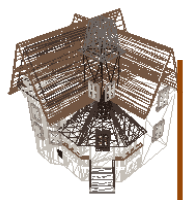
Partage de données

Jean-Gabriel Ganascia
Équipe ACASA - LIP6 – CNRS (UMR 7606)
Université Pierre et Marie Curie – Sorbonne Universités – Labex OBVIL
Institut Universitaire de France
Président COMETS – Membre CERNA
4, place Jussieu, 75252 Paris Cedex 05, FRANCE
Jean-Gabriel.Ganascia@lip6.fr

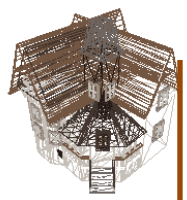


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS



Synoptique

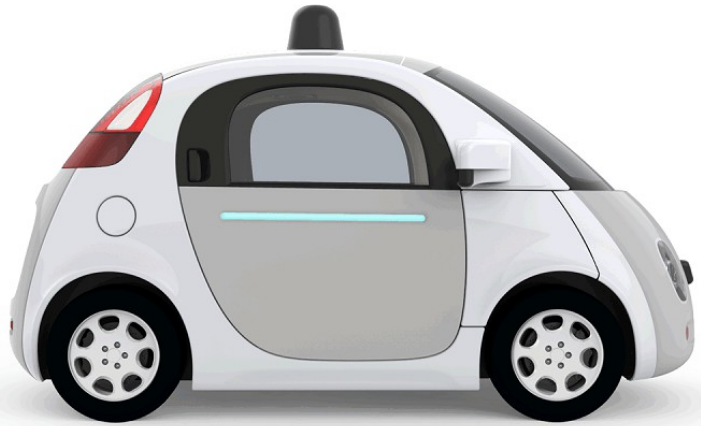
1. **Nécessité des données et du partage**
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS

Nécessité et importance des données

- Robots (et bots)
- Véhicules autonomes
- Biométrie
- Vision
- Reconnaissance de la parole
- Traitement et Compréhension du langage naturel
- e-scie
- …



925M+ People on Groups each month

900M People on WhatsApp each month

1.55B People on Facebook each month

800M People on Messenger each month

400M People on Instagram each month

La masse des masses de données

Quantité d'information

- *1 livre, 1 million de caractères*
  → 1000 Ko = 1Mo (106 octets)
- *Catalogue des livres et imprimés de la BNF:*
  *14 millions d'ouvrages = **14To***
  1 Go = 103Mo = 109o
  1 To = 1012o (1000 milliards)
- ***Volume total du web en 2015**:*
  7 Zeta-octets = 7 1021o

½ milliard de BNF!

Production des données en nombre de BNF

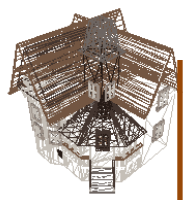
- Twitter produit 7 To/jour c'est-à-dire ½ BNF!
- Échange photos sur Flickr: 2milliards/jour = 500 BNF
- Radio télescope *Murchison Widefield Array* (Australie):
  - **données brutes** 7000

Quantité de données - performances

- **DeepFace**: 4.4 millions d'images représentant 1000 images pour 4030 personnes.
- **Facenet**: 200 millions d'images et 8 millions d'identités uniques

- FaceFirst,
- Face-Six,
- DeepFace (Facebook) 97,25%,
- FaceNet (Google) 99,63%!

Exemples d'images pour 6 identités



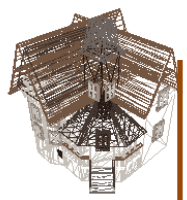
La logique des masses de données

Trois caractéristiques:

- **Les 3V**
  - Volume
  - Variété
  - Vélocité
- **Acquisition continue**
  - « Crowdsourcing »
  - requêtes,

