoming points: the status of data and meta-data
- Discussion
    - Mark ASCH (MENESR) : link to the Research Data Alliance in order to share our expertise and knowledge with other domains in health (https://rd-alliance.org)

11h20-12h00 Institutional aspects: What are the institutional actions to promote data sharing?

- **Christine Balagué** (VP CNNum, FR)
- C. Balagué a exposé les travaux du Conseil du numérique. Importance de la confidentialité et du droit de retrouver pour le patient ses données et de les partager à sa guise (voir project Fing en France ou the US Blue Button). Importance de gérer au cas pas cas (voir partage de données personnelles pour le suivi de la progression d'une épidémie type EBOLA). Importance d'une uniformisation européenne qui tient compte des spécificités historiques face à l'ouverture large nord-américaine. Modèle économique à trouver face à Google, Apple etc ... Nécessité de s'adapter à l'évolution rapide dans le champ du numérique. Importance de la formation des acteurs de santé sur le numérique et le partage de données. Les aspects abordés concernaient entre autre:
    - How institutions can promote Data sharing?
    - CNNum is an independent advisory commission launched by the French Ministry of Digital contents
    - digital consumer empowerment: from individual based sources to net based sources (Demand-> Information-> Network-> Crowd)
    - Toward massive information sharing era (VVV : Volume, Veracity, Variety)
    - New paradigm: Our lives are in data
    - Big data challenges: Collection, Aggregation, Analysis
    - Data challenge in health sector
        - huge generation of data
        - value of data is very high
        - data regulation : 2 principle:

        - protection of fundamental rights of individuals (actual is passive consent but gives very few power to the individuals as for dominant e-platforms)
        - digital at the service of public goods
      - "autodetermination informationnelle (Conseil d'Etat 2014" (individual right to decide of communication and use of personal data): how this will extend to EUROPEAN level?
      - in US : US Blue Button (PIMS: Personal Information Management Systems)
    - Data Sharing supposes Interoperability and portability standards, Anonymization, security of environment, Authorization for access, Explicit Consent
    - Allow public actors to ask for a constitution of "Public Goods"
    - Notion of "Commons": digital commons are source of economic and social innovation, a core concept in Open Data
    - Challenge: educate and train the health sector (people, institutions) to digital challenges
      - relation patent-doctor, peer-to-peer communities, telemedecine, qualified self, robotization, decision making, new Internet actors (Google, Apple, Amazon, …), augmented human being, Internet of things, transhumanism
    - connexions to Ethics committees of Allistene (CERNA) and TheraLab, a PIA on Big Data (with some aspects on heath domain)
- Discussion

Déjeuner

14h-15h00 Technical aspects: What are the existing solutions? What are the main difficulties?

- **P. Mouillard** (Vigisys, FR)
- P. Mouillard a exposé la vision de Vigylis qui accompagne des startups dans l'utilisation du numérique en santé, les aspects abordés concernaient notamment:
    - what is big data? lots of data, complex data, dataflows and time series.
    - big data does not mean qualitative data for data mining
    - Medical Imaging is not just diagnosis and decision making, but also prevention & screening, training & planning, real time imaging, atlas and follow-up of pathologies or treatment assessment
    - Big Data in medical imaging is very different to e-commerce: fewer instances but more data *per capita*
    - Sharing medical images: medical practice become collective, multi-modality, tele-medicine & tele-diagnosis, nosology and semiotics needs to be refined, research more and more focused on specific diseases, sharing and testing image processing algorithms
    - Static Big Data (retrospective and prospective studies with statistical analysis) vs Dynamic Big Data ("forever ongoing studies, auto-adaptive and machine learning systems)
    - Imaging data: images are not just pixels, Big data needs image processing. Metadata is key for relevant retrieval and selection.
    - Clinical data is always needed for use of image data
    - technological trends: data encapsulation, distributed storage, Virtualization (Processing and storage), Front/back API (Front apps for pre-processing, anonymisation), Security (including data striping for data storage), SaaS with Web App

