

# Définition informatique de la donnée numérique et perception des contraintes de propriété intellectuelle

Tristan Allard, IRISA / univ. Rennes 1  
Pascale Sébillot, IRISA / INSA Rennes

Institut de Recherche en Informatique et Systèmes Aléatoires

# Donnée numérique : définition

- **Tout est donnée (?)**
  - Quelque chose de **fourni** est donnée
  - Quelque chose de **produit** à partir de cela devient donnée
- **En fait, des acceptions différentes au sein de l'informatique, avec un certain flou dans la terminologie**
- **Définition de la donnée par opposition à information et savoir (connaissance)**
  - Donnée : fait, valeur brute
  - Information : donnée + sémantique
  - Savoir : information + confiance/pertinence/expérience
  - Déclinaison en Science de l'information (pyramide DIKW), en *Business Intelligence* (cercle vertueux de la donnée)...

# Donnée numérique : définition

- **Définition par opposition entre données structurées et données non structurées**

pièce	largeur	poids
écrou	12	50
...		

tableurs, bases de données,  
*Open Data*, data.gouv.fr...

Derniers  
développements  
dans l'affaire...



des suites de caractères,  
des successions de pixels...

- **Organisation de l'exposé**

- Caractérisation et portée de la donnée personnelle (T. Allard)
- Propriété intellectuelle et contenus multimédias (P. Sébillot)

# Donnée personnelle... Quésaco ?

- **Point de vue juridique (article 4 (1) du RGPD) : définition très large d'une donnée personnelle (cf. ci-dessous)**  
« toute *information* se rapportant à une personne physique identifiée ou *identifiable*, (...) notamment par référence à **[LISTE]** (...) »
- **Point de vue 'base de données'**
  - « *information* » : (ici) des octets associés à un « contexte »  
*Par ex. : (Age (nombre entier), 0110 1011)*
  - « *identifiable* » : potentiel(s) croisement(s) avec d'autre(s) fichier(s) (cf. transparents suivants)
  - « *notamment par référence à [LISTE]* »

« Nom, numéro d'identification, localisation, identifiant en ligne, éléments de son identité physique, physiologique, génétique, psychique, économique, culturelle, sociale »

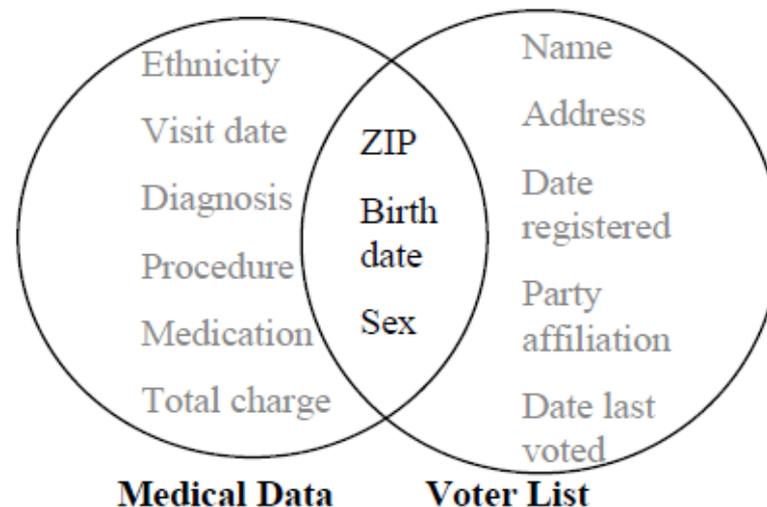
**RGPD, article 4(1)**



# Le processus de ré-identification par l'exemple

## ■ Cas du gouverneur Weld [Sweeney02]

- Données médicales collectées par le GIC (responsable achat assurance santé pour employés de l'état du Massachusetts)
    - Pas d'identifiant direct
    - Données transmises à des chercheurs, vendues à des industriels
  - Liste de votants du Massachusetts, accessible publiquement
- Croisement : 6 avaient la même date de naissance que le gouv. Weld, 3 d'entre eux étaient des hommes, dont 1 de même code postal



