Formulaire - GOD

Leonard Cseres | January 15, 2026

PI L'ensemble des droits accordés à des créations intellectuelles

PI littéraire et artistique
- moraux
- partimoniaux

PI industrielle
1. nouveauté
2. caractère inventif
3. application industrielle

Brevet Titre de PI

Copyleft Allowed to use but must use the same original license

CC Properties
- BY: attribution (always)
- SA: share alike (derivatives) (− combined licenses difficult)
- ND: no derivatives (− data preprocessing friction)
- NC: no commercial use (− not compatible Wikipedia)

CC Licenses
- BY (CC-BY 4.0)
- BY SA
- BY ND
- BY NC
- BY NC SA
- BY NC ND

CC licenses cannot be changed after creation for the same content under license

CC0
Fully open, for facts or data. (no attribution)
TODO public domain dedication and public domain mark

Cardre legal Vaud (LRrD), Suisse (LPD, LRH), Europe (RGPD or GDPR in eng.)
- LRH: …recheche sur l'être humain
- RGPD: …règlemennt euro. sur la protection des données.

Data Management Plan (DMP)

1. Plan
- gestion de projet
- responsable de données

2. Collect
- is/how data is reused
- file formats
- how data is collected
- data integrity/reproducibility

3. Organize/analyse
- filenames
- sorting
- languages

4. Preserve/archive
- how long to preserve
- how is data saved
- rule:
  - 3: copies of data
  - 2: storage media
  - 1: copy stored offsite
  - 1: offline immutable
  - 0: errors/backup verification

5. Publish/share
- FAIR (Easy to find, Accessible, Interoperable, Reusable)

6. Reuse
- What do do with data after
- Who is responsible
- Does data need to be cleaned

Personal Data
- Information linked to a physical person identified or identifiable
- Direct identification (name, last name)
- Indirect identification (identification nb., phone, biometric data, physical info, ip, …) data)

Sensible Data Political opinions/religious/philo., health/biometric, sex, military, pregnancy, condamnations

Non-sensible Data Name/last name, birth date, address, banking data

Data Processing Any type of data handling

Anonymise Modify in an irreversible way

Pseudo-anonymise Break the link between the person and their data without additional information

Préposé fédéral protection données transparence (PFPDT)
- Monitor data processing from federal admission, state companies and private firms
- Advises them for projects
- Has the names of the "conseillers à la protection des données" (Portail DPO/conseiller DPO)
- Investigates whether companies follow the rules
- Receives data breaches announcements

Registre des traitements (LPD)
1. qui est le responsable
2. quel est le but du traitement
3. qui sont les personnes concernées et les données personnelles traitées
4. qui sont les destinataires
5. combien de temps seront conservées les données
6. quelles mesures ont été prises pour garantir la sécurité des données
7. si transmission à l'étranger : dans quel état et quelles garanties il offre

Annonces de violation de sécurité
- Violation of high risk for the personality or fundamental rights of the concerned people
- Announce to PFPDT
- Concerned people need to be averted if it is necessary for their protection

Legal Basis
- Application of GDPR: CEPD ou EDPB (comité Européen de la Protection des Données)
- French speaking autorities
  - France: CNIL (commission nationale informatique et libertés)
  - Belgique: CPVP
  - Luxenbourg: CNDP
  - Suisse: le Préposé fournit un guide GDPR

GDPR conformity
1. Create a registry of our data processing
   - written medium
   - communicated to whoever asked if public sector, else not obligated
   - Tenir 2 registres (recommandation de la CNIL)
   - Forme:
     1. Objectif
     2. Data category
     3. Who has access
     3. How long are they processed, then archived
2. Sort our data
   - What data does the company really need?
   - Is sensible data processed, are they allowed?
3. Respect personal rights
   1. droit d'accès
   2. de rectification
   3. d'opposition
   4. d'effacement
   5. droit à la portabilité
   6. droit à la limitation du traitement
   - Transparency:
     - Why we collect
     - Why we can process
     - Who has access
     - How long are they conserved
     - How can we exert our right
     - What is the legal basis
4. Secure our data
   - Dépendent de la sensibilité des données et des risques pour les personnes

Analyse Impact Protection Données (AIPD)
- AIPD = PIA, Privacy Impact Assessment
- **Before** collecting data
- Why: identify risks

At least 2 of the following points:

1. Évaluation d'aspects personnels ou notation d'une personne (exemple : score financier)
2. Exclusion du bénéfice d'un droit, d'un service ou contrat (exemple : liste noire)
3. Prise de décision automatisée
4. Surveillance systématique de personnes (exemple : télésurveillance)
5. Traitement de données sensibles (exemples : santé, biométrie)
6. Traitement de données concernant des personnes vulnérables (exemple : mineurs)
7. Traitement à grande échelle de données personnelles
8. Croisements d'ensembles de données
9. Usages innovants ou nouvelles technologies (exemple : objet connecté)

High risk -> consult PFPDT or conseiller DPO

- **DSA** = Digital Services Act (août 2023)
  - obligations pour combattre la désinformation et la haine en ligne, protéger la liberté d'expression, limiter les recommandations, etc.
- **DMA** = Digital Markets Act (mars 2024)
  - assurer une concurrence équitable, ne pas enfermer les utilisateurs dans une plateforme
  - interdiction de croiser sans consentement des données collectées à travers différentes plateformes d'une même société pour le ciblage publicitaire
- Ce sont deux nouvelles législations européennes pour protéger les citoyens
  - concernent 20 à 25 très grandes entreprises
    - réseaux sociaux, places de marché, moteurs de recherche (inclus GAFAM)
    - liste évolutive selon le nombre d'utilisateurs dans l'UE (> 45 millions)

Formulaire de Collecte
Les informations recueillies sur ce formulaire sont enregistrées dans un fichier informatisé par [**identité et coordonnées du responsable de traitement**] pour [**finalités du traitement**]. La base légale du traitement est [**base légale du traitement**].
Les données collectées seront communiquées aux seuls destinataires suivants : [**destinataires des données**].
Les données sont conservées pendant [**« durée de conservation des données prévue » ou « critères permettant de la déterminer »**].
Vous pouvez accéder aux données vous concernant, les rectifier, demander leur effacement ou exercer votre droit à la limitation du traitement de vos données. **(en fonction de la base légale, mentionner : « Vous pouvez retirer à tout moment votre consentement au traitement de vos données » ; « Vous pouvez également vous opposer au traitement de vos données » ; « Vous pouvez également exercer votre droit à la portabilité de vos données »)**
Consultez le site cnil.fr pour plus d'informations sur vos droits.
Pour exercer ces droits ou pour toute question sur le traitement de vos données dans ce dispositif, vous pouvez contacter **(le cas échéant, « notre délégué à la protection des données » ou « le service chargé de l'exercice de ces droits ») :** [**adresse électronique, postale, coordonnées téléphoniques, etc.**]

Métadonnées Données à propos de données (1ère apparition: 1968, Philip Bagley)

3 types de métadonnées
1. **Descriptives**: décrivent ressource pour identification/recherche (titre, auteur, sujet, mots-clés)
2. **Structurelles**: composantes et liens (structure BD: tables, colonnes, clés, indexes)

3. **Administratives**: infos techniques (version, date archivage, droits, GDPR compliance, infos fichiers)

Utilité métadonnées
- Aider utilisateurs à trouver/découvrir ressources selon critères
- Organisation ressources numériques
- Identification numérique et archivage
- Analyse trafic réseaux

Exemples métadonnées But donnée, date/heure création, créateur/auteur, emplacement, taille fichier, qualité, source, moyen de création

Métadonnées photos numériques
- **Exif**: CIPA/JEITA (appareil, exposition, vitesse, focale, balance blancs)
- **IPTC**: International Press Telecom Council (infos presse)
- **XMP**: Adobe/ISO (extensible metadata platform)
- **PLUS**: Picture Licensing Universal System (licences)
- Géolocalisation GPS, tags/mots-clés, détection visages/objets

Stockage métadonnées
- **Interne**: dans même fichier que données
  - + garantit cohérence, facile à mettre à jour avec données
  - − difficile à analyser/indexer/cataloguer
- **Externe**: fichier séparé/BD
  - + recherche efficace, regroupement (repositories), traitement performant
  - − risque d'incohérences
- **Formats**: texte (XML, lisible, volumineux) vs binaire (compact, nécessite logiciel)