LIP6

UPMC SORBONNE UNIVERSITÉS



Synoptique

1. Nécessité des données et du partage
2. **Logique du partage**
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

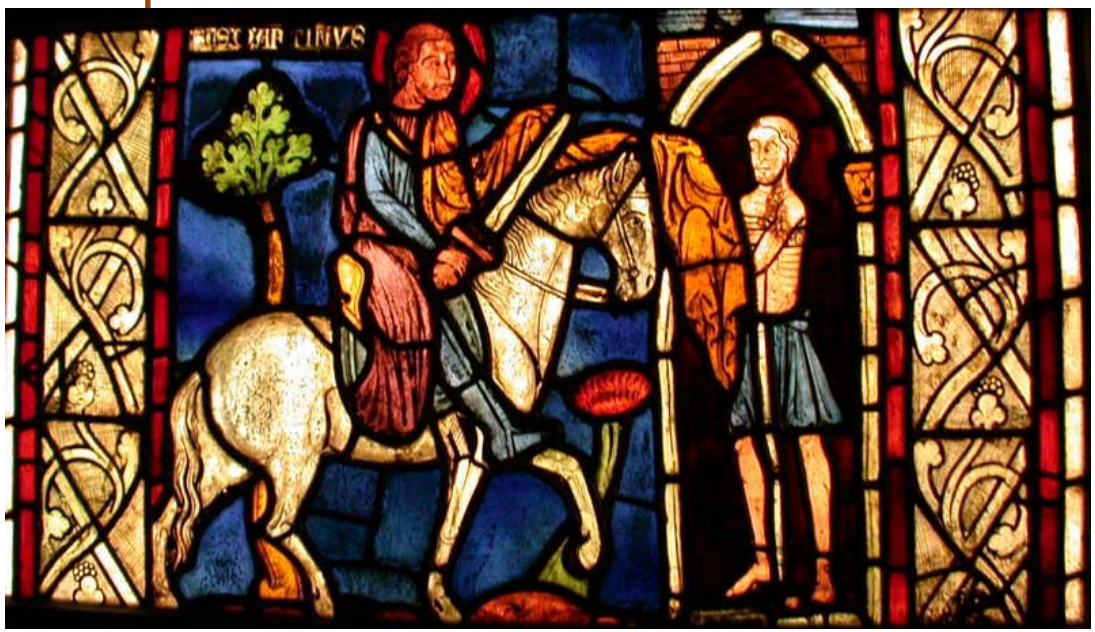
LIP6

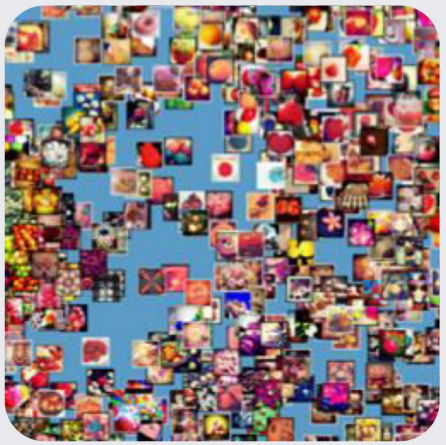
UPMC SORBONNE UNIVERSITÉS

La logique du partage

Du partage matériel, où l'on perd l'usage de ce que l'on met en

… Au partage numérique où rien n'est p

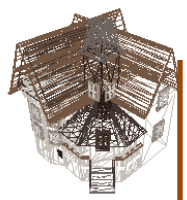




Photos Shared Per Day

2×10^9

Le partage numérique est « gagnant – gagnant »!



