Rapide aperçu des grands thèmes abordés à l'école thématique E-ENVIR 2019

***

Sommaire

***

Document produit à l'occasion de la deuxième édition de l'école thématique E-ENVIR 2021. Il vise à présenter brièvement le contenu des cours afin d'avoir une navigation facilitée dans les ressources disponibles au lien suivant : https://e-envir.sciencesconf.org/

Document rédigé par V. Girard - Mai 2021

Cadre juridique de la science ouverte

L'émergence de la Science ouverte a pour conséquence l'ouverture des produits de la recherche publique (base de données - BDD, cartographies, logiciels et écrits scientifiques), à la fois dans un effort de reproductibilité de la science et d'un accès aux informations (au sens de données et publications) par un large public.

Avec l'émergence de la Science ouverte, la législation a progressivement évolué laissant apparaître avec la loi pour une république numérique en 2016 une trentaine de codes relatifs à l'accès et l'usage libres et gratuits des produits de la recherche. Toutes les données au sens large ne sont pas automatiquement mises à disposition ; elles répondent également au principe « aussi ouvert que possible, aussi fermé que nécessaire » avec en dernière date la modification de la loi « informatique et libertés » de 2018 faisant référence notamment au RGPD. Les données relatives à la protection de l'intérêt de l'Etat, des droits de la personne humaine et des biens peuvent également bénéficier d'exceptions à l'ouverture.

Pour les BDD de références nationales, les producteurs de données s'engagent à fournir des données de haut niveau de qualité, alors que les utilisateurs participent à cette montée de qualité de la donnée (signalement des erreurs, etc.) via la production de services sur la base de ces données.

Les licences de diffusion - le modèle français « Etalab » ou le modèle international « Créative Commons » - conditionnent les modalités d'usages et sont à étudier avec attention.

Consulter la présentation de Stéphanie Rennes de INRAE, « Produire, utiliser, et diffuser les données en sciences ouvertes : présentation du cadre juridique »

Cycle de vie des données / Principe FAIR / Plan Gestion des Données

Le cycle de vie des données est l'ensemble des étapes de gestion, conservation et diffusion des données de recherche. Une bonne appréhension du cycle de la vie des données, notamment à travers un PGD, permet de gagner du temps (et parfois de l'argent) et en particulier dans les projets collaboratifs, au regard de l'accroissement de la quantité des données, d'éviter la perte de données uniques et faciliter la reproductibilité, etc.

Le principe FAIR (*Findable, Accessible, Interopérable, Reusable*) permet de garantir une utilisation optimale des données et métadonnées à la fois par les hommes et les machines, et s'applique tout au long du cycle de vie de la donnée.

Le PGD est un document évolutif établi en début de projet, impliquant non seulement les collaborateurs de niveau 1 (thématiciens) que les collaborateurs de niveau 2 (juriste, archiviste, éditeur…). Le PGD mentionne les besoins matériels et humains, les responsabilités, etc. L'outil DMP-OPIDOR permet une entrée simplifiée à différents modèles institutionnels, avec des aides à chaque étape, et se définit par produits de recherche.

Consulter la présentation de Yvette Lafosse et Coralie Wysoczynski du CNRS-INIST, « Introduction au cycle de vie des données, aux principes FAIR et au Plan de Gestion des Données ou DMP »

Consulter l'outil DMP-OPIDOR : https://dmp.opidor.fr/plans

Métadonnées

Évaluer la qualité des données avec OpenRefine

La prise en compte de la qualité des données est assez facile à percevoir dès qu'on évalue le coût de la non-qualité qui augmente au fil de la distribution/du partage de la donnée. L'évaluation de la qualité de la donnée consiste en général à faire disparaître les erreurs communes (doublons, incohérences, valeurs manquantes, …). Parmi les **critères de qualité**, on retrouve la disponibilité des données (accessibilité, trouvabilité), la cohérence des données, leur traçabilité, la sécurisation, l'exhaustivité, la fraîcheur. Il convient alors de bâtir avec ses collaborateurs une véritable stratégie pour assurer la qualité des données (unicité des données, gestion des droits d'accès, choix de référentiels…) et parfois utiliser les normes existantes (ISO 19115 pour les métadonnées géographiques, ISO 8601 pour la date et heure, etc.).

