HAL
open science

LODEX : des données structurées au web sémantique

Stéphanie Gregorio, Alain Collignon, François Parmentier, Nicolas Thouvenin

To cite this version:

Stéphanie Gregorio, Alain Collignon, François Parmentier, Nicolas Thouvenin. LODEX : des données structurées au web sémantique. Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019), Jan 2019, Metz, France. hal-01990444

HAL Id: hal-01990444

https://hal.science/hal-01990444

Submitted on 23 Jan 2019

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.

LODEX : des données structurées au web sémantique

Stéphanie Gregorio, Alain Collignon
François Parmentier, Nicolas Thouvenin

Inist-CNRS, 2, Allée du Parc de Brabois, CS 10310, 54519 Vandœuvre-lès-Nancy
prenom.nom@inist.fr, https://www.inist.fr

Résumé. Lodex est un logiciel *open source* dédié à la valorisation de données structurées. Il facilite la curation et la sémantisation de données brutes pour les connecter au web de données via les normes et les standards du web sémantique. Il propose, en plus de la création automatique d'uri, la génération d'identifiants pérennes normalisés via le système des ark.

1 Introduction

Les bibliothèques produisent depuis longtemps dans leurs catalogues des données structurées et contrôlées, qu'elles exposent sur le web. Le web sémantique est présenté comme étant le web pour lequel les ordinateurs interprètent les métadonnées afin de mieux assister l'utilisateur dans sa recherche de l'information (Berners-Lee et al., 2001). L'Inist-cnrs a lancé une expérimentation visant à publier, selon les normes du web sémantique, des données extraites du fonds istex [1] (plus de 20 millions de publications scientifiques). Cette expérience a eu comme incidence le développement de lodex, outil permettant de mettre en ligne des jeux de données dans le respect des normes et standards du w3c.

Dans cet article nous présenterons brièvement l'archive istex puis nous développerons l'outil lodex qui a pour but de publier des données extraites de cette archive et ainsi faciliter l'accès et la diffusion des données acquises et produites. Cette publication est réalisée via un site dédié (https://data.istex.fr/) et un SPARQL endpoint (https://data.istex.fr/sparql/) contenant un graphe global des données istex.

2 L'archive ISTEX

Le projet istex a pour objectif de permettre à la communauté scientifique française d'accéder à une bibliothèque numérique pluridisciplinaire en texte intégral regroupant l'essentiel des publications scientifiques mondiales. Ce réservoir de publications scientifiques est bien entendu à destination des documentalistes et chercheurs ayant un besoin documentaire. C'est également une ressource unique pour tous les chercheurs gravitant autour des thématiques de la fouille de textes, du tal (Traitement Automatique de la

1. https://www.istex.fr

Langue), de la Recherche d'Information…La mise en ligne de ces informations en texte intégral structuré permet de développer des fonctionnalités d'extraction de connaissances basées sur les technologies de la fouille de textes.

Ces enjeux ont été un déclencheur pour proposer une documentation dynamique et interopérable du fonds ISTEX, et pour publier sous forme de jeux de données toutes les informations non présentes dans les documents. Ces derniers respectent les normes du web sémantique grâce à l'utilisation d'un outil dédié.

3 L'outil LODEX

3.1 Cadre de réflexion

Dans le contexte présenté ci-dessus nous avons identifié différents utilisateurs pouvant intervenir lors de ce processus de publication. L'internaute consulte les ressources sur la toile et peut prendre le rôle de *data consumer* lorsqu'il télécharge des informations. Le documentaliste *data manager* sélectionne, affine et publie des données en toute autonomie. L'informaticien ou le documentaliste joue la fonction d'administrateur *data administrator* du système.

Puis nous avons défini schématiquement un processus intellectuel de publication des jeux de données (Fabry et al., 2017). Pour l'établir, nous avons rapproché notre réalité de terrain avec les notions théoriques du web sémantique appliquées en milieu documentaire (Bermès et al., 2013). En particulier nous nous sommes penchés sur le caractère hétérogène des ressources et son incidence sur le protocole à mettre en œuvre.

Prenant en compte la typologie des utilisateurs ainsi que notre processus de publication, nous avons souhaité disposer d'un outil permettant de :

— publier selon des normes du web sémantique des tableaux comportant des données brutes,
— faciliter la transformation en données structurées,
— aider à aligner les données à publier avec des données similaires ou connexes,
— explorer le jeu de données publié pour valoriser et référencer chaque ressource.

Dans un environnement professionnel en pleine mutation, ayant vu naître de nouvelles activités dans les bibliothèques (ou centres de documentation), la curation, la modélisation, la normalisation, le modèle RDF sont au cœur des préoccupations des *data managers*. Ceci a eu pour incidence l'émergence d'outils dédiés à ces activités comme par exemple LODRefine et Catmandu (Harlow, 2015). Datalift(Scharffe et al., 2012) en est un autre exemple. Le concept *élévation des données*, permettant de passer d'un fichier tabulé à un fichier RDF nous a fortement séduits. Cependant, la fonctionnalité d'exposition des données sur le web était peu satisfaisante. Plus près de nos préoccupations, le logiciel CubicWeb dédié aux techniques du web sémantique est utilisé dans le développement de l'application `data.bnf.fr` (Le Bœuf, 2013). Le logiciel CubicWeb, présente de nombreuses fonctionnalités pouvant nous être utiles, cependant l'usage de ce *framework* nécessite l'appui technique de la société Logilab, par conséquent, nous nous sommes orientés vers le développement d'une solution logicielle libre appelée LODEX.

