Utilisation du standard DCAT à l'Insee

Guillaume Duffes
Institut national de la statistique et des études économiques (Insee)

Résumé

L'Institut national de la statistique et des études économiques (Insee) joue un rôle crucial en collectant, produisant, analysant et diffusant de nombreuses informations sur l'économie et la société françaises et doit assurer la visibilité et la découvrabilité de ses données. L'Insee a recours au vocabulaire DCAT pour ses jeux de données qui constituent le point d'entrée principal des résultats statistiques, Ainsi l'Insee s'est doté de son propre outil de gestion de ces objets DCAT, l'application Bauhaus, publiée en open source.

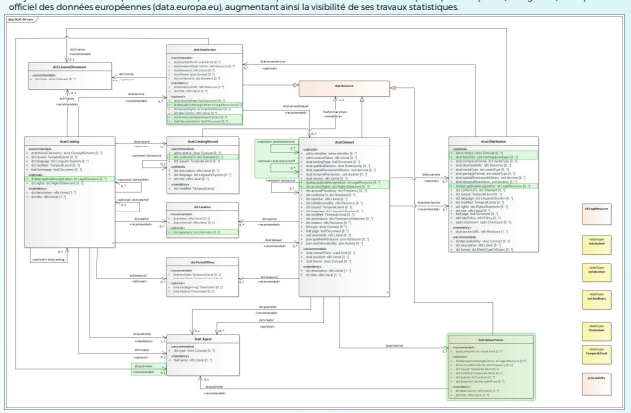
L'Insee, diffuseur de données

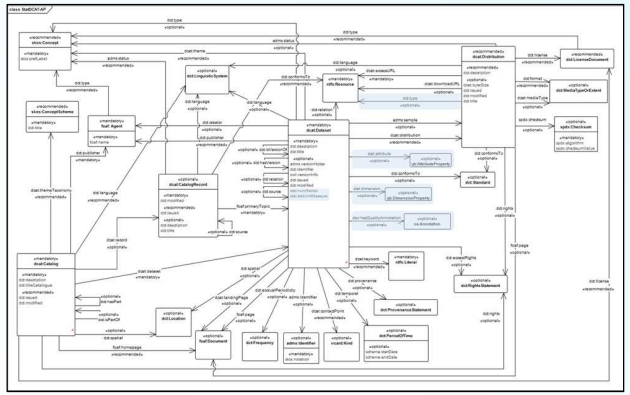
Tous les ans, l'Insee organise et pilote un nombre important d'enquêtes dont la plus célèbre est le recensement de la population menée avec les communes. Celle-ci permet de connaître les évolutions socio-démographiques à un niveau territorial fin pour, par exemple, évaluer les besoins en infrastructures qui en découlent. Réalisée sur toute l'année, l'enquête emploi vise, au-delà de son indicateur phare, le taux de chômage, à décrire le marché du travail et son évolution. En complément, d'autres enquêtes auprès des particuliers sur différents sujets (ressources, dépenses, logement, formations et compétences, patrimoine…) sont menées pour éclairer le débat public français et européen. Auprès des entreprises, l'Insee recueille des informations tant sur leur structure (taille, secteur, main-d'œuvre…) que sur leur activité (carnet de commandes, investissement…). Découvrez la représentation illustrée des sujets des enquêtes menées par l'Insee en 2023.

L'ensemble de ces enquêtes, mais aussi des sources externes fournies par d'autres administrations ou organismes privés donnent lieu à l'élaboration et la mise à disposition d'un nombre conséquent de jeux de données, plus de 450 pour les seuls actuellement diffusés sur les site Internet de l'Insee (les microdonnées entre autres ne sont mises à dispositions que de certains publics dans des conditions très strictes).

