

« Science ouverte » et données de la recherche



bsn

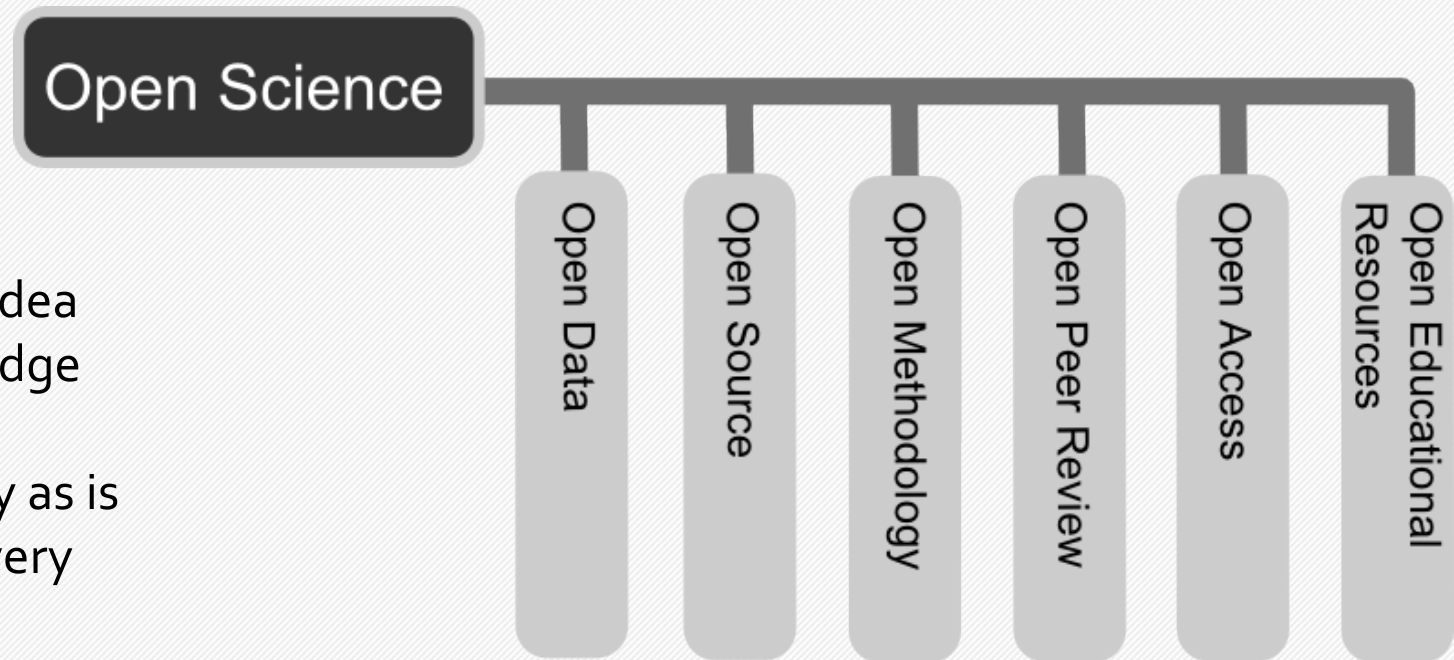
CIRM - 17.10.2016 / **Michel Roland**

1. « Open Science »

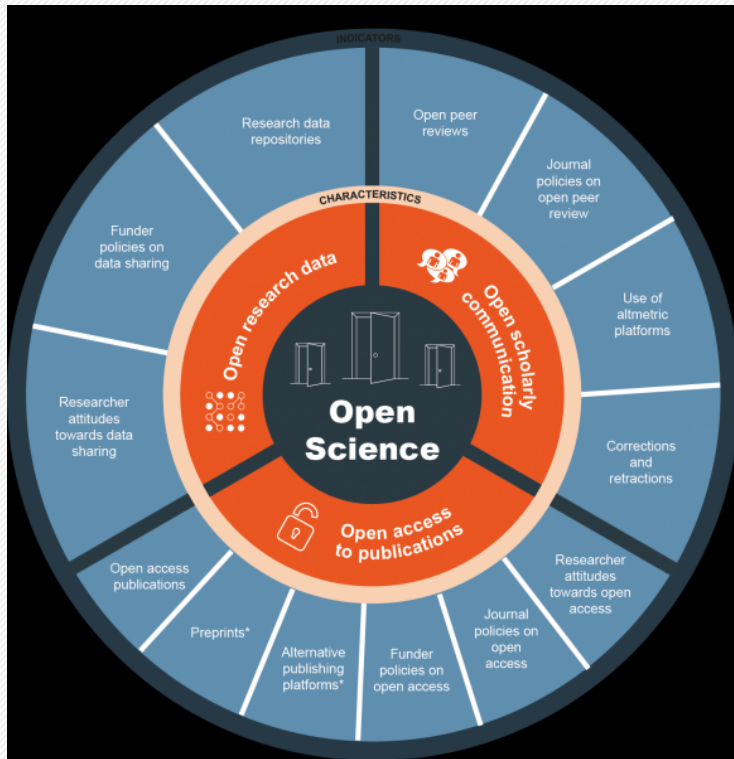
« Open Science »

“Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process.”

[Michael Nielsen](#)



« Open Science » et « Open Access »



Open Science Monitor
(Commission Européenne)

Un thème émergent:
les données de la recherche

Données de la recherche: Pourquoi l'ouverture ?

- du point de vue du chercheur ...
- du point de vue de la société ...
- du point de vue de la science ...

ENJEUX ET BÉNÉFICES

POUR LES CHERCHEURS

De nouvelles exigences et opportunités

- Le partage des données peut être **obligatoire** pour obtenir une **subvention** ou pour **publier** un article.
- Il permet aussi au chercheur de **valoriser ses données** et d'**accroître sa visibilité**.

Exemples

Les bénéficiaires d'un financement H2020 doivent planifier le dépôt de leurs données...

Les données sont citables et la paternité du chercheur est reconnue.

POUR LA SCIENCE

De nouvelles perspectives

Le partage des données permet aux chercheurs de s'appuyer sur un **socle plus important** et ouvre la voie à de **nouvelles méthodes d'investigations**.

Utiliser des bases de données de référence, approcher de différentes façons un même jeu de données...

Une meilleure validation

- L'**accès aux données améliore les conditions de validation des articles**. Aussi le dépôt des données est **de plus en plus demandé par les comités de lecture**.
- La réutilisation des données par d'autres chercheurs peut amener au **réexamen des résultats initiaux**.

POUR LA SOCIÉTÉ

La réutilisation des données permet un **meilleur usage de l'argent public** et dynamise les relations entre la recherche et la société.

Implication des citoyens dans le cadre des [Sciences participatives](#)

Le saviez-vous ?

La croissance très importante du nombre de données dans le domaine scientifique mais aussi en général ([Big data](#)) présente à la communauté scientifique **des champs et des exigences nouvelles** ([Data Driven Science](#), [Humanités Numériques](#))

2 finalités principales

- **Validation (science reproductible)**
- **Réutilisation (science cumulative)**

2 finalités principales

Retour de la réutilisation sur la validation



En s'appuyant sur vos données, ils peuvent lancer de nouvelles recherches. Ils peuvent aussi déceler un problème et invalider votre travail.

2. « Données de la recherche » *(de quoi parle-t-on au juste?)*

Une « définition » de l'OCDE

"les « données de la recherche » sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche.

Ce terme ne s'applique pas aux éléments suivants : carnets de laboratoire, analyses préliminaires et projets de documents scientifiques, programmes de travaux futurs, examens par les pairs, communications personnelles avec des collègues et objets matériels (par exemple, les échantillons de laboratoire, les souches bactériennes et les animaux de laboratoire tels que les souris). L'accès à tous ces produits ou résultats de la recherche est régi par d'autres considérations que celles abordées ici."

(Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics)

Une « définition » de l'ANDS

*« Research Data: Data are facts, observations or experiences on which an argument, theory or test is based. Data may be numerical, descriptive or visual. Data may be raw or analysed, experimental or observational. **Data includes: laboratory notebooks**; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. »*