Consulter la présentation de Chloé Martin du CNRS-BBEES, « Évaluer la qualité de ses données, exemple d'utilisation d'OpenRefine ».

Consulter l'outil OpenRefine : https://openrefine.org/ ; https://fr.wikipedia.org/wiki/OpenRefine

Principaux standards pour les (méta)-données en environnement

Les données de l'environnement sont issues des observations (occurrence d'espèces, photos, biométrie…) d'analyses (et protocoles), de capteurs (in situ, télédétection) et de modèle/simulation, et diffusées dans des formats clés (tabulaire : csv, SQL, NetCDF, raster ou vecteur ; au format darwin core…). Les métadonnées sont soit encapsulées dans la donnée (NetCDF, JPEG), soit dans un fichier à part (ex. le fichier xml d'un shapefile), et relève de différents niveaux d'explicitation (métadonnées de découvertes, de visualisation, pour l'accès et la requête de données, etc.).

Pour découvrir les jeux de données, les grands catalogues tels que Google (data search), Zenodo, GBIF…peuvent être consultés. Quant aux plus petits catalogues, ils sont moissonnés par les grands catalogues. Le recueil des métadonnées dans ces catalogues repose sur le choix de standards et normes. Notamment, les principaux standards de métadonnées à considérer pour s'assurer de l'interopérabilité des métadonnées produites et éviter toute re-saisie ; on parle de mapping entre les normes. Et de la même manière, l'utilisation de standard pour les métadonnées s'appuie sur l'interopérabilité sémantique (usage des vocabulaires contrôlés). Les datapaper sont des métadonnées littéraires permettant d'afficher une publication et un doi (important pour les carrières du personnel de la recherche).

Consulter la présentation de Julien Barde de l'IRD, « Les principaux standards de métadonnées à l'échelon international : comment s'y retrouver et comment exploiter pour la recherche en sciences de l'environnement ».

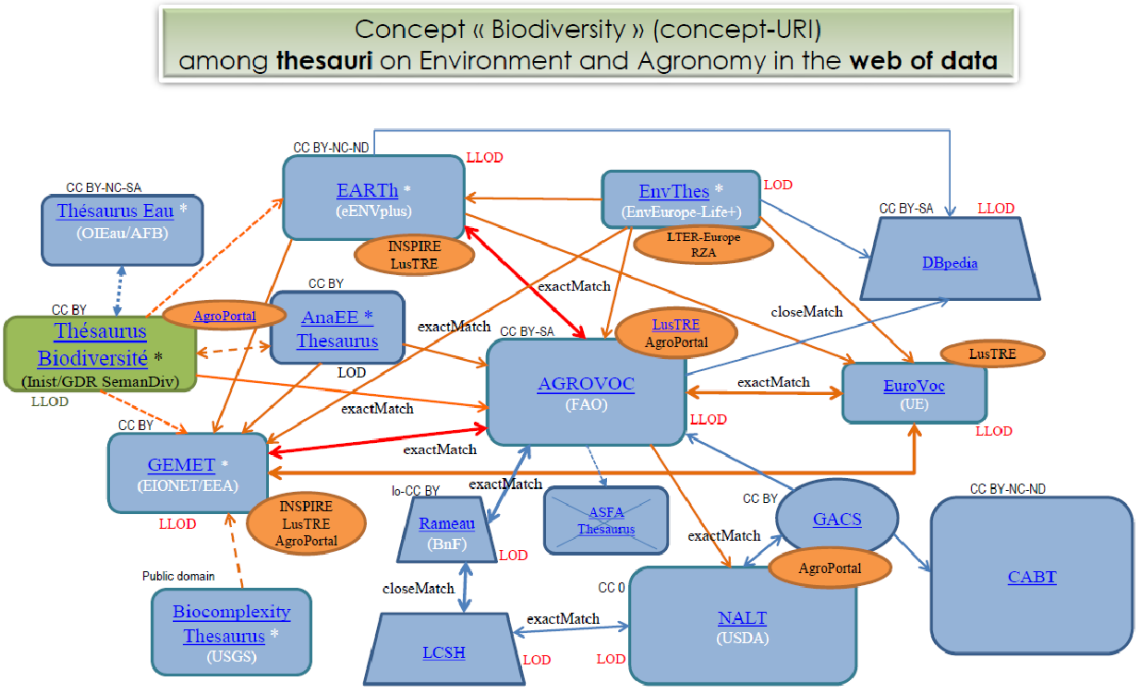
Consulter la présentation de Eric Garnier du CNRS-CEFE, « Intérêt et apports scientifiques de la synthèse de données : importance des ressources terminologiques », pour un exemple sur la nécessité de construire une base de données harmonisées sur les traits des plantes

