Enjeux de la donnée universitaire, de sa collecte à son exploitation

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'avenir portant la référence ANR-20-IDES-0001.

Qui suis-je ?

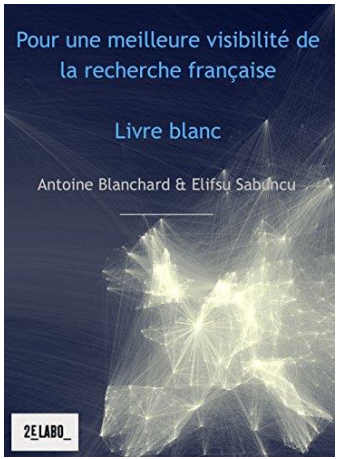
Antoine Blanchard, Responsable du service Données, décisionnel, datalab (3D) de l'université de Bordeaux

- **en charge de l'équipe décisionnelle du Pôle Pilotage et aide à la stratégie**
- **en charge du projet Datalab UBx qui vise à développer, à l'échelle du programme ACT, des actions de stratégie data, de gouvernance des données et de développement des usages de la donnée**

2020 – 2023 : consultant *open data* et science ouverte (société coopérative Datactivist)

2014 – 2020 : chargé de projets et de programmes d'innovation numérique pour l'IdEx Bordeaux (université de Bordeaux)

2009 – 2014 : consultant communication et diffusion des sciences (Deuxième labo)



1. Des données pour quoi faire ?

université de BORDEAUX

Définitions

Données

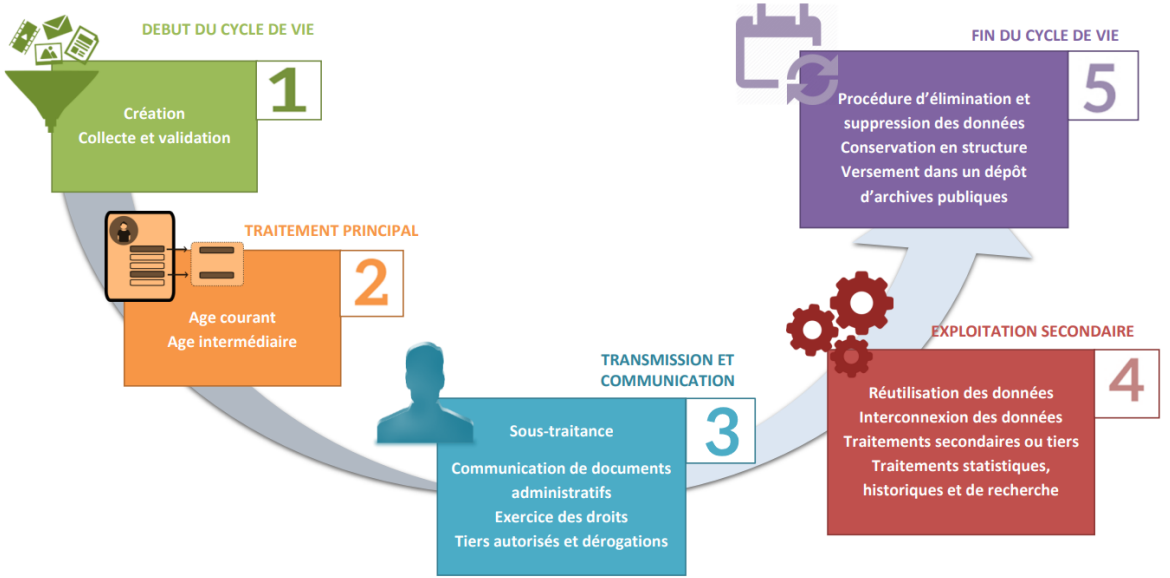
La donnée est la représentation d'une information sous une forme conventionnelle destinée à faciliter son traitement.

– Arrêté du 22 décembre 1981, « Enrichissement du vocabulaire de l'informatique », *Journal officiel de la république française* du 17 janvier 1982 (numéro complémentaire)

Une donnée n'est jamais donnée, elle est obtenue : elle naît toujours d'une opération.

