Sept étapes pour publier des données ouvertes et liées

Cécilia Fabry, Clotilde Roussel, Alain Collignon, Elise Moreau, François Parmentier, Nicolas Thouvenin

▶ **To cite this version:**

Cécilia Fabry, Clotilde Roussel, Alain Collignon, Elise Moreau, François Parmentier, et al.. Sept étapes pour publier des données ouvertes et liées. I2D – Information, données & documents, 2017, 1, pp.12-14. hal-01541517

HAL Id: hal-01541517
https://hal.science/hal-01541517v1

Submitted on 19 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Sept étapes pour publier des données ouvertes et liées

Toute personne chargée d'un projet de mise en ligne de données liées et ouvertes trouvera des repères méthodologiques dans ce guide rythmé par une alternance de propos théoriques et d'études de cas.

Les évolutions technologiques du Web ont permis de passer d'un web de documents (navigation hypertextuelle) au web de données (liens entre les données elles-mêmes, espace unifié). Ce passage a pu se faire grâce à l'avènement de différents standards issus du web sémantique comme RDF, OWL, SPARQL et URI[1]. Les bibliothèques se sont saisies de ces opportunités pour faire migrer leurs catalogues[2], comme l'indique, par exemple, data.bnf.fr[3]. Au-delà de ces établissements, des organismes publics[4] mettent en accès libre des jeux de données et des organismes de recherche[5] nourrissent la réflexion concernant la publication de données de recherche.

Le processus à suivre

Le processus qui suit est illustré par Linked Open Data Experiment (LODex)[6,] conçu sur l'archive Istex[7] (voir encadré) en construisant des lacis de données alignées et interopérables dans l'esprit et les standards du web sémantique (Linked Open Data – LOD).

Le projet Istex est un vaste programme d'acquisition de ressources scientifiques visant à créer une bibliothèque numérique aux meilleurs standards internationaux, accessible à distance par tous les membres des établissements de l'enseignement supérieur et de la recherche.

[1] http://fr.slideshare.net/AntidotNet/web-smantique-web-de-donnes-web-30-linked-data-quelques-repres-pour-sy-retrouver
[2] Emmanuelle Bermès, avec la collab. d'Antoine Isaac et Gautier Poupeau. *Le Web sémantique en bibliothèque*. Électre-Cercle de la Librairie, 2013 (coll. Bibliothèques)
[3] http://data.bnf.fr
[4] https://opendata.paris.fr/page/home/ ; https://www.data.gouv.fr/fr
[5] http://fr.slideshare.net/paventurier/les-donnes-de-la-recherche-enssib2013paventurier
[6] http://lodex.inist.fr/2016/03/presentation-de-lodex
[7] www.istex.fr

Fig. 1: Représentation schématique du processus intellectuel (Source : Inist-CNRS)

Étape 1 : Choix de la licence

Les données publiées sur le Web nécessitent de clarifier les conditions et possibilités juridiques de leurs réutilisations. Il s'agit de garantir et sérier la sécurité juridique en déclarant une licence reconnue d'utilisation des jeux de données exposés. L'usage conventionnel de publication sur le Web recommande de retenir une licence ouverte telles que la licence Etalab[8] ou les licences Creative Commons[9].

LODex : Les jeux de données traitées pour l'expérimentation sont issus de l'archive Istex. Elles héritent donc du régime de la licence ouverte/open licence Etalab pour maintenir la cohérence globale d'accès et d'utilisation de ces ressources.

Étape 2 : Recueil des données (ou constitution du jeu)

Cette étape clé du processus est le fruit d'une réflexion préalable sur le choix des données à publier : sont-ce des données bibliographiques, des données issues d'un fichier d'autorité (auteurs, thésaurus, etc.), des données de recherche ? Quel est l'état de ces données ? etc. Deux facteurs importants sont à examiner : l'unicité et le fort potentiel à valeur ajoutée des données à recueillir, donc à publier.

LODex : L'originalité de cette expérimentation est de publier des données autres que des données bibliographiques, raison pour laquelle la collecte des données s'est déroulée en étroite collaboration avec les autres équipes Istex.