Dictionnaire des données / Metadata Repository
Nécessaire pour comprendre données numériques/structurées
- Ex: colonne 13 chiffres → ISBN; BD patient: longueur max nom, jeu caractères, obligatoire?, accès?
- Contenu: field name, table/DB, title, description, type/size, dimension (min/max), is-required/read-only, default, source, access, relationships

Registre de métadonnées (ISO 11179, 15000-3)
Géré par architecte des données, maintenu par data stewards
1. Sémantique de chaque élément (définitions)
2. Contraintes techniques (longueur, format, relations)
3. Informations administratives
- Systèmes: IBM InfoSphere, Informatica Data Catalog, Oracle MDS, SAS Metadata Server

Métadonnées Data Warehouses
- Définissent éléments insérés et relations
- Assurent cohérence données de sources variées
- **Common Warehouse Metamodel** (CWM): spécification échanges plateformes hétérogènes (IBM, Oracle, SAS, UBS)

Standards métadonnées
- **Darwin Core**: biologie (occurrences espèces, spécimens collections)
- **DOI**: Digital Object Identifier (ressources réseau, persistence, interopérabilité)
- **Dublin Core**: ressources réseau interopérables
- **EBUCore**: audiovisuel (broadcasters)
- **ISO 19115**: données géographiques
- **ISO/IEC 11179**: organisations (registre éléments données)
- **MARC**: bibliothèques (Machine Readable Cataloging)
- **MPEG-7**: multimédia (ISO/IEC descripteurs contenus)
- **RDF**: Resource Description Framework (ressources web)

DublinCore (DCMI, 1995) Standard interopérable pour métadonnées descriptives. 15 éléments facultatifs et répétables:
- **Contenu**: Title, Subject, Description, Source, Language, Relation, Coverage
- **Propriété intellectuelle**: Creator, Contributor, Publisher, Rights
- **Instanciation**: Date, Type, Format, Identifier
- Encodage: texte, HTML/XHTML (balises meta/link), XML, RDF
- **DC qualifié**: éléments supplémentaires (audience, provenance, détenteur droits)
- Utilisé par BnF pour Gallica (8M documents): augmente visibilité catalogues

---

Rôles clés gouvernance données Chief Data Officer | Data Owners | Data Stewards | Data Custodians | Comité GDD

Comité de gouvernance des données
- Accompagne déploiement, définit stratégie, approuve règles/standards
- Rôle intermédiaire direction/opérationnels
- Se réunit périodiquement (ex: mensuellement)
- **Responsable**: director of data governance, IT, BI, ou CIO/CTO
- **Membres**:
  1. Spécialistes IT (data warehousing, BI, architectes données)
  2. Responsables secteurs activité (sensibles qualité/partage données)
  3. Interface IT/business (gestionnaires, analystes)
  4. Direction (CIO, CFO) pour légitimité

Équipe de gouvernance des données Si équipe dédiée (opérationnel, ≠ comité décisionnel):
- **Manager**: conçoit, implémente, suit Master Data Management
- **Architecte**: supervise conception/implémentation mesures
- **Analyste**: identifie tendances
- **Spécialiste conformité**: vérifie respect normes légales

Chief Data Officer (CDO)
- Responsable gouvernance données, dirige comité (souvent)
- Fonction essentielle pour entreprises data science/big data
- Missions:
  - Mettre en œuvre stratégie data, homogénéiser actions, créer synergies
  - Valoriser actifs data (gains financiers, productivité, efficacité)
  - Contrôler état données (stockage, backups, niveaux sécurité)
  - Assurer respect lois (RGPD), faciliter décisions basées données
- ≠ DPO (CDO = dirigeant, DPO = indépendant)

Data Owner
- Prend décisions sur les données
- Spécifie exigences format/qualité
- Souvent cadres supérieurs divisions:
  - Responsable approvisionnement = owner données fournisseurs
  - Responsable RH = owner données employés

Data Steward
- **Technique/Custodian**:
  - Modèles, cycles de vie, maintenance, mise à jour, intégrité
  - Connaissance technique SI
  - Homogénéité technique (outils, architecture, traitements), éviter silos
- **Fonctionnel** (producteurs/consommateurs):
  - Soutient départements/domaines métiers (client, fournisseur, produit)
  - Définit principes création, mesure qualité, documente dans catalogue

Data Protection Officer (DPO) Conseiller à la protection des données (Suisse: LPD, Europe: RGPD)