Définitions

Cycle de vie des données



– AAF et Sup'DPO, *Les étapes de la vie d'une donnée au regard des règlementations « CRPA », « RGPD » et « Patrimoine »*, http://supdpo.fr/wp-content/uploads/2019/07/Aurore-SupDPO-Vie-dune-donne%CC%81e-v1.pdf

Définitions

Organisation des données

Table de données : structure en lignes et colonnes qui stocke des attributs (en colonnes) relatifs à des objets (en ligne)

Base de données : ensemble de tables liées entre elles (par ex. SQL)

Entrepôt de données (*data warehouse*) : collection de données métier orientées sujet, modélisées avec la granularité la plus fine et le plus de dimensions d'analyse possible

Lac de données (*data lake*) : stockage non structuré de données variées dans leur état brut

Quelles sont les données de l'université ?

Données transactionnelles

Un acte de gestion impliquant l'université produit un ensemble de données qui sont enregistrées dans le système d'information (SI) métier.

- **examen** : date, UE, note, identifiant étudiant…
- **recrutement** : dates de début et fin de contrat, rémunération, type de contrat
- **convention de partenariat** : date d'entrée en vigueur, organisations signataires…

Quelles sont les données de l'université ?

Données de référence

Données partagées par l'ensemble des processus qui soutiennent les activités courantes d'un domaine d'affaires.

– Ville de Montréal, *Directive sur la gouvernance des données*, juin 2022

- données nomenclatures : nomenclatures RH…
- données maîtres : référentiel des structures de l'établissement, des locaux…

Les données de référence sont les données qui ont une durée de vie plus longue que celle du processus de gestion.

Quelles sont les données de l'université ?

Données décisionnelles

Données consolidées permettant des analyses statistiques et l'édition de rapports afin de piloter les activités, d'élaborer des stratégies et de suivre leur implémentation.

– Nicolas Koudlansky, *Présentation de la solution Sinaps*, webinaire Amue : Paris, 31 mars 2022

- données sur les inscriptions des étudiants…

Réutilisent les données transactionnelles et les données de référence.

Quelles sont les données de l'université ?

Données de monitoring / temps réel

Données acquises et transmises en temps réel sans l'intervention humaine, typiquement à l'aide d'un enregistreur de données (*data logger*).

- données de consommation des fluides…
- données de connexion à un examen en ligne…
- données de comptage du trafic de véhicules, du franchissement des portiques de bibliothèque…

Quelles sont les données de l'université ?

Données administratives

Les données administratives sont les renseignements recueillis par l'université dans le cadre de ses activités courantes de gestion :

- finances, budget, achat, comptabilité…
- RH, santé et sécurité au travail…
- affaires juridiques et institutionnelles…
- patrimoine et environnement
- stratégie, pilotage, évaluation, amélioration continue…
- recherche, formation et innovation…

– Philippe Lahire, Fabienne Bonetto, Sylvie Haouy, et Stéphane Martinez. « Vers une gouvernance des données au service des missions de l'établissement et de son pilotage ». Lab'U #3 Contrôle de gestion et gouvernance de la donnée, Amue : Paris, 21 mars 2022

Quelles sont les données de l'université ?

Données de recherche

Les données de la recherche sont 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 les résultats de la recherche.

– OCDE, *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*, 2007

- données de capteurs, données haut débit (génomique…)
- données d'observation (astronomie…)
- données expérimentales (physique, chimie…)
- données de modélisation, de simulation
- données d'enquête, statistiques (démographie, sociologie, économie…)
- données du web, des réseaux sociaux (sociologie, science politique…)
- texte et multimédia (histoire, arts, lettres, langues…)

Quelles sont les données de l'université ?

Données pédagogiques

Données produites par les enseignants et les apprenants dans le cadre des activités d'apprentissage, typiquement dans les outils pédagogiques (LMS, plateformes de MOOC et e-portfolios).

- données qualitatives : réponses à des formulaires…
- traces : interactions d'un usager avec l'environnement d'apprentissage…

L'analyse des données pédagogiques (*learning analytics*) vise à améliorer les résultats des apprenants, renforcer leur engagement, optimiser leurs expériences d'apprentissage.

– Florence Cherigny et coll., *L'analytique des apprentissages avec le numérique*, GTnum2 / Direction du numérique pour l'Éducation (DNE-MENJ), mars 2020

Que faire avec ses données ?