    - Imaging data archtecture: non vendors PACS, Open Source visualization software, access the PACS from the patient's bed, link to in-house Medical Record System, connexion to research and processing add-ons
    - Difficulties: pooling distributed images, computation cost, who wants to share? extraction of relevant features (biomarkers), data format (image, clinical), open systems, heterogeneity of data, Ethics, Security, Who pays for this?
    - Perspectives: new imaging sources, connected objects, dynamic imaging and time series, robotic surgery, multimodal in vivo imaging, PACS app for iPhone
- **D. Kennedy** (INCF, NITRC, USA)
- David Kennedy a rappelé la nécessité scientifique à partager les données notamment autour de la reproductibilité des résultats publiés, la position de l'INCF et les efforts faits à l'échelle internationale pour le partage de données et de résultats de Neuroimagerie. En parallèle des efforts qui doivent être faits à par les institutions sponsorisant les études et les journaux pour inciter au partage, il faut faciliter le partage en créant des outils adaptés qui s'ajoutent aux outils existants. Des efforts sont faits par l'INCF pour intégrer de tels outils dans SPM ou FSL et assurer la provenance des données produites. Les aspects abordés concernaient notamment:
    - Why to Share?
        - Key aspect for scientific reproducibility: irreproducibility costs money, costs trust and at the end costs lives
        - ConceptFAIR Data (hhtp://www.datafairport.org): Findable, Accessible Interoperable and Re-usable: These principes are critical to the protection of public investments
    - Framework for Sharing
        - Stages for information sharing: Data (data repositories)), Analysis (standard workflow description), Results (meta-data repositories), Interpretation (Literature repositories
    - Existing solutions
        - Image raw data level (ADNI with IDA, NDAR with NIMH, HCP with XNABT), NITRC (XNAT), COINS, LORIS, OpenfMRI
        - Derived Data: Statistical maps (NeuroVault), Activation Foci (BrainMap, SuMSDB, ...)
        - NITRC: Neuroinformatics Tools and Resources Clearinghouse (http://nitrc.org) NITRC-CE is available on the amazon web services marketplace (AWS Marketplace)
        - NIDM: INCF Neuroimagig Data Model (Keator et al. 2013): apps for SHARING and querying NI-DM repositories
    - Data Sharing Barriers
        - Incentive barriers:
            - funders and publishers beginning to require sharing (e.g. PubMed central "success story" for open publication)
            - Education & simplification needed
        - Costs Barriers
            - sharing costs money, who pays for this?
            - what is the magnitude of the data persistance insurance (5-10% of the cost of acquisition)
        - Credit Barriers: Nested DOIs (e.g. new publication using shared data cite the original DataElement DOI in the new Dataset DOI a part of the new publication DOI)

        - Publication Barriers
- Discussion

Pause

15h30-16h30 Pros and cons of working architectures: positive and negative aspects of sharing to large scale in practice? user acceptance.

- **Wiro Niessen** (Rotterdam cohorte, NL)
- Wiro Niessen a exposé le projet de Rotterdam autour de la large cohorte sur le vieillissement normal. Il souligne la difficulté de supporter financièrement les challenges permettant de valider des chaînes de traitement. Les aspects abordés concernaient notamment les enjeux et solution pour le partage de données, à savoir:
    - standardization and workflows
        - lack of standardization, lack of automation, state-of-the-art algorithms not widely available, lack of automatic QA, lack of tools for data sharing
    - performance s, challenges and "open medical computing": toward OPEN MEDICAL IMAGE 2.0 for sharing data AND processing tools
    - IT infrastructures enabling linking imaging to other data
    - Sharing data: depends to the perimeter of sharing (Who and how long?)
    - imaging genetics and radiomics
        - derive image genetics to biobank
    - Rotterdam study:
        - >12k MRI acquired with tissue quantification, lesion assessment, segmentation & shape, microstructure & function, incidental brain finding
        - XNAT for Dicom images, processed images and annotations, Image processing unit for processing and OpenClinical for clinical data

- **Gabriel Robert + Dimitri Papadopoulos** (Imagen Cohort, FR, UK)
- G. Robert et D. Papadoupoulos ont exposé le projet IMAGEN et notamment les points difficiles sur le long terme pour assurer la gestion et la collection des données en imagerie et génétique avec des acteurs qui nécessairement ont évolués au cours du temps. Les aspects abordés ont concernés:
    - Phenotypic variability + small gene effects + gene environment interactions = power issues
    - 8 acquisition centers.
    - Initial data management plan: centralized in CEA-Neurospin using XNAT + extensions with image processing
    - lacks of persistance of the infrastructure (move from XNAT to CubicWeb Brainomics data management solution):
        - several subtle changes made an overall major change (subject identifier, new people in acquisition centers, unmaintained equipments for acquisition, lost access to software for data collection)
    - lesson learned:
        - use portal for data collection and automatic error detection early
        - use standard software
        - do not underestimate difficulties in data collection ad curation
        - data dissemination and data collection are linked
        - data may change over time
    - CubicWeb: query langages using SPARQL/RDF and RQL extension: faster query than XNAT with local extension

- Imagen is big data but not open open, Localizer fMRI project is an opendata extension (http://brainomics.cea.fr/localizer)

- 16h30 – 17h Closing discussion:
- Les aspects abordés concernaient:
  - Prior Issues
    - Standards for sharing data
      - Do we need standard for data model (e.g. ontologies) ?
      - Standard for raw/derived data format ?
      - Do we need standard for interoperability (on concepts, on communication) ?
      - Do we wait for standards or do we go forward ?
    - Quality control in Data Sharing
      - Ratio cost / added value
      - Automatic vs Human Quality Control
  - Infrastructure Issues:
    - Data bases
      - Big Data Centers vs Distributed Storage vs Peer to Peer storage (bioTorrent)
    - High Performance Computing (HPC) and Big Data for *in vivo* imaging
      - Is HPC infrastructure generic or specific to usages?
      - Cloud computing: Clusters vs Grids vs Crowd Computing ?
    - Image Processing Code Sharing ?
      - Is it different than Data Sharing?
      - Is Crowd Science meaningful for in vivo imaging?
  - Socio-economic issues:
    - Data Sharing credit ?
      - Co-authorship *vs* citation *vs* payment ?
      - Co-production of derived data production
      - Data rights
    - Economic model
      - Cost for using shared data ?
      - Who support the cost ?
      - Can data sharing become a business?
    - Is regulation / ethics a real limit?
    - How funding agencies/institutions can help to promote data sharing?