Synoptique

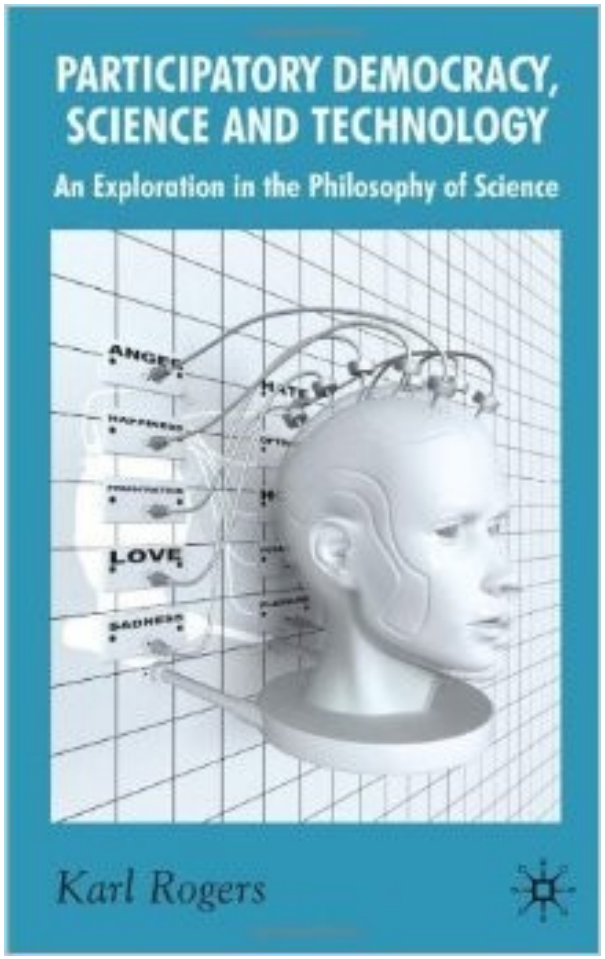
1. Nécessité des données et du partage
2. Logique du partage
3. **Vertus du partage**
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS

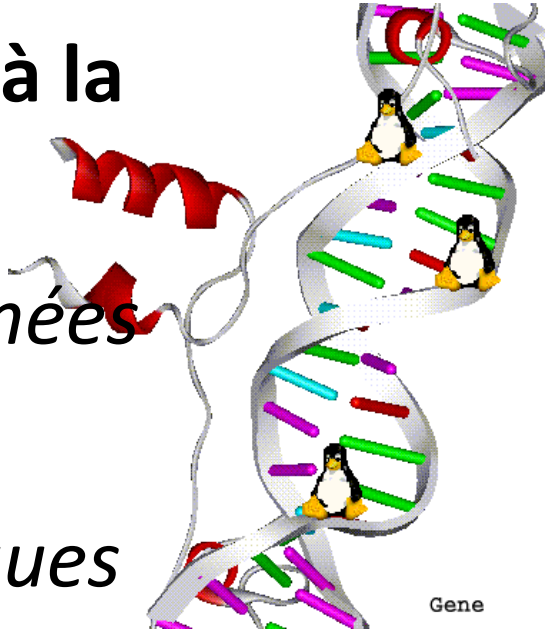
Vertus du partage et de la participation

- Donner accès à tous à la connaissance (*démocratisation*)
- Accélérer la production des connaissances
- Sensibiliser la population à la démarche scientifique
  - Science et société



Vertus du partage …

Le partage ne coûte rien – il est à la mode



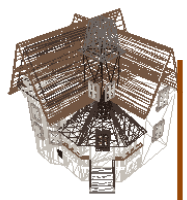
- « Data Sharing »: *partage données sc.*
- « Open data »: *données publiques ouvertes*
  - Les données sont le **pétrole de l'avenir**!
- « Open access »: *accès ouvert (publications)*

facebook IBM Google DeepMind

- …
  - Le partage est-il **équitable**?

LIP6

UPMC SORBONNE UNIVERSITÉS

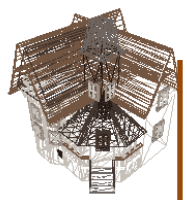


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. **Histoire du partage**
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS



Histoire du partage

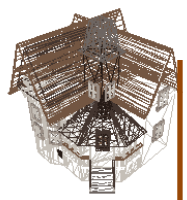
Partage de données scientifique:

- Origine: biologie, *Bermuda principles*, 1996
- Déclaration de Berlin, 2003 et 2005
- CNRS, 2014

Vient des communautés scientifiques

- Astrophysique, biologie, …
- Médecine ?? – industrie pharmaceutique?

Objectifs



Synoptique

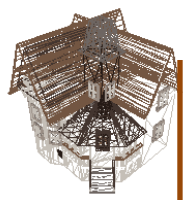
1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS

Partage: de quoi parle-t-on?