Usage primaire : gestion de l'activité

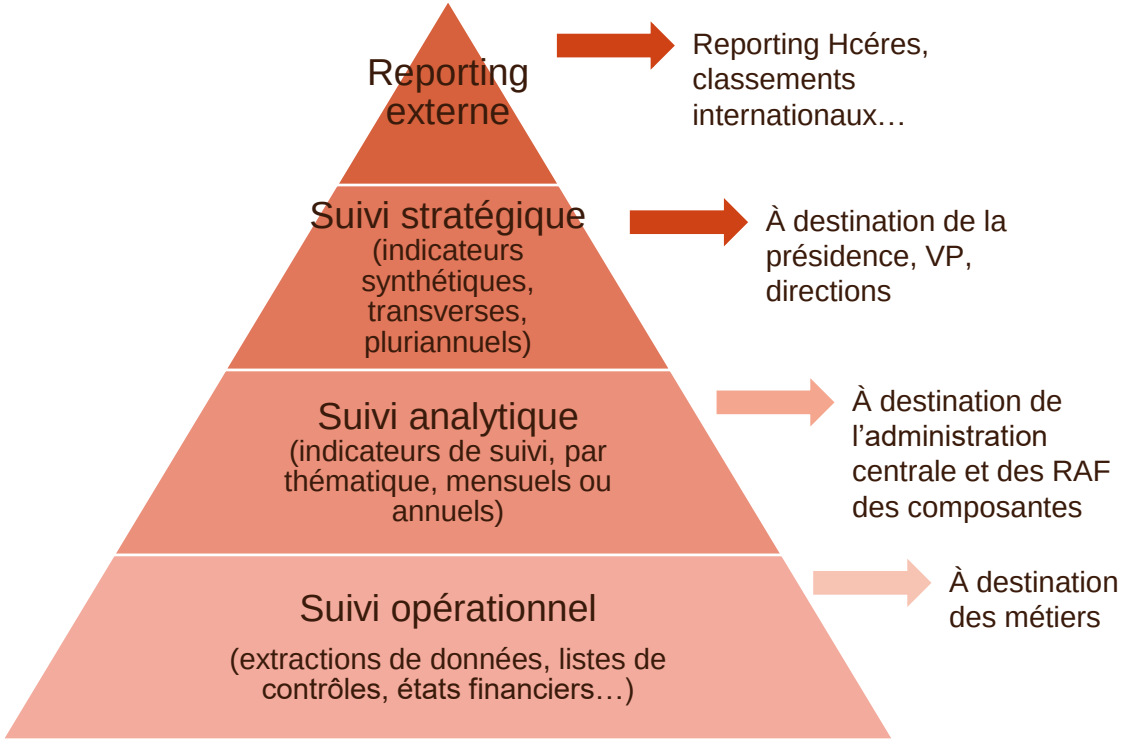
Pour remplir ses fonctions, l'Université met en œuvre des processus de gestion qui s'appuient sur des données.

- admettre l'étudiant à un examen selon ses résultats

Ces données ont une valeur probante (cf. normes *ISO 30300 Systèmes de gestion des documents d'activité* et *ISO 15489-1 Gestion des documents d'activité*) et doivent permettre de justifier les droits des personnes morales ou physiques.

Que faire avec ses données ?

Usage secondaire : suivi de l'activité



– *Présentation du projet SIROCCO*, juin 2023

Que faire avec ses données ?

Usage secondaire : recherche

- recherches empiriques sur le système d'enseignement supérieur et de recherche (ex. CPESR, RESUP)

- « métascience » ou « recherche sur la recherche » avec une visée normative pour faire évoluer les pratiques scientifiques afin de les rendre plus efficaces, de meilleure qualité, mais aussi de changer la culture de la recherche