[8] https://www.etalab.gouv.fr/licence-ouverte-open-licence
[9] http://creativecommons.fr/licences

Étape 3 : Protocole de constitution de l'URI

À partir des données collectées en amont, l'URI se construit en prenant en compte deux éléments déterminants et signifiants :

- la définition du nom de domaine – localisation – autorité d'adressage (éviter toute formulation liée à l'obsolescence des termes qui le composent) ;
- la définition de l'identifiant permettant de repérer la ressource elle-même, quel que soit le système d'identifiant retenu (DOI, ARK[10], etc.). Il doit répondre à deux critères : unicité et pérennité.

LODex : Comme nom d'autorité d'adressage (ou localisation), nous avons retenu le nom du jeu de données en deux mots séparés par un tiret suivi de .lod.istex.fr. Pour la partie identifiant proprement dite, nous avons retenu le système ARK. En complément à l'autorité nommante, le nom ARK est constitué dans notre cas d'un *subpublisher* (sur 3 caractères) par jeu de données suivi d'une séquence alphanumérique (sans voyelle). Exemple: URI pour le jeu de données dédié aux entités nommées : http://named-entity.lod.istex.fr/ark:/67375/RXP-00000000-0

Étape 4 : Analyse des données

La première étape du processus ou curation de publication de données est d'analyser ces données. Cette étape doit se dérouler en étroite collaboration avec les producteurs ou experts de données, car ce travail d'éditorialisation et normalisation est à haute valeur ajoutée.

LODex : Cette étape consiste à analyser rigoureusement les données internes (format, quantité, qualité, etc.) comme, par exemple, le format (texte ou standard) des différentes cellules du tableau, l'encodage des caractères, etc.

Étape 5 : Enrichissement des données

Nous appelons enrichissement toute information pouvant apporter du sens à la donnée initiale, comme une traduction, une définition, un contexte d'utilisation, des données de gestion. Lors de cette étape, nous pouvons créer des liens avec d'autres ressources. Dans ce cas, nous parlons d'alignement avec des référentiels contrôlés contenant des concepts de même nature. Ce type de lien permet un rapprochement avec des entités similaires et par conséquent augmente la visibilité du jeu de données sur le Web en multipliant les points d'entrée.

LODex : Après avoir enrichi les données initiales, nous avons réalisé un alignement manuel avec des ressources extérieures telles que la Classification décimale Universelle (CDU) et le thésaurus MeSH (*Medical Subject Headings*). En parallèle, une étude pour mettre en place un outil d'alignement automatique est en cours. Pour ceci, un apprentissage est effectué sur un corpus particulier correspondant aux entités nommées issues d'Istex et notamment sur les noms géographiques (*place name).*

Étape 6 : Modélisation des données - Schéma conceptuel

Après avoir recueilli, analysé/contrôlé et enrichi les données, l'étape suivante consiste à organiser ces dernières dans un modèle de données afin de garantir une interopérabilité sémantique. Construire un modèle de données pour le jeu de données signifie que l'on sélectionne des classes et des propriétés idoines dans des ontologies existantes. Pour cela, il faut sélectionner des vocabulaires du web sémantique recensés dans le Linked Open Vocabulary - LOV[11].

[10] https://wiki.ucop.edu/display/Curation/ARK
[11] https://lov.okfn.org/dataset/lov

LODex : Pour publier les jeux de données, deux ontologies ont surtout été utilisées : DCMI Metadata Terms (dcterms) et Skos (Simple Knowledge Organization System). L'étape d'alignement privilégie les relations d'équivalence skos:closeMatch et skos:exactMatch. Exemple de cette modélisation (fig. 2).