Désignation obligatoire si: a) Organisation publique (sauf judiciaire) b) Suivi régulier/systématique grande échelle (activité base): marketing personnalisé, fidélité, pub comportementale, profilage/notation, géoloc, dispositifs portables/connectés c) Traitement grande échelle données sensibles (activité base)
- **Recommandé** autres organisations. Peut être interne/externe, mutualisé

Qualifications DPO (art. 37.5 RGPD):
- Indépendant, éviter conflit d'intérêts
- ≠ directeur général/opérationnel/financier/SI/RH (ne décide pas finalités/moyens)
- Expert législation/pratiques protection données, connaissance SI
- Rapporter haut niveau, animer équipe, communiquer efficacement
- Profils: 28% IT, 28% juridique, reste: admin/finance/audit

Missions DPO (couvrir ensemble traitements):
- Informer/conseiller responsable traitement et employés
- Contrôler respect règlementation (LPD/RGPD)
- Contrôler (parfois tenir) registres activités traitement
- Conseiller/vérifier AIPD
- Contact/coopération avec autorité contrôle (PFPDT, CNIL)
- Répondre aux personnes concernées
- ⇒ **DPO pas responsable** en cas non-conformité organisation

Moyens pour DPO:
- Communication désignation, implication questions protection, accès données/traitements
- Ressources: temps, finances, équipe. Indépendance: position, pas sanctions

---

Activités gestion données Architecture, Modélisation/design, Stockage, Sécurité, Interopérabilité, Documents, MDM, BI, Métadonnées, Assurance qualité

Métiers données (13 rôles)
1. **Data Analyst**: collecte/nettoie/agrège (SQL), visualisations, présente
2. **Database Administrator**: performance, disponibilité, sécurité BD
3. **Data Modeler**: modèles partagés, perspective d'ensemble
4. **Software Engineer**: systèmes logiciels, implémente spécifications
5. **Data Engineer**: pipelines sources → warehouse, nettoyage/qualité
6. **Data Architect**: optimise pipelines flux, transforme pour insights
7. **Statistician**: hypothèses/tests, reporting quantitatif
8. **BI Analyst**: besoins business, rapports dirigeants
9. **Marketing Scientist**: data scientist marketing/publicité/clients
10. **Business Analyst**: spécifications systèmes/reporting, interactions BD
11. **Quantitative Analyst**: modèles complexes, flux, intégrité
12. **Data Scientist**: extrait/transforme, prédictions, insights, décisions données
13. **ML Engineer**: systèmes ML production, pipelines

Compétences Data Scientist Maths, stats, ML, prog., visu, comm. **Faculté centrale**: comprendre business (3+ compétences: IT, maths, business)

Remarques métiers
- ⇒ Titres imprécis, charges similaires
- "Sexiest job 21st century" (HBR) mais compliqué: rapprocher métier, impacts business
- Valeur: maîtrise chaînes traitement (pas nouveaux algo). Préparation difficile automatiser

Cycle de vie des données
1. **Acquisition**: saisie manuelle, senseurs, achat/open data
2. **Prétraitement**: qualité, origine (*lineage*), confidentialité
3. **Stockage**: *data warehouse/mart/lake*, chiffrement, sauvegarde
4. **Utilisation**: consultation, visualisation, analyse, BI
5. **Archivage**: conservation (pas traitement), accès rares
6. **Destruction**: suppression totale, conformité RGPD

Data Management Plan (DMP) Document évolutif définissant gestion, description, stockage, protection données
Obligatoire: projets recherche (SNSF, UE, USA)
Utile: structurer GOD dans toute organisation

Éléments principaux
1. Spécifier données collectées (volume, type, sources, contraintes)
2. Organisation et gestion (outils, conformité)
3. Stratégie sauvegarde (risques, conformité, effacement)
4. Politique accès/partage (droits, licence, RGPD)
5. Responsabilités (qui responsable, rôles)

Projets *data-driven* Principe: commencer par questions métier, pas par technologie
- projet **métier**, pas purement IT
- résultats souvent pas à court terme

Mise en place: ateliers idéation avec utilisateurs métier
- questions simples, petite échelle
- faire émerger besoins depuis métier
- "*Si quelqu'un vous donnait le résultat, qu'en feriez-vous?*"