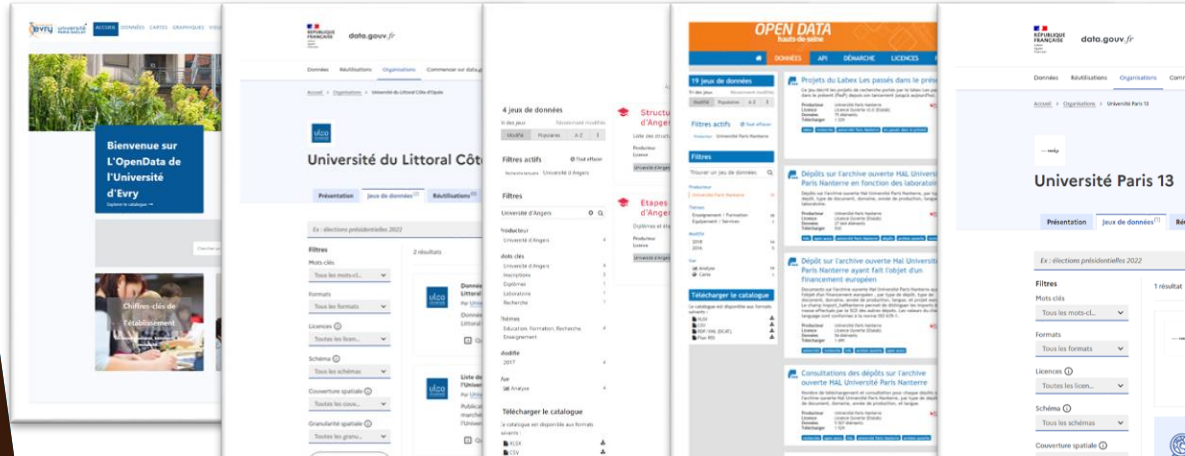
– Célya Gruson-Daniel et Maya Anderson-González, *Étude exploratoire sur la « recherche sur la recherche » : acteurs et approches*, Comité pour la science ouverte, 2021

Que faire avec ses données ?

Usage secondaire : open data

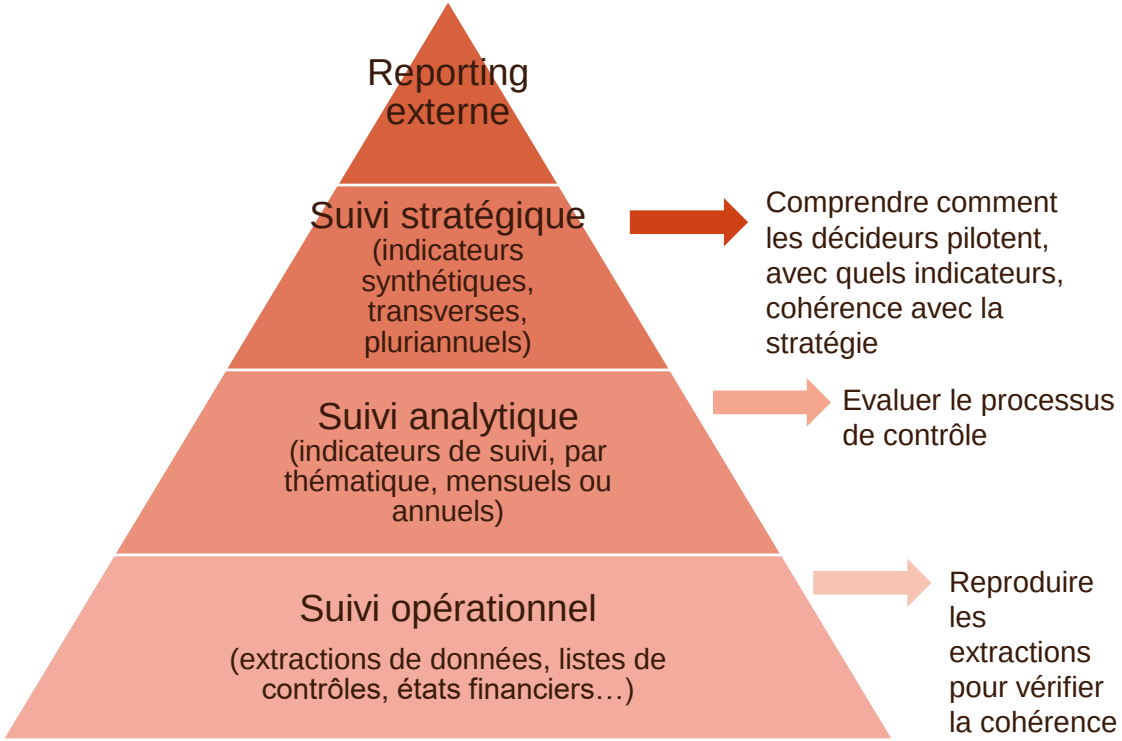
Depuis la loi pour une République numérique (2016), les administrations sont tenues de « publier en ligne les documents administratifs suivants (…) :

