arbido

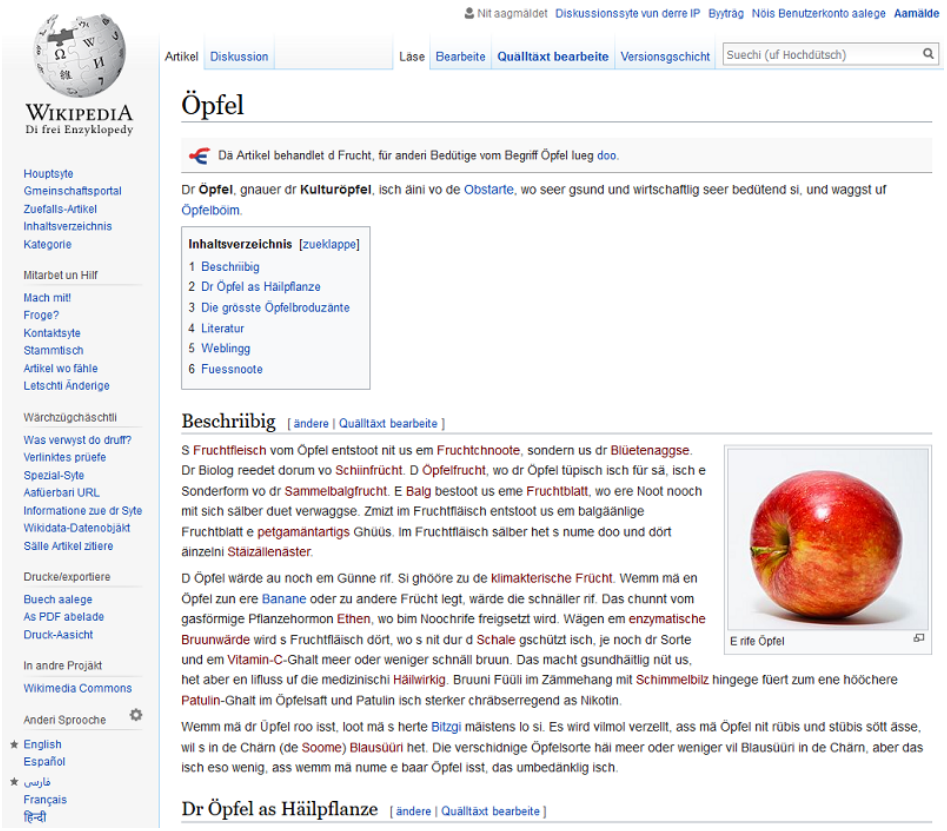
2017/3 Metadaten – Datenqualität

[Neko] [Rama], Wikimédien, photographe, dessinateur

Données structurées, la puissance de Wikidata au service de Wikimedia Commons

Parmi les projets de la Wikimedia Foundation, un nouveau projet cherche à uniformiser les métadonnées pour structurer la «jungle» de Wikimedia Commons de façon similaire à la base de connaissance structurée en web sémantique Wikidata.

Si tout le monde aujourd'hui utilise l'encyclopédie libre Wikipédia, il existe d'autres projets de la Wikimedia Foundation moins connus du grand public. Parmi ceux-ci, Wikimedia Commons, la médiathèque libre qui regroupe des images, sons et vidéos; et Wikidata, une base de connaissance structurée en web sémantique. Si ces projets ont une existence et une vie propre, leur fonction s'illustre intuitivement par les services qu'ils rendent à Wikipédia.

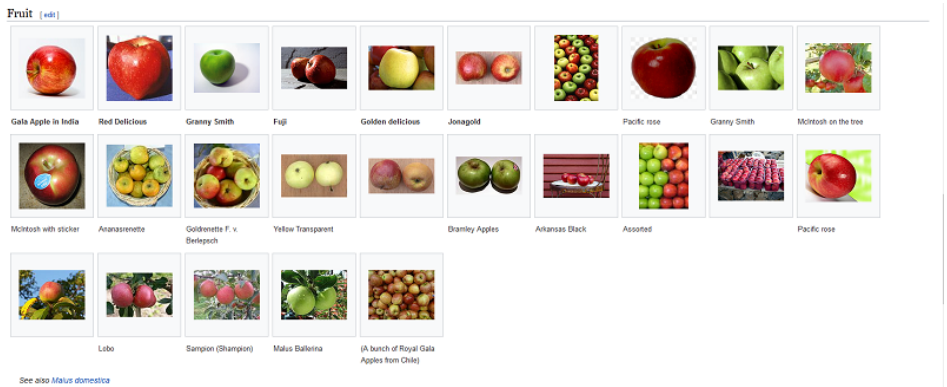


Un article Wikipédia utilisant des données de Wikidata (liste des langues) et de Commons (image) https://als.wikipedia.org/wiki/Apfel

De la jungle de Commons...