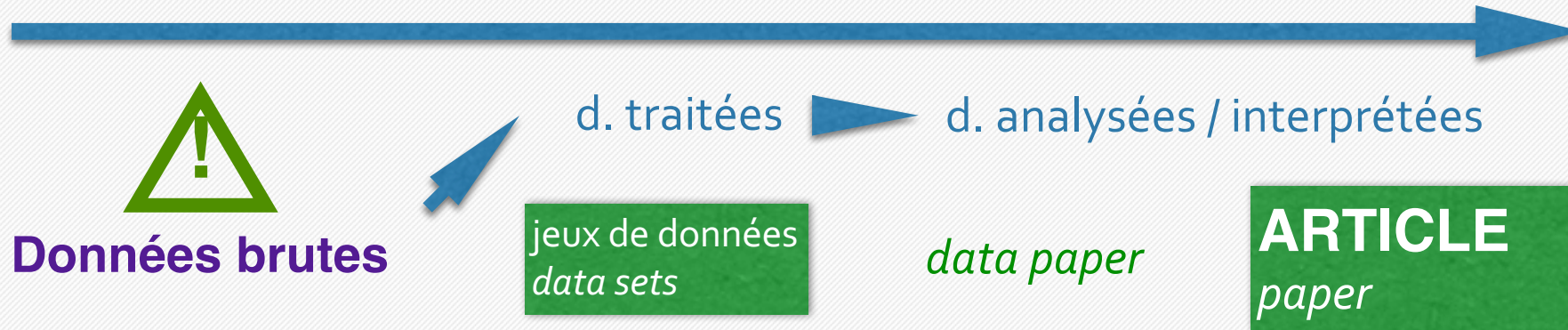
(What is research data?)

Une mise au point de Sylvie Fayet

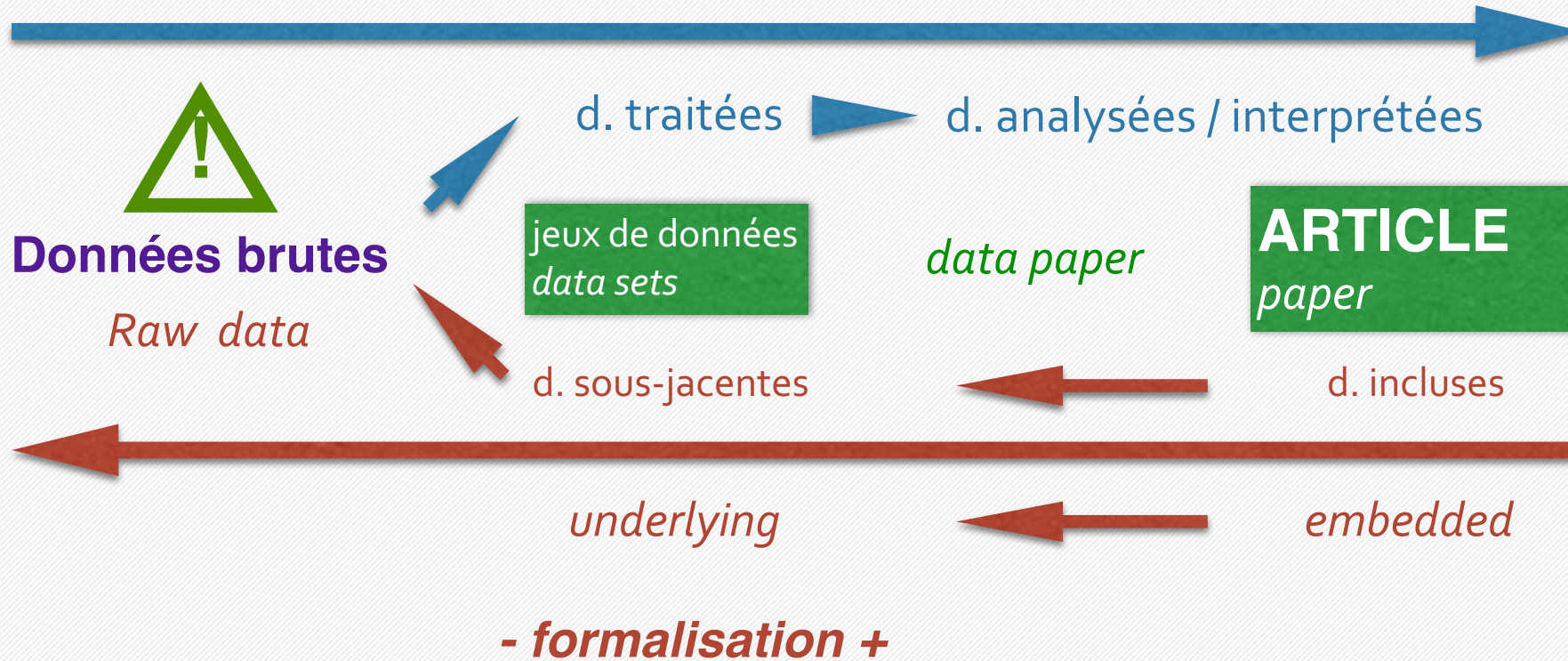
"A minima, on s'accorde implicitement sur l'idée suivante : quand on évoque « les données de la recherche », on désigne des chiffres, relevés, mesures, résultats d'expérience, réponses à des enquêtes, statistiques, comptages, et autres données quantitatives sur la base desquels va s'élaborer une hypothèse, et/ou qui serviront à infirmer ou valider cette hypothèse... **bref essentiellement du quantitatif, que l'on pourra traiter, trier, exploiter, visualiser de manière homogène.** La publication de telles données fait déjà partie, dans certaines disciplines du moins, des canons de la rédaction d'un article scientifique (par exemple, la partie « Materials and methods » dans les recommandations pour la rédaction d'articles dans des revues médicales)."