- Nature des données
  - astronomie, génome, images, parole, …
- Types de données:
  - Données primaires:
    - radio-téléscope 7 Po/mn!
  - Données secondaires
    - dérivées des données primaires
    - Filtrées, annotées, enrichies interprétées

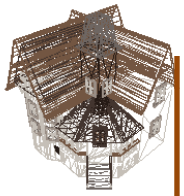


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

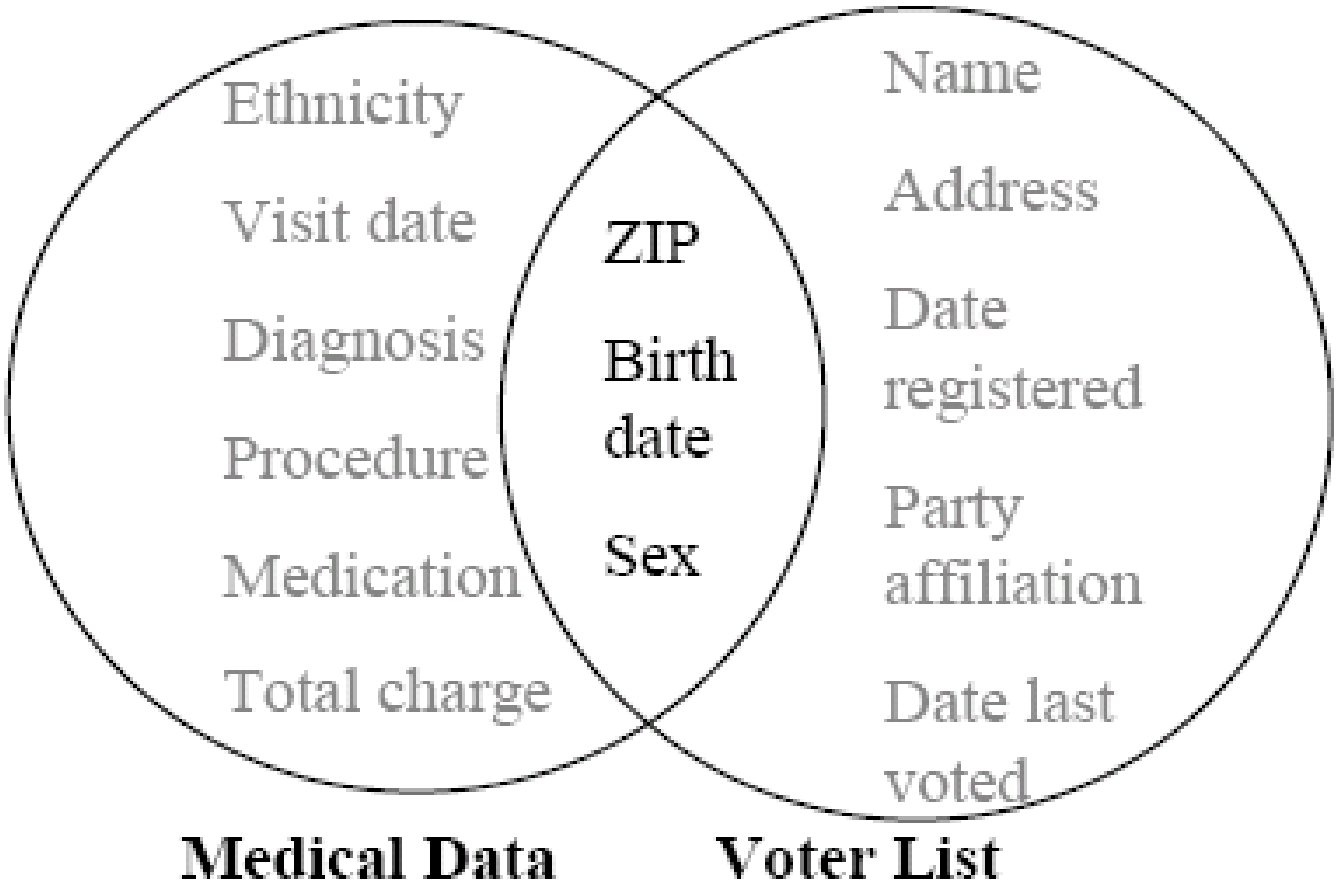
UPMC SORBONNE UNIVERSITÉS

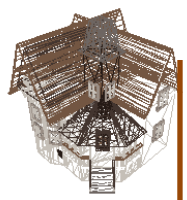


Limites

- **Données sensibles et données à caractère personnel**
  - Médecine, interfaces, images, sciences sociales, etc.
  - Anonymisation: difficultés → croisement