- Les bases de données, mises à jour de façon régulière, qu'elles produisent ou qu'elles reçoivent et qui ne font pas l'objet d'une diffusion publique par ailleurs ;
- Les données, mises à jour de façon régulière, dont la publication présente un intérêt économique, social, sanitaire ou environnemental. »



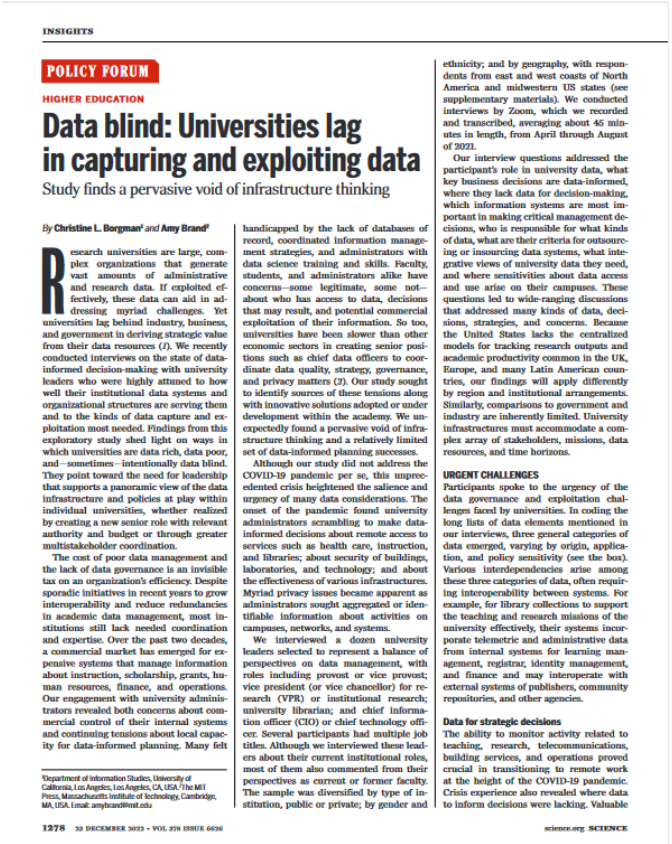
Que faire avec ses données ?

Usage secondaire : audit interne



Des données sous-exploitées

Que faire avec ses données ?



INSIGHTS

POLICY FORUM

HIGHER EDUCATION

Data blind: Universities lag in capturing and exploiting data

Study finds a pervasive void of infrastructure thinking

By Christine L. Borgman[1] and Amy Brand[2]

Research universities are large, complex organizations that generate vast amounts of administrative and research data. If exploited effectively, these data can aid in addressing myriad challenges. Yet universities lag behind industry, business, and government in deriving strategic value from their data resources (*1*). We recently conducted interviews on the state of data-informed decision-making with university leaders who were highly attuned to how well their institutional data systems and organizational structures are serving them and to the kinds of data capture and exploitation most needed. Findings from this exploratory study shed light on ways in which universities are data rich, data poor, and—sometimes—intentionally data blind. They point toward the need for leadership that supports a panoramic view of the data infrastructure and policies at play within individual universities, whether realized by creating a new senior role with relevant authority and budget or through greater multistakeholder coordination.

The cost of poor data management and the lack of data governance is an invisible tax on an organization's efficiency. Despite sporadic initiatives in recent years to grow interoperability and reduce redundancies in academic data management, most institutions still lack needed coordination and expertise. Over the past two decades, a commercial market has emerged for expensive systems that manage information about instruction, scholarship, grants, human resources, finance, and operations. Our engagement with university administrators revealed both concerns about commercial control of their internal systems and continuing tensions about local capacity for data-informed planning. Many felt handicapped by the lack of databases of record, coordinated information management strategies, and administrators with data science training and skills. Faculty, students, and administrators alike have concerns—some legitimate, some not—about who has access to data, decisions that may result, and potential commercial exploitation of their information. So too, universities have been slower than other economic sectors in creating senior positions such as chief data officers to coordinate data quality, strategy, governance, and privacy matters (*2*). Our study sought to identify sources of these tensions along with innovative solutions adopted or under development within the academy. We unexpectedly found a pervasive void of infrastructure thinking and a relatively limited set of data-informed planning successes.