Par rapport aux outils similaires, cet outil se concentre sur trois priorités : masquer la complexité des triplets au format RDF, donner envie de structurer son information en augmentant les données (visualisation, interconnexion, *etc.*) et faciliter la mise à jour ou l'ajout d'information sans refaire un long processus de publication. LODEX a été développé avec des technologies JavaScript. C'est un logiciel libre dont le code source est accessible sur GitHub [2] et sous licence CeCILL.

2. `https://github.com/Inist-CNRS/lodex`

3.2 Le back office

Son *back office* permet de réaliser toutes les fonctionnalités nécessaires au traitement ou *stylage* d'un jeu de données.

Après avoir importé un jeu de données dans un des formats acceptés (`.csv`, `.tsv`, `.xml`, `.json`, …), l'outil propose six grandes étapes permettant le processus de publication :

1. Informations générales.
2. Comment la valeur est créée.
3. Transformations appliquées à cette valeur.
4. Sémantiques.
5. Comment et où elle est affichée.
6. Recherche.

Nous allons détailler les singularités de LODEX, sans nous attarder sur l'ensemble du processus qui sera développé lors de la démonstration du logiciel.

Suite à l'import d'un fichier, l'outil génère automatiquement un URI (*Uniform Resource Identifier*), identifiant requis pour le web sémantique. Par défaut, LODEX crée un `uid://` (*Unique Identifier*). Si votre organisation a opté pour le système d'identification ARK[3], l'URI se génère automatiquement en fonction de la présence des paramètres `naan` et `subpublisher` dans le fichier de configuration.

Une attention particulière a été portée à la fonctionnalité « Transformations appliquées à cette valeur » car elle donne la possibilité au *data manager* de réaliser une curation automatisée de ses données. L'outil LODEX propose différents *transformers* permettant de standardiser le contenu du jeu de données. Par exemple, LODEX permet de transformer la valeur du champ en un booléen, de remplacer une chaîne de caractères par une autre ou bien encore d'ajouter une chaîne de caractères à la fin de la valeur du champ…

L'étape 4 « Sémantiques », permet de renseigner la propriété ou prédicat des triplets (un triplet est composé de trois parties : sujet - prédicat - objet). La saisie y est facilitée par autocomplétion avec les différentes ontologies présentes dans le *Linked Open Vocabularies* (LOV). LODEX exporte les structures nécessitant des nœuds blancs en leur créant des identifiants uniques. Nous avons identifié deux cas :

1. Annoter un autre champ : par exemple pour préciser la source d'une définition.
2. Composer ce champ : au sens du web sémantique, composer ce champ à partir de plusieurs champs. Par exemple, une adresse est composée d'un nom de rue, d'une ville, d'un pays.

Après curation, *sémantisation*, le jeu de données est publié via le *front office* (dans notre cas `https://data.istex.fr/`). Différents exports aux formats du web sémantique sont possibles (Turtle pour sa lisibilité ; N-Quads et N-Triple pour leur simplicité et JSON pour son application courante dans le web). Ces exports permettent d'alimenter un *triplestore* (`https://data.istex.fr/sparql/`).

Une documentation permettant la prise en main de l'outil ainsi que son utilisation est accessible à l'adresse suivante `https://user-doc.lodex.inist.fr/`. Des tutoriels viendront la compléter.

3. `http://www.bnf.fr/fr/professionnels/issn_isbn_autres_numeros/a.ark.html`

4 Conclusion

L'objectif principal de notre approche est de mettre à disposition un outil intuitif afin de valoriser un jeu de données via le web de données ou *Linked Open Data*. L'outil LODEX qui présente la caractéristique de publier des tableaux bruts selon des normes du web sémantique révèle les particularités suivantes :

— faciliter la transformation de données structurées en données sémantisées,
— aider à aligner les données à publier avec des données similaires ou connexes,
— exposer le jeu de données pour valoriser et référencer chaque ressource.

Dans le nouveau paradigme de la science ouverte et plus particulièrement celui des données ouvertes, l'outil LODEX peut être un excellent allié afin de publier des données selon les principes FAIR [4].

Références

Bermès, E., A. Isaac, et G. Poupeau (2013). *Le Web Sémantique en bibliothèque*. Electre-Ed. du Cercle de la Librairie, Paris.

Berners-Lee, T., J. Hendler, et O. Lassila (2001). The semantic web. *Scientific American*, p. 29–37.

Fabry, C., C. Roussel, C. A., M. E., P. F., et T. N. (2017). Publier des données liées et ouvertes en sept étapes. *I2D - Information, données & documents 54*, p. 12–14.

Harlow, C. (2015). Data munging tools in preparation for RDF: Catmandu and LODRefine. *Code{4}lib Journal 30*, p. 1–12.

Le Bœuf, P. (2013). Customized OPACs on Semantic Web: the OpenCat prototype. In *IFLA Satellite Meeting*, Singapore.

Scharffe, F., L. Bihanic, G. Képéklian, G. Atemezing, R. Troncy, F. Cotton, F. Gandon, S. Villata, J. Euzenat, Z. Fan, B. Bucher, F. Hamdi, P.-Y. Vandenbussche, et B. Vatant (2012). Enabling linked data publication with the datalift platform. In *Proc. AAAI workshop on semantic cities*.

Summary

LODEX is an open source software dedicated to the valuation of structured data. It facilitates the curation and semantisation of raw data to connect them to the web of data via standards of the semantic web. It offers, in addition to the automatic creation of URIs, the generation of standardized perennial identifiers via the ARK system.

4. Findable, Accessible, Interoperable and Reusable