7 étapes projet basé données
1. **Poser problème**: question business, objectif, niveau minimal requis
2. **Obtenir données**: internes > partenariat > achat/open data/scraping
3. **Explorer données**: analyse exploratoire, visualisation, sémantique métier
4. **Transformer**: nettoyage (*outliers*, valeurs manquantes), nouveaux attributs
5. **ML**: modèles (simple → complexe), validation croisée, optimisation
6. **Présenter**: visualisations, résultats actionnables, implications métier
7. **Production**: implémentation, monitoring, réentraînements périodiques

Principes à suivre:
- écrire fonctions réutilisables
- historiques traitements
- gestionnaire versions (Git, DVC)
- documenter dès définition

ML statistique vs Deep Learning

	ML stat.	**DL**
Volume données	milliers	millions+
Fonctionnement	dirigé par analyst	dirigé par donnée
Attributs	sélection manuelle	représentation auto

Deep learning: bonne représentation auto, mais beaucoup de données nécessaires

---

CRISP-DM *Cross-Industry Standard Process for Data Mining* - méthodologie standard pour projets data mining (6 phases: business understanding, data understanding, data preparation, modeling, evaluation, deployment)

Biais liés aux mégadonnées 8 niveaux de biais:
1. Sociétaux: données reflètent biais existants
2. Mesure: traits non pertinents pour propriété visée
3. Représentativité: données mal échantillonnées
4. Étiquetage: erreurs d'annotation
5. Algorithmiques: modèle capture mal la propriété
6. Évaluation: test/métriques déséquilibrés
7. Déploiement: environnement différent
8. Feedback: résultats influencent nouvelles données

Biais vecteurs de mots (*embeddings*)
- Réseaux neurones prédisent mot suivant → vecteurs basse dimension (300)
- Modèles pré-entraînés: word2vec, GloVe, fastText
- **Problème**: apprennent stéréotypes (man:programmer :: woman:homemaker)

Solutions:
1. Réentraîner sur données non-biaisées
2. Débiaiser vecteurs par calcul
3. Ajouter méta-information (genre) pour traduction

IA générative et toxicité
Risques éthiques:
- langage toxique/grossier
- discrimination genre/race/groupe social
- divulgation données personnelles
- tromper utilisateurs

Nettoyer les donnée a un coût énergétique

Solutions alignement:
1. *Adversarial testing*: red team interne/externe
2. Modèle préférences humaines (RLHF)
3. Algorithmes: PPO, DPO, *Rejection Sampling*
4. Filtrage lors production

Exemple Tay (Microsoft 2016): chatbot désactivé après 16h, reflet tweets reçus

Biais vision par ordinateur Problèmes:
- Reconnaissance faciale: + erreurs sur personnes couleur/asiatiques
- Détection genre: 65% précision femmes couleur vs 99% hommes blancs
- Étiquetage sexiste: femmes → sourire/menton, hommes → officiel
- Objets biaisés: ordinateur présent → "homme" même si femme

Causes: représentation insuffisante groupes, stéréotypes banques images

Consommation énergétique données Data centers:
- 200-500 TWh/an (1-2% électricité mondiale)
- 5% émissions GES
- Consommation: serveurs + refroidissement (égalité), puis disques + réseau
- 13% consommation électrique mondiale en 2030?

Big data & IA:
- Temps calcul doublé tous 3.5 mois
- AlexNet → AlphaGo Zero: calculs ×300′000 en 6 ans

Modèles langage géants Coûts entraînement:
- 1 modèle (120h GPU): \$52-175 cloud, \$5 électricité
- 4789 modèles: \$103k-350k cloud, \$9870 électricité

	Llama-2	**GPT-4**
Paramètres	70 mds	220 mds (est.)
Tokens	2×10^{12}	13×10^{12}
petaFLOPS-jour	10′000	200′000
GPUs A100	2000 (1 mois)	20′000 (2 mois)

Équivalences 1000 petaFLOPS-jour:
- 500 MWh électricité
- 100′000 frs
- 10^{12} mots = 15M livres (Bibliothèque Nationale France)

Cerebras Wafer Scale Engine CS-2 (2021):
- 850′000 cœurs IA
- 2.6 billions transistors
- 40 Go SRAM sur puce
- 20 kW consommation (40 ménages suisses)
- Prix: \$60k/semaine (\$1.65M/an)
- Équivalent 250 GPU parallélisés