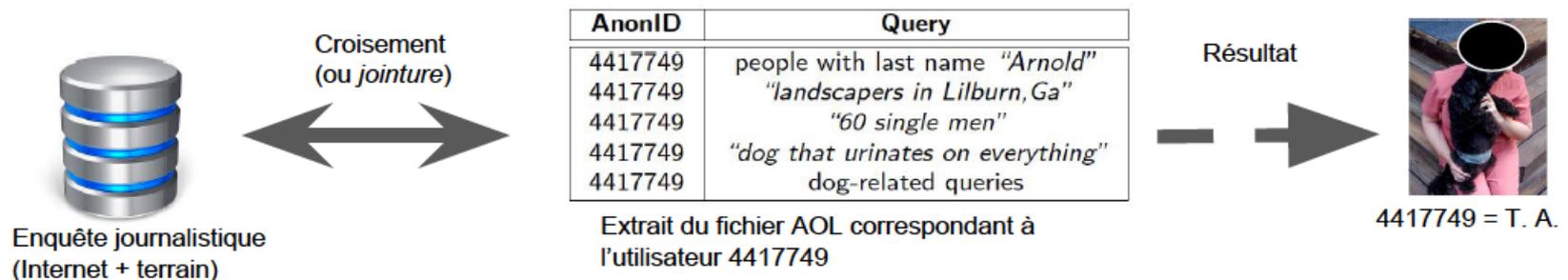
# Confirmation par diverses études

- **Capacité discriminante de divers types d'information**
  - [Golle06] : **63% de la population US** possède une combinaison unique de {genre, code postal à 5 digits, date de naissance complète}
  - [deMontjoy13] : **4** couples aléatoires {date et heure, localisation de l'antenne la plus proche} **identifient 95% des individus** de leur jeu de données.
  - [deMontjoy15] : résultats similaires sur des données d'achat dans des boutiques {jour d'achat, boutique}
  - *Etc. (Netflix [Naryanan09]).*
- **Corrélation entre données**
  - Logs d'appel et prédiction de traits de personnalité (par ex. 'extraversion', 'ouverture') [deMontjoy13]
  - *Etc.*

# Questionnement PI et données personnelles #1

## ■ Accès et usage des logs AOL

- En 2006
  - Publications par AOL des mots-clés des recherches Internet de 658 000 utilisateurs (sans identifiant direct, « *démo* » fichier AOL)
  - Ré-identification de l'utilisateur 4417749 par des journalistes du NY Times



- Aujourd'hui
  - Données toujours accessibles sur Internet, mais pas sur le site d'AOL
  - Droit de les stocker ? Les interroger ? Les republier ? Publier des statistiques ?

## Questionnement PI et données personnelles #2

- **Données Facebook accessibles publiquement**
  - Parcours du graphe Facebook et collecte des informations accessibles par tous
  - « *Démo* » *fichier FB*
  - Droit de les stocker ? Les interroger ? Les publier ? Publier des statistiques (par ex., nombre d'amis par personne) ?  
Quid d'une version sans les noms des profils ?

# Propriété intellectuelle et contenus multimédias

- **Travail sur des contenus / données multimédias** → **des questionnements à chaque étape du processus**
- **Illustration via un projet : LIMAH**   
*(Linking Media in Acceptable Hypergraphs)*



# LIMAH (LInking Media in Acceptable Hypergraphs)



## ■ Objectifs

- Faciliter l'exploration de vastes collections de documents multimédias
- Structurer automatiquement la collection grâce à des liens créés entre documents, fondés sur leur proximité sémantique (même thématique, conséquence de, réaction à, réponse à...)
- Étudier en quoi cette structure modifie les usages et est acceptable

## ■ Partenaires



# LIMAH – Le corpus



- **Corpus d'actualités, multimédia et multi-sources**
  - **Pages web d'actualités** (~5000 documents)
  - **Podcasts d'émissions de radio** (~1500 documents)
  - **Vidéos** (~300 documents)
  - **Réseaux sociaux et commentaires** (~16000 tweets)
- **Corpus récupéré via des flux RSS entre le 20 mai 2015 et le 8 juin 2015**
- **Questions soulevées**
  - Droit à l'aspiration de données sur le web ?
  - Statut du corpus (appartenance, droit à diffusion...)

# LIMAH – Les transformations du corpus



## ■ **Prétraitements du corpus**

- Transcription automatique de la parole présente dans les podcasts radio et les vidéos
- Découpage en phrases et mots ; ajout d'étiquettes de catégories grammaticales aux mots...

## ■ **Représentation des documents pour comparaison**

- Représentation = index formés des mots les plus fréquents ou discriminants, des entités nommées...
- Mots ramenés à une forme de base (racinisation)

## ■ **Questions soulevées**

- Statut de ces nouvelles données (corpus enrichi, transformé ; index)
- (Méta)données « identifiantes » ou pas