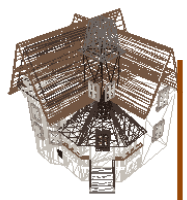
Difficultés de l'anonymisation





Limites

- **Données sensibles et données à caractère personnel**
  - Médecine, interfaces, images, sciences sociales, etc.
  - Anonymisation: difficultés → croisement
  - loi *Informatique et Libertés* du 6 janvier 1978, mod. 2004
  - CNIL: principes de finalité et de proportionalité (*peu approprié au traitement des grandes masses de données*)

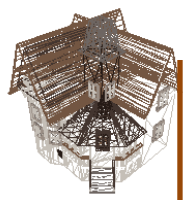


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. **Données ouvertes**
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS



« Open Data » - *Données publiques*

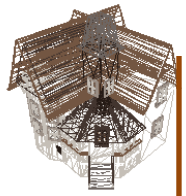
Initiatives gouvernementales

- Europe, Directive sur la réutilisation des informations du secteur public, 2003
- États-Unis, Obama, Governement 2.0, accès publiques aux archives administratives, 2008
- France ETALAB, gère l'*Open data* public sous l'autorité du Premier ministre Circulaire du 17 septembre 2013 - https://www.etalab.gouv.fr/

Nature des données

LIP6

UPMC SORBONNE UNIVERSITÉS

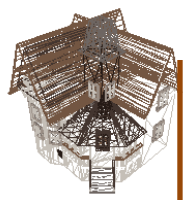


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. **Accès ouvert**
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS



Accès ouvert – « Open Access »

Définition: fournir accès en ligne gratuit à l'**« information scientifique »** - déclaration de Budapest, 2002

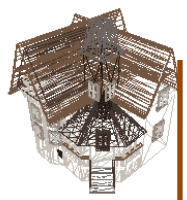
« Information scientifique »:

a) *Articles de recherche*, éventuellement évalués par les pairs

b) *Données de recherche* ayant servi aux publications (données primaires et secondaires)

LIP6

UPMC SORBONNE UNIVERSITÉS

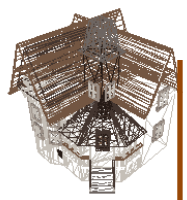


Accès ouvert aux publications

- **Auto-archivage (« green »)**: les articles évalués sont déposés en ligne **par les auteurs** sur un réceptable avant ou après publication. *Exemple: HAL,* ArXiv
- **Publication en accès ouvert (« gold »)** les articles sont mis à disposition des lecteurs en accès ouvert (et gratuit). Financement par les institutions ou les auteurs.