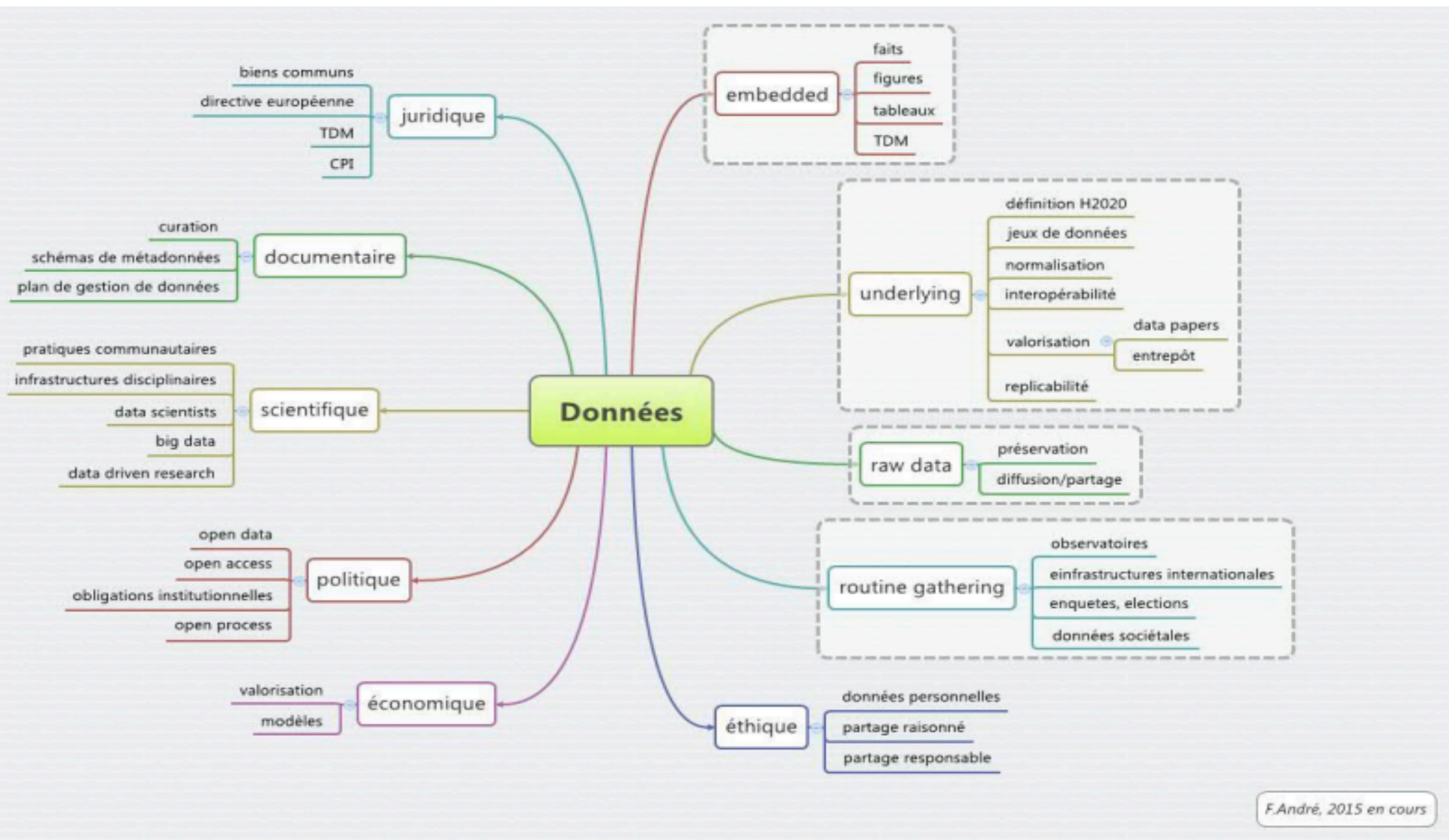
(« Données » de la recherche, les mal-nommées)