# LIMAH – L’hypergraphe et l’interface de visualisation



Hyper Media Analytics

Jun 3, 2015



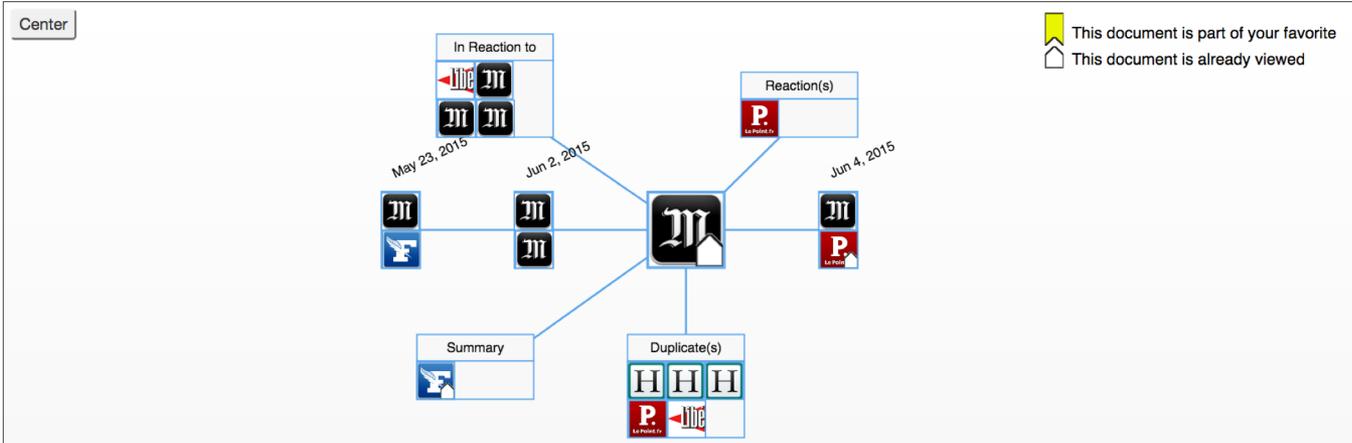
**ABOUT THIS ARTICLE**  
 Type: PRESSE  
 Publisher: Le Monde  
 Category: Article  
 Link: Le Monde.fr  
 Published on: Jun 3, 2015  
 Author: Jean-Michel Bezat  
 Related documents: 18

**WORD TAG**

48	
17	Areva
11	EDF
8	euros
8	milliards
6	activités
6	réacteurs
5	NP
5	filière
3	maintenance

**NAMED ENTITIES**

- Areva
- Areva NP
- CEA Industrie
- CGN
- EDF



**FILTER GRAPH**

- PRESSE
  - Le Figaro
  - Huffington Post
  - Le Point
  - Liberation
  - Le Monde
- VIDEO
  - France Television
- RADIO
  - RMC
  - Radio France
- TWITTER
  - Twitter

## L’Etat se prononce pour la cession des réacteurs nucléaires d’Areva à EDF

EDF « a vocation à devenir actionnaire majoritaire » de l'activité réacteurs d'Areva, et l'Etat « recapitalisera » l'ex-fleuron du nucléaire, a annoncé l'Elysée.

Le gouvernement a pris une première décision, mercredi 3 juin, sur l’avenir de la filière nucléaire. Au cours d’une réunion à l’Elysée, François Hollande, Manuel Valls et les quatre ministres concernés par le dossier – Emmanuel Macron (économie), Ségolène Royal (énergie), Michel Sapin (finances) et Laurent Fabius (affaires étrangères) –, ont donné leur feu vert au projet de rachat par EDF d’une part majoritaire de l’activité réacteurs nucléaires d’Areva, indique *Le Figaro* sur son site Internet. Un arbitrage qui a été confirmé peu après dans un communiqué publié par la présidence de la République.

Dans un premier temps, indique l’Elysée, « les activités de conception, gestion de projets et commercialisation des réacteurs neufs d’EDF et d’ Areva seront rapprochées dans une société dédiée ». Elles concernent 1 200 personnes (sur 44 000 salariés). Ce rapprochement « permettra une politique d’exportation ambitieuse » et « le renouvellement futur du parc nucléaire français », qui doit faire l’objet d’un « grand carénage » dont le coût est estimé à 55 milliards d’euros d’ici à 2025.

**RELATED ARTICLES**

**DUPLICATE(S)**

Rapprochement avec EDF, emplois... l'avenir d'Areva se joue ce mercredi

SUR LES BLOGS  
Investir dans le nucléaire: et à la fin, c'est le contribuable qui paye

L'Etat veut sauver Areva quoi qu'il en coûte

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01

# LIMAH – L'interface



Hyper Media Analytics

Search Add to favorites My favorites Toggle view Account

**Jun 3, 2015**

**ABOUT THIS ARTICLE**  
 Type: PRESSE  
 Publisher: Le Monde  
 Category: Article  
 Link: Le Monde.fr  
 Published on: Jun 3, 2015  
 Author: Jean-Michel Bezat  
 Related documents: 18

Center

This document is part of your favorite  
 This document is already viewed

**FILTER GRAPH**

- PRESSE
  - Le Figaro
  - Huffington Post
  - Le Point
  - Liberation
  - Le Monde
- VIDEO
  - France Television
- RADIO
  - RMC
  - Radio France
- TWITTER
  - Twitter

**RELATED ARTICLES**  
 DUPLICATE(S)

## ■ Questions soulevées

- Statut du lien (hyperlien), liberté de lier
- Rôle des acteurs
- Statut de « l'œuvre » obtenue
- Biais dû à l'affichage (perception particulière « forcée »)

## Remarques conclusives

- **Beaucoup de problèmes de propriété intellectuelle et/ou liés à la vie privée**
  - Sujets d'études informatiques
  - Aspects sous-jacents à des études informatiques
- **Spécificités de l'« objet numérique » vis-à-vis de la propriété intellectuelle**
  - Facilité à le dupliquer / transférer / modifier
  - Difficulté à le supprimer