LIP6

UPMC SORBONNE UNIVERSITÉS



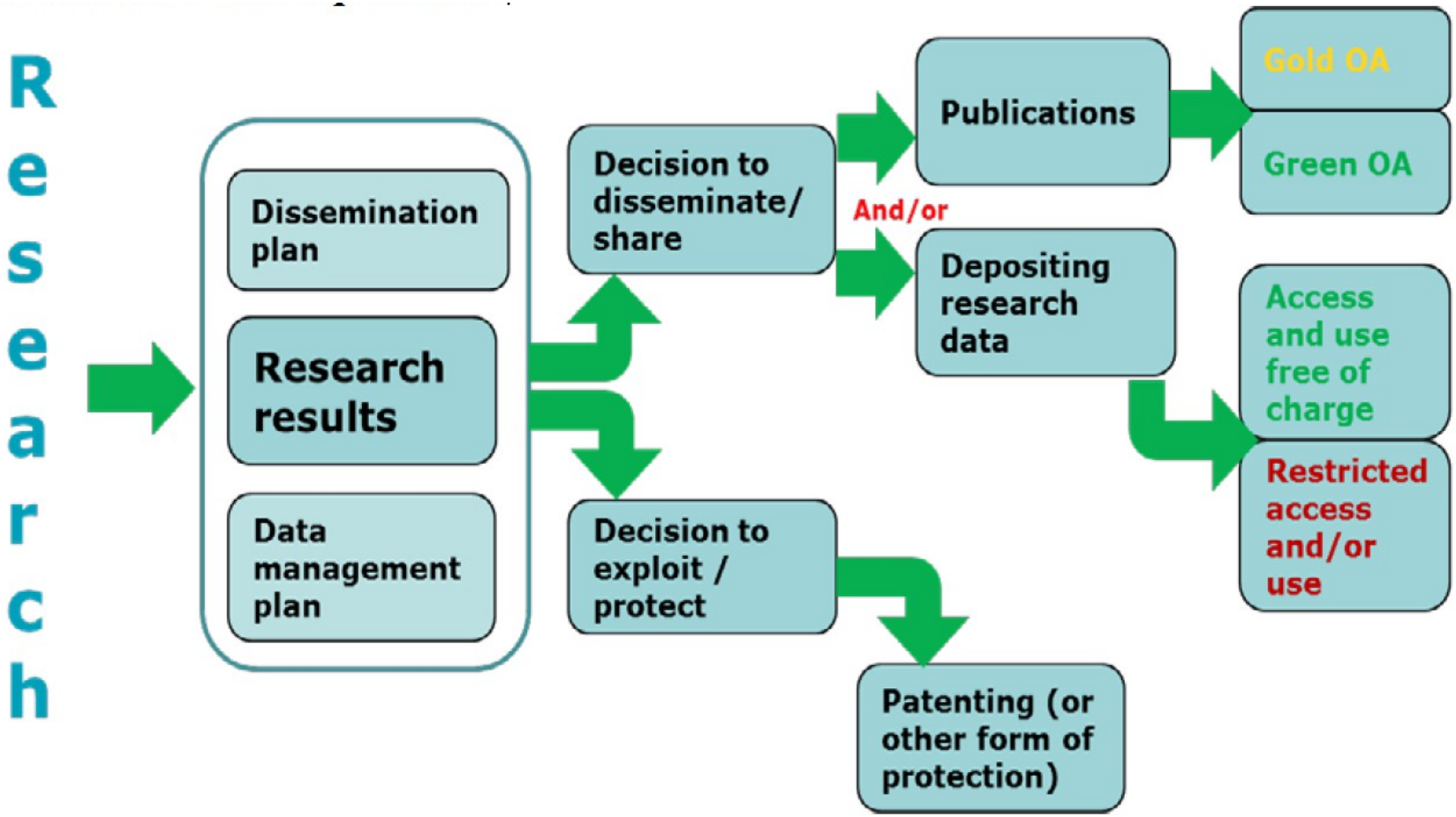
Accès ouvert aux données

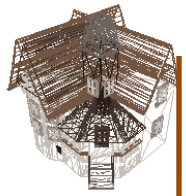
- Droit d'accèder aux données et de les réutiliser sous un ensemble de conditions spécifiées dans chaque cas.
- Les données sont accessibles et peuvent être exploitées, fouillées, reproduites et diffusées, sans charge financière.

LIP6

UPMC SORBONNE UNIVERSITÉS

Schéma d'ensemble



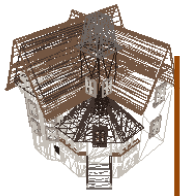


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. **Gratuité du partage**
10. Sciences participatives
11. *Crowdsourcing*
12. Références

LIP6

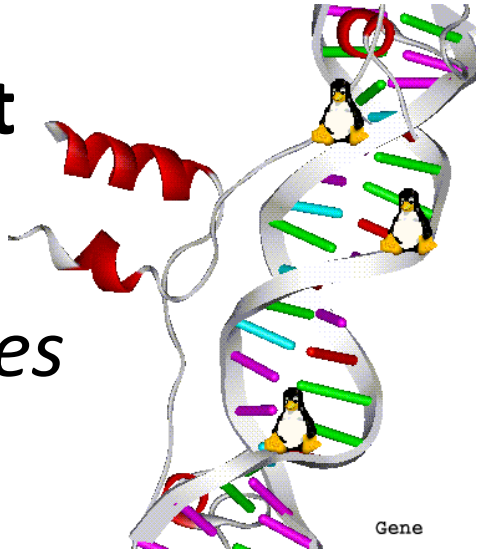
UPMC SORBONNE UNIVERSITÉS



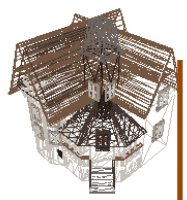
Gratuité du partage …

Le partage ne coûte rien…

… mais il n'est pas nécessairement gratuit!