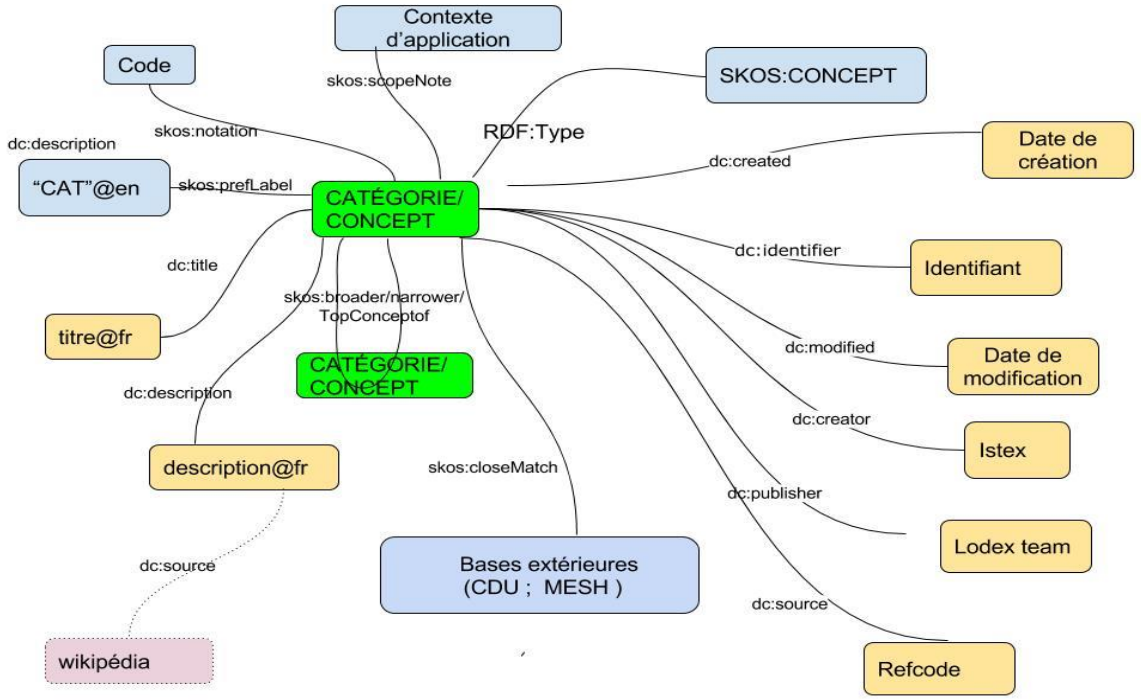

Fig. 2 : Exemple de modélisation du jeu de données « catégorie Inist » (Source : Inist-CNRS)

Étape 7 : Publication du jeu de données

Les 6 étapes précédentes effectuées avec succès, les jeux de données validés peuvent être publiés sur le Web, dans le respect des standards du web sémantique, sous une licence ouverte. Cette étape de publication arrive en bout de chaîne, comme un autre format de sortie qui permet de rendre accessibles et réutilisables vos données par les internautes et, en particulier, les communautés scientifiques.

LODex : Cinq jeux de données ont été publiés au cours de cette expérimentation. L'installation d'un *triplestore* permettant les requêtes Sparql est prévue dans la suite de cette expérimentation.

Perspectives

Ce guide décrit un processus d'éditorialisation des données ouvertes et liées et ainsi invite à une publication contextualisée sur le Web. L'ouverture des données permet d'enrichir les ressources, de construire de nouvelles interfaces intelligentes[12], de mieux les documenter grâce aux ontologies. Par conséquent, elle ouvre les champs du possible vers de nouvelles coopérations en créant un nouveau *continuum*.■

Cécilia Fabry cecilia.fabry@inist.fr, Clotilde Roussel clotilde.roussel@inist.fr, Alain Collignon alain.collignon@inist.fr
Ingénieurs documentalistes Inist-CNRS

Elise Moreau elise.moreau@inist.fr, François Parmentier francois.parmentier@inist.fr, Nicolas Thouvenin nicolas.thouvenin@inist.fr
Ingénieurs informaticiens Inist-CNRS

[12] E. Dzalé, K. Yeumo, S. Aubin, C. Mader, P. Aventurier, S. Cocaud. « Publication en Linked Open Data de données expérimentales sur la chenille processionnaire du pin ». *In* : Atelier IN-OVIVE (INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement), Lille, 02-03 juillet 2013
http://prodinra.inra.fr/record/195427