Thesaurus

Un thesaurus est un catalogue de concepts, donnant accès à des identifiants uniques, des définitions, des traductions ou équivalence dans d'autres langues, une hiérarchisation des termes.

L'INIST a travaillé sur un recensement des vocabulaires contrôlés dédiés à des disciplines des sciences de l'environnement (biodiversité notamment), notamment pour identifier les atouts et inconvénients de chacun, relativement aux questions de standards, d'alignement avec d'autres thesaurus et de dynamique d'actualisation (outils de gestion ; ex. vocbench, openthseo, skomos, etc.).

Des portails d'accès aux ressources sémantiques (ex. AgroPortal) sont également des entrées potentielles pour identifier le vocabulaire contrôlé partagé.



Un exemple d'alignement du concept « biodiversité" à travers différents thesaurus.

Consulter la présentation de Dominique Vachez CNRS-INIST, « Panorama des thésaurus existants dans le domaine des sciences de l'environnement » et « Technologie pour la gestion des thésaurus et de leur interopérabilité : standards, outils de gestion et alignement ».

Web sémantique

Le web 3.0 semble laisser la place à une autonomie grandissante des requêtes et échanges d'information entre machines, notamment grâce au web sémantique ou encore nommé *web of linked data*. Le web sémantique permet de donner du sens aux informations en utilisant des représentations formelles (structurées) et normalisées des connaissances (ontologies). Le web sémantique repose sur 3 principes : l'utilisation d'un URI (Uniform Resource Identifiers), de HTTP URIs et de standards (RDF, SPARQL…) … A titre d'exemple, AnaEE a fait le choix de mobiliser des technologies du web sémantique pour la gestion et l'exploitation de la connaissance sur les données, essentiellement par les machines.

Consulter la présentation de Danielle Ziebelin de l'UGA-LIG, « Introduction au web sémantique », la présentation de Véronique Chaffard de l'IRD-IGE et Isabelle Braud INRAE « Projet de portail de données THEIA-OZCAR », et la présentation de Christian Pichot de

INRAE « Gestion et valorisation sémantiques de données de biodiversité et d'études d'écosystèmes dans l'infrastructures ANAEE-France ».

Publier et accéder aux données

Entrepôt des données

Les entrepôts de données sont des services en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion des jeux de données. Il en existe des disciplinaires, des institutionnels et des généralistes (ouvert à toutes disciplines). Ces entrepôts interagissent avec d'autres dispositifs numériques tels que les plateformes de données, les archives de données, et les annuaires de données (intégrateurs). Ils offrent ainsi une meilleure visibilité des données.

Consulter la présentation de Damien Boulanger de l'IRD "Les entrepôts de données : Ou comment rendre les données trouvables, accessibles et réutilisables ?", présentant notamment le choix de l'IRD de mettre en place son propre entrepôt.