Although our study did not address the COVID-19 pandemic per se, this unprecedented crisis heightened the salience and urgency of many data considerations. The onset of the pandemic found university administrators scrambling to make data-informed decisions about remote access to services such as health care, instruction, and libraries; about security of buildings, laboratories, and technology; and about the effectiveness of various infrastructures. Myriad privacy issues became apparent as administrators sought aggregated or identifiable information about activities on campuses, networks, and systems.

We interviewed a dozen university leaders selected to represent a balance of perspectives on data management, with roles including provost or vice provost; vice president (or vice chancellor) for research (VPR) or institutional research; university librarian; and chief information officer (CIO) or chief technology officer. Several participants had multiple job titles. Although we interviewed these leaders about their current institutional roles, most of them also commented from their perspectives as current or former faculty. The sample was diversified by type of institution, public or private; by gender and ethnicity; and by geography, with respondents from east and west coasts of North America and midwestern US states (see supplementary materials). We conducted interviews by Zoom, which we recorded and transcribed, averaging about 45 minutes in length, from April through August of 2021.

Our interview questions addressed the participant's role in university data, what key business decisions are data-informed, where they lack data for decision-making, which information systems are most important in making critical management decisions, who is responsible for what kinds of data, what are their criteria for outsourcing or insourcing data systems, what integrative views of university data they need, and where sensitivities about data access and use arise on their campuses. These questions led to wide-ranging discussions that addressed many kinds of data, decisions, strategies, and concerns. Because the United States lacks the centralized models for tracking research outputs and academic productivity common in the UK, Europe, and many Latin American countries, our findings will apply differently by region and institutional arrangements. Similarly, comparisons to government and industry are inherently limited. University infrastructures must accommodate a complex array of stakeholders, missions, data resources, and time horizons.

URGENT CHALLENGES

Participants spoke to the urgency of the data governance and exploitation challenges faced by universities. In coding the long lists of data elements mentioned in our interviews, three general categories of data emerged, varying by origin, application, and policy sensitivity (see the box). Various interdependencies arise among these three categories of data, often requiring interoperability between systems. For example, for library collections to support the teaching and research missions of the university effectively, their systems incorporate telemetric and administrative data from internal systems for learning management, registrar, identity management, and finance and may interoperate with external systems of publishers, community repositories, and other agencies.

Data for strategic decisions

The ability to monitor activity related to teaching, research, telecommunications, building services, and operations proved crucial in transitioning to remote work at the height of the COVID-19 pandemic. Crisis experience also revealed where data to inform decisions were lacking. Valuable

[1]Department of Information Studies, University of California, Los Angeles, Los Angeles, CA, USA. [2]The MIT Press, Massachusetts Institute of Technology, Cambridge, MA, USA. Email: amybrand@mit.edu

1278 22 DECEMBER 2022 • VOL 378 ISSUE 6626

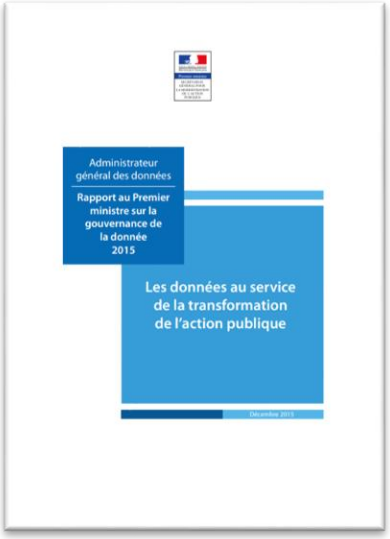
science.org SCIENCE

– Christine L. Borgman et Amy Brand, « Data blind: Universities lag in capturing and exploiting data », *Science* **378**(6626):1278-81, 23 décembre 2022

Que faire avec ses données ?