Ordre (chrono) logique



Ordre (chrono) logique





Données collectées / données produites

Numéro du site	Arbre	Année notation	Chancres	Cime	Température moyenne hivernale (degré)	Température moyenne estivale (degré)
1	1	2007	0	1	4,1	19,1
1	2	2007	1	2	4,1	19,1
1	3	2007	0	2	4,1	19,1
1	4	2007	0	4	4,1	19,1
1	...	2007	0	3	4,1	19,1
1	50	2007	1	2	4,1	19,1
1	1	2008	1	1	3,5	17,8
1	2	2008	1	1	3,5	17,8
1	3	2008	0	2	3,5	17,8

Données collectées:

- Questions juridiques et éthiques
- Big Data
- TDM

Données produites:

- Interopérabilité des recherches
- Partage > dépôt
- Conservation > archivage pérenne

3. Enjeux

3.1. Validation

Crise de la validation : le « Watergate du clonage »

Tout commence le 1er juin 2005 –soit plus de cinq mois avant ce que disent les [chronologies](#) publiées jusqu'ici. Ce jour-là, un courriel est envoyé à l'émission *PD Notebook*, émission de journalisme d'enquête diffusée sur le réseau sud-coréen MBC (Munhwa Broadcasting Corporation). [Selon ce qu'a raconté l'un des réalisateurs](#), Kim Bo Seul, l'auteur écrit que "sa conscience le trouble" à cause de problèmes liés aux recherches

Patient-Specific Embryonic Stem Cells Derived from Human SCNT Blastocysts

Woo Suk Hwang^{1,2,*}, Sung Il Roh³, Byeong Chun Lee¹, Sung Keun Kang¹, Dae Kee Kwon¹, Sue Kim¹, Sun Jong Kim³, Sun Woo Park¹, Hee Sun Kwon¹, Chang Kyu Lee², Jung Bok Lee³, Jin Mee Kim³, Curie Ahn⁴, Sun Ha Paek⁴, Sang Sik Chang⁵, Jung Jin Koo⁵, Hyun Soo Yoon⁶, Jung Hye Hwang⁶, Youn Young Hwang⁶, Ye Soo Park⁶, Sun Kyung Oh⁴, Hee Sun Kim⁴, Jong Hyuk Park⁷, Shin Yong Moon⁴, Gerald Schatten^{7,*}

<http://www.sciencemag.org/site/feature/misc/webfeat/hwang2005/>

Crise de la validation : le « Watergate du clonage » 2



Le 4 décembre, MBC [suspend PD Notebook](#) et décide de ne pas diffuser le deuxième reportage.

L'histoire a donc failli mourir à ce moment. Mais c'était sans compter la façon dont fonctionne la science. Le 5 décembre, un courriel anonyme est envoyé à un service d'information hyper-spécialisé destiné aux jeunes chercheurs sud-coréens, BRIC (Biological Research Information Center). Intitulé, en anglais, "The show must go on", le courriel invite les lecteurs à chercher des photos dupliquées dans l'article de mai 2005, ajoutant: "j'en ai trouvé deux!"

Plus de 200 messages suivent en moins de deux jours, plusieurs identifiant effectivement des photos qui ne sont que des copies les unes des autres. Le 7 décembre, Hwang reconnaît avoir, "par inadvertance", dupliqué les photos. Mais déjà, sur le forum du BRIC, un autre courriel pointe des données d'ADN des soi-disant clones qui ne seraient que des copies d'ADN d'embryons déjà connus. Ces informations émergent dans les médias sud-coréens, puis font le tour du monde.

3.2. Réutilisation, etc.

The End of Theory?