- « Data Sharing »: *partage données sc.*
  Gratuit!
- « Open data »: *données publiques ouvertes,* **non nécessairement gratuites!**
- « Open access »: *accès ouvert*



Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. **Sciences participatives**
11. *Crowdsourcing*
12. Références

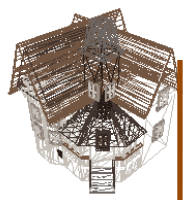
LIP6

UPMC SORBONNE UNIVERSITÉS

Sciences participatives

- Origine: années 50 ornithologie, Bonney dans un laboratoire, université de Cornell à New-York,
- Participation du pu
l'a

- *Citizen Science* terme forgé par Alan Irwin en 1995
- Aujourd'hui, *crowdsourcing*

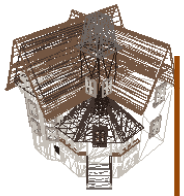


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. ***Crowdsourcing***
12. Références

LIP6

UPMC SORBONNE UNIVERSITÉS



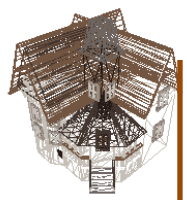
Crowdsourcing

- *Crowd*: foule
- formation: modèle *out-sourcing* (externalisation)
- exemple: Amazon Mechanical Turk
- Analogie: *crowd-f*
  Questi
  finan

amazon mechanical turk
Artificial Artificial Intelligence

LIP6

UPMC SORBONNE UNIVERSITÉS

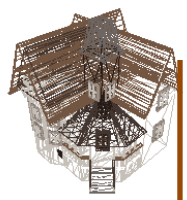


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. **Sciences participatives**
11. *Crowdsourcing*
12. Références

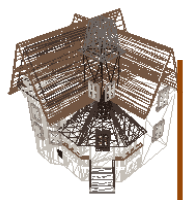
LIP6

UPMC SORBONNE UNIVERSITÉS



Modalités participation

1. **Recueil d'observations**,
   citoyens = capteurs ou processeurs élémentaires
   - **Modalité active**: le citoyen contribue consciemment
   - **Modalité passive**: le citoyen porte des capteurs, ex. santé connectée
2. « science distribuée », observation (ou calcul) + **interprétation**
3. Participation à la conception du proje

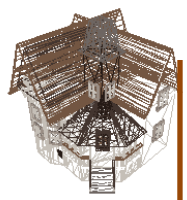


Questionnements éthiques

- **Propriété des données recueillies et des résultats des recherches effectuées**
  - Propriété morale et patrimoniale
  - Reconnaissance du travail
- **Motivation et rétribution des contributeurs** :
  - rétribution motivante…
  - Paradoxe: une forte rétribution diminue la qualité des résultats
- **Validation des contributions** :

LIP6

UPMC SORBONNE UNIVERSITÉS

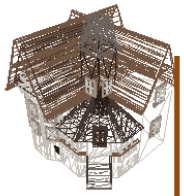


Synoptique

1. Nécessité des données et du partage
2. Logique du partage
3. Vertus du partage
4. Histoire du partage
7. Données ouvertes
8. Accès ouvert
9. Gratuité du partage
10. Sciences participatives
11. *Crowdsourcing*
12. **Références**

LIP6

UPMC SORBONNE UNIVERSITÉS



Bibliographie

- *Les enjeux éthiques du partage des données scientifiques, rapport COMETS, 2015 (site COMETS)*
- *Une science ouverte dans une république numérique - Guide stratégique d'application, CNRS, DIST*
- *Déclaration internationale sur le libre accès de Budapest* le 14 février 2002, connue sous le sigle *BOAI (Budapest Open Access Initiative*, http://www.budapestopenaccessinitiative.org/
- *Déclaration de Berlin sur le libre accès à la connaissance dans les sciences*, 2003 puis 2005