Une limite : la gouvernance des données

Ch. 2 « **Le manque de gouvernance des données comme frein au potentiel des données** »



« Focalisé sur la fiabilité, la sécurité et la maîtrise des coûts, l'Etat a négligé l'interopérabilité, l'accessibilité et la capacité d'usage, et a donc toléré une culture de silos, des divergences de formats avec des qualités excessives ou au contraire dégradées, une sous-traitance excessive et une perte globale de souveraineté et d'autonomie sur ses propres données. »

– Administrateur général des données, *Les données au service de la transformation de l'action publique. Rapport au Premier ministre sur la gouvernance de la donnée*, 2015

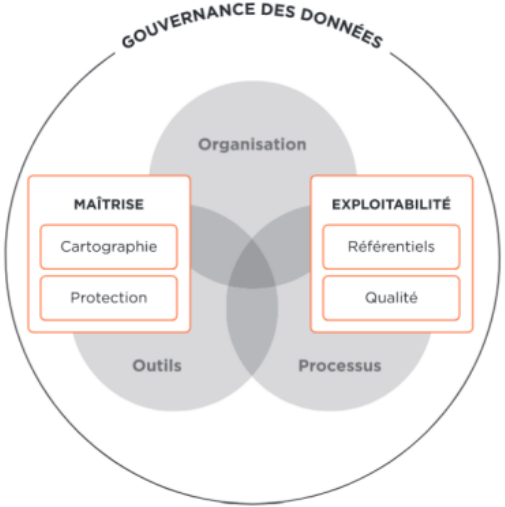
2. Pourquoi gouverner ses données ?

université de BORDEAUX

Gouvernance des données

La gouvernance regroupe l'organisation, les outils et les processus visant à accroître la maîtrise et l'exploitabilité des données.

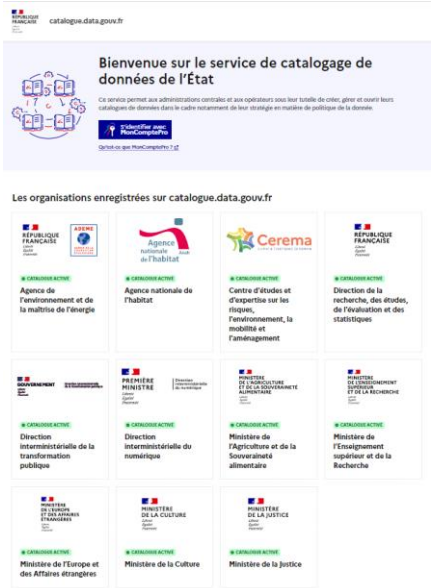
– Mick Lévy, *Sortez vos données du frigo. Une entreprise performante avec la Data et l'IA*, Dunod, 2021



Maîtriser ses données

Identification

Il est nécessaire de recenser le patrimoine de données afin de le connaître et le caractériser (et il y a parfois des surprises !).



La cartographie des données puis leur catalogage sont des démarches complexes qui s'inscrivent dans la durée. Les premiers recensements à opérer concernent les données personnelles au titre de la mise en conformité RGPD, puis les jeux de données destinés à être publiés en open data.

– Aymeric Buthion et coll., *Gestion des données : quels outils et quelle stratégie pour les territoires ?*, Banque des territoires, novembre 2020

Maîtriser ses données

Authenticité, fiabilité, intégrité

Authenticité : on peut prouver que la donnée est bien ce qu'elle prétend être, qu'elle a été créée ou envoyée par celui qui prétend l'avoir créée ou envoyée, au moment prétendu.

Fiabilité : le contenu peut être considéré comme la représentation complète et exacte des opérations, activités ou faits qu'il atteste, sur lequel on peut s'appuyer lors d'opérations ou d'activités ultérieures.

Intégrité : la donnée est dans un état non altéré.