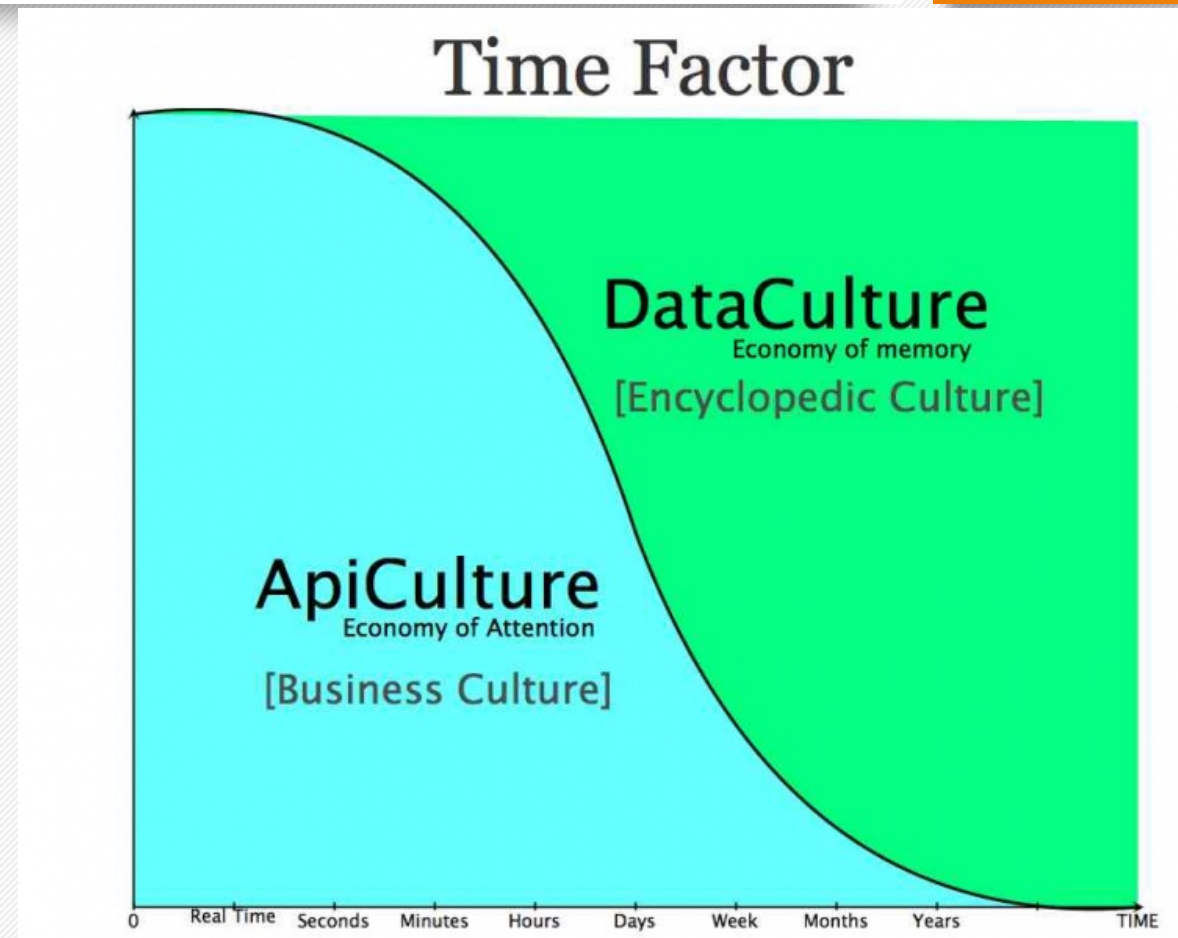
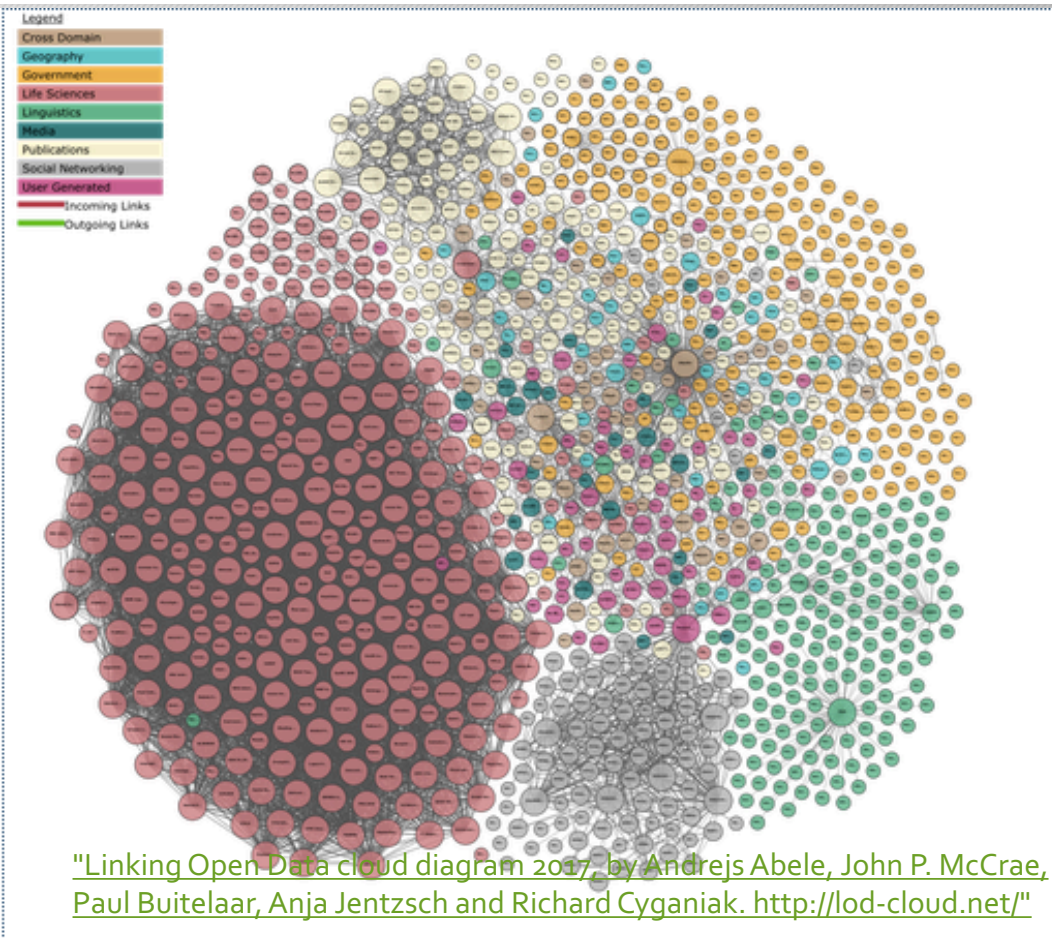
4^e paradigme (Jim Grey, 2006):