DCAT et les profils liés

L'Insee a recours aux vocabulaires RDF pour gérer une grande partie de ses métadonnées relatives aux processus métiers. Le jeu de données constituant le point d'entrée principal des résultats statistiques, il est donc essentiel que sa description soit facilement accessible et interopérable avec d'autres catalogues de données. Afin d'atteindre cet objectif, l'Insee a adopté le Data Catalog Vocabulary (DCAT), un standard du web sémantique publié par le W3C. En s'appuyant sur des classes permettant de définir un catalogue (Catalog), un jeu de données (Dataset) ou encore un fichier (Distribution, représentant un moyen d'accéder à un Dataset, par exemple via un fichier CSV téléchargeable) et sur des extensions du standard DCAT, dont le profil d'application DCAT-AP, et plus particulièrement son extension pour les données statistiques StatDCAT-AP, publiées par la Commission Européenne, l'Insee vise des premières publications compatibles avec ces standards d'ici fin 2024. Cela lui permettra de rendre ses jeux de données interopérables avec des portails comme la plateforme ouverte des données publiques françaises (data.gouv.fr) ou le portail officiel des données européennes (data.europa.eu), augmentant ainsi la visibilité de ses travaux statistiques.

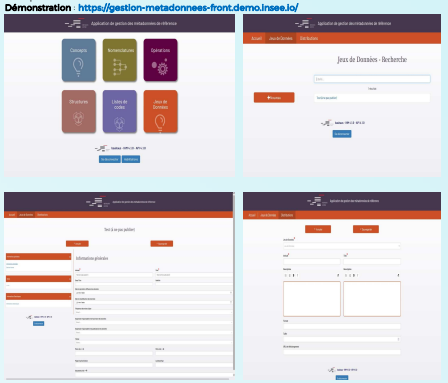




Bauhaus, outil de gestion des objets DCAT

Pour faciliter la transition vers l'interopérabilité avec les portails nationaux et européens, l'Insee s'est doté de son propre outil de gestion de métadonnées DCAT, l'application Bauhaus, publiée en open source. Cet outil a été conçu pour que les étapes de conception et de construction de ces métadonnées ne soient plus réservées aux experts du standard. Grâce à Bauhaus, l'Insee peut gérer efficacement ses métadonnées, rendant la documentation et l'accès aux jeux de données plus simples et plus accessibles pour une plus large catégorie d'agents. En intégrant ces technologies avancées, l'Insee renforce son engagement à améliorer l'accessibilité et la visibilité de ses données statistiques, répondant ainsi aux besoins croissants des utilisateurs en quête d'informations structurées, fiables et interopérables.

Démonstration : https://gestion-metadonnees-front.demo.insee.io/



Triplets RDF publiés

Exemple de triplets DCAT

Prochaines étapes

Le vocabulaire *DCAT* et ses profils d'application sont riches, seule une partie très minoritaire a été utilisée dans cette première mise en oeuvre.

- L'Insee utilise déjà des définitions de structures de données (DSD) telles que définies par le modèle d'information *SDMX (Statistical Data and Metadata Exchange)* et notamment dans leur version RDF exprimée par le vocabulaire *RDF Data Cube* (Recommandation W3C, préfixe *qb*). L'extension *statDCAT-AP* permet de lier des dimensions et des attributs d'un cube de donné à un dcat:Dataset grâce aux prédicats *stat:dimension* et *stat:attribute*.
- De même, les évaluations de la qualité des jeux données peuvent être enrichies grâce au *Data Quality Vocabulary (DQV)*. Les classes dcat:Dataset et dcat:Distribution y occupent une place centrale. Le profil d'application *statDCAT-AP* reprend même la propriété dqv:hasQualityAnnotation dans sa spécification. L'Insee publie des rapports qualité pour la plupart de ses opérations statistiques dont les composants pourraient donc être caractérisés et liés au dcat:Dataset correspondant.
- L'information géographique occupe une place centrale dans les données statistiques publiées par l'Insee. Le profil d'application geoDCAT-AP offre une syntaxe RDF pour les métadonnées de la directive européenne Inspire tout en étendant le profil d'application DCAT. Il pourrait donc s'agir d'ajouter les métadonnées géographiques essentielles dans de nombreux catalogues de jeux de données de l'Insee.