- calcul et comparaison de l'empreinte (*hash*) d'un fichier de données dans un système d'archivage

– Norme *ISO 15489-1 Gestion des documents d'activité*, 2016

Maîtriser ses données

Sécurité informatique

Risques liés à la sécurité :
- fuite ou revente de données
- modifications non désirées, effacement de données

Mesures techniques de sécurité :
- chiffrement des données au repos et en transit
- pseudonymisation et cloisonnement des données identifiantes
- authentification forte des accès (double authentification, ou code temporaire, ou carte à puces…)
- contrôle proactif et régulier des traces (outils pour détecter les comportements inhabituels sur les postes de travail, journalisation des accès…)

Mesures organisationnelles de sécurité :
- revue des accès
- gestion des habilitations
- contrôle des droits

– Erik Boucher de Crevecoeur, « Cyber-risques et mesures de sécurité. Gouvernance de la protection des données », *La protection des données de santé : du soin à la recherche*, Université de Bordeaux : Bordeaux, 23 juin 2022

Maîtriser ses données

Confidentialité

Risques liés à la confidentialité :

- risque financier : sanction administrative de la CNIL, action collective…
- risque juridique : responsabilité pénale des dirigeants, action judiciaire d'un agent ou d'un usager
- risque réputationnel : mise en demeure ou sanction CNIL publique, action judiciaire, fuite de données dévoilées par la presse
- risque patrimoine informationnel : perte de données suite à une violation, retrait d'une certification.

En 2022 la CNIL a reçu 4 000 notifications de violation de données et 12 000 plaintes, pour 101 M€ d'amendes infligées !

– Erik Boucher de Crevecoeur, « Cyber-risques et mesures de sécurité. Gouvernance de la protection des données *», La protection des données de santé : du soin à la recherche*, Université de Bordeaux : Bordeaux, 23 juin 2022

Garantir l’exploitabilité de ses données

Disponibilité

La donnée peut être localisée, récupérée, communiquée et interprétée dans un temps raisonnable.

- récupérer et traiter les champs relatifs à l’inscription administrative dans le logiciel de scolarité
- récupérer et traiter les données relatives à la consommation énergétique dans l’outil en ligne de suivi

– Norme *ISO 15489-1 Gestion des documents d'activité*, 2016

Garantir l’exploitabilité de ses données

Complétude, cohérence

Complétude : la couverture des données correspond au périmètre attendu et les valeurs sont bien renseignées

- toutes les thèses soutenues doivent être présentes avec une date de soutenance renseignée (pas de données manquantes)

Cohérence : les données doivent être équivalentes, quel que soit le système d’information ou l’emplacement où elles sont stockées.

- le nom de l’étudiant doit être saisi une seule fois et appelé partout où c’est nécessaire, plutôt que donner lieu à des saisies multiples

Garantir l’exploitabilité de ses données

Validité, précision

Validité : les valeurs possèdent les caractéristiques attendues (intervalle borné, valeur conforme aux règles de gestion, typage…).

- le dernier caractère d’un numéro INE sert de clé de contrôle

Précision : les données ont la granularité voulue.

- la table de suivi des mobilités étudiantes donne le pays de destination, pas seulement le continent

Garantir l’exploitabilité de ses données

Actualité, fraîcheur

Actualité : les données reflètent l’état des choses telles qu’elles sont à l’instant.

- des processus efficaces produisent des données actuelles

Fraîcheur : le délai entre le moment où les données sont produites et celui où elles sont disponibles est le plus réduit possible.

- les automatisations optimisent la fraîcheur des données disponibles

3. Comment gouverner ses données ?

université de BORDEAUX

Comment gouverner ses données ?

L’université de Bordeaux peut s’appuyer sur une note de la Direction des archives universitaires sur la gouvernance des données (mai 2021).

Voyons comment s’y sont prises l’Université de Rennes et Aix Marseille Université !

Les 3 points à retenir

- Les données des universités ont des provenances et des usages variés
- La gouvernance des données permet d'accroître la maîtrise et l'exploitabilité des données.
- Sans cette gouvernance, le potentiel des données sera freiné

Le conseil de l'expert

→ Lisez « Sortez vos données du frigo » de Mick Lévy, qui vulgarise bien le sujet et propose des fiches pratiques

www.u-bordeaux.fr

@univbordeaux

@universitedebordeaux

@univbordeaux

@universitedebordeaux

Appli mobile U&me

@univbordeaux

antoine.blanchard@u-bordeaux.fr

université de BORDEAUX