"ideas like recursion, parallelism and abstraction taken from computer science will redefine modern science. Implicit in the idea of a fourth paradigm is the ability, and the need, to share data. In sciences like physics and astronomy, the instruments are so expensive that data must be shared. Now the data explosion and the falling cost of computing and communications are creating pressure to share all scientific data. » (J. Wing in: John Markoff, "A Deluge of Data Shapes a New Era in Computing, "New York Times, 14 décembre 2009)

The End of Theory (Chris Anderson, 2008):

The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

Web sémantique / Web de données



Humanités numériques



« a widening rift between print- and digital-based scholars...

Needed are approaches that can locate digital work within print traditions, and print traditions within digital media, without obscuring or failing to account for the differences between them.. »

(How We Think: Digital Media and Contemporary Technogenesis / Katherine Hayles, 2012)

4. DoRANum

THEMATIQUES



ENJEUX - BÉNÉFICES

Pourquoi partager les données ? Qu'est-ce que l'Open Science ?



DMP - PLAN DE GESTION DES DONNÉES

Pourquoi et comment rédiger un plan de gestion des données ?



ID PERMANENTS - DOI

Comment créer un jeu de données ?



MÉTADONNÉES - FORMATS

Comment décrire mes données ?



ASPECTS JURIDIQUES ET ÉTHIQUES

Que puis-je partager ? Que puis-je réutiliser ?



DEPOT - ENTREPÔT

Où et comment déposer mes données ?



ACCÈS - VISUALISATION

Où et comment retrouver les données qui m'intéressent ?



DATA PAPERS - DATA JOURNALS

Pourquoi et comment publier mes données dans un article scientifique ?



STOCKAGE - ARCHIVAGE

Quelles données conserver à long terme et comment ?

FICHES SYNTHÉTIQUES

VOIR



INFOGRAPHIES

VOIR



TUTORIELS

PLAY



MAIS AUSSI...

VOIR



<http://doranum.fr>



bsn



Merci de votre attention !



Contacts:

Paolo Lai paolo.lai@inist.fr

Michel Roland michel.roland@unice.fr