Commons constitue une archive unique où l'iconographie est disponible comme si elle était enregistrée en local sur la Wikipédia concernée: ainsi, il n'y a plus besoin de téléverser une même image sur les Wikipédia en différentes langues pour qu'elle s'affiche aussi bien en français et en allemand qu'en anglais. Cette centralisation permet aussi de mieux gérer les informations associées aux fichiers: licences, descriptions, etc. De là s'est constituée une communauté qui encourage la qualité des images par des concours et des labels décernés par les pairs, et la quantité avec des partenariats, des événements encourageant la photographie (Wiki Loves Monuments, Wiki Loves Earth, Wikicheese, etc.).

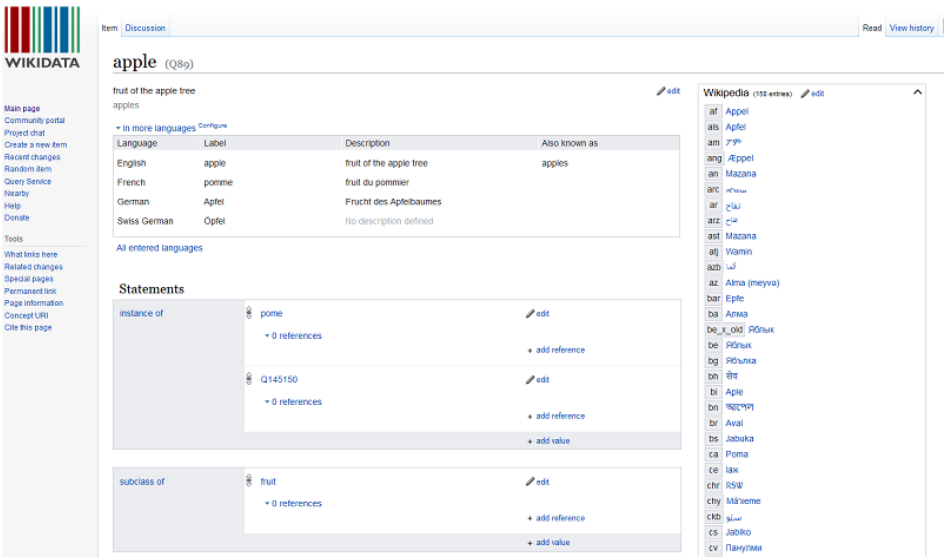
Interface de Wikimedia Commons pour la catégorie «Apples»
https://commons.wikimedia.org/wiki/Category:Apples



Sélection d'images (fruits) dans Wikimedia Commons
https://commons.wikimedia.org/wiki/Apples#Fruit

... à la structure de Wikidata

Wikidata est une base de connaissance, qui utilise le logiciel Wikibase pour stocker ses informations sous forme de triplets identifiant-propriété-valeur (par exemple, l'objet Q684661 « Jet d'eau de Genève» a une propriété «localisation» dont la valeur est «Genève»). À l'origine, elle visait à centraliser les «interwikis», c'est-à-dire les liens qui relient les articles de Wikipédia équivalents en différentes langues (par exemple fr:pomme est lié à en:apple et à als:Öpfel), tout comme Commons centralise les images. On s'est bientôt rendu compte que l'endroit où étaient rassemblés ces interwikis pouvait contenir aussi des informations: ainsi un objet numéroté Q89 a-t-il été créé, associé à un libellé en plusieurs langues («pomme» pour le français, «apple» pour l'anglais, «Apfel» pour l'allemand et ainsi de suite). Une quantité illimitée d'autres informations sous forme de triplets peut s'y associer. Par exemple, pour transcrire l'idée que «une pomme peut avoir la couleur rouge», on construit le triplet «pomme»-«couleur»-«rouge» (sur Wikidata, les identifiants sont Q89-P462-Q3142); pour mentionner qu'une pomme pourrait aussi bien être verte, on ajoute simplement un autre triplet Q89-P462-Q3133, et ainsi de suite.



Interface de Wikidata pour l'objet Q89 https://www.wikidata.org/wiki/Q89

Recherches personnalisées avec Wikidata

L'arrangement des données dans Wikidata a des avantages multiples, à commencer par le fait que Wikidata est intrinsèquement multilingue, de sorte que tout projet qui l'emploie l'est également; et la possibilité de faire des recherches complexes en utilisant le langage SPARQL, qui permet de générer les listes d'objets qui présentent des propriété et satisfont des conditions fournies par l'utilisateur (par exemple «les fruits qui peuvent avoir une couleur rouge mais pas verte, et ont des pépins plutôt qu'un noyau»). De plus, de nombreux outils de visualisation permettent de disposer les listes automatiquement sur des carte si elles comportent des coordonnées («lieux de naissance des autrices qui ont étudié à l'Université d'Edimburgh»), sur une ligne temporelle si elles comportent des dates («dates de formation des universités européennes»), etc.

Le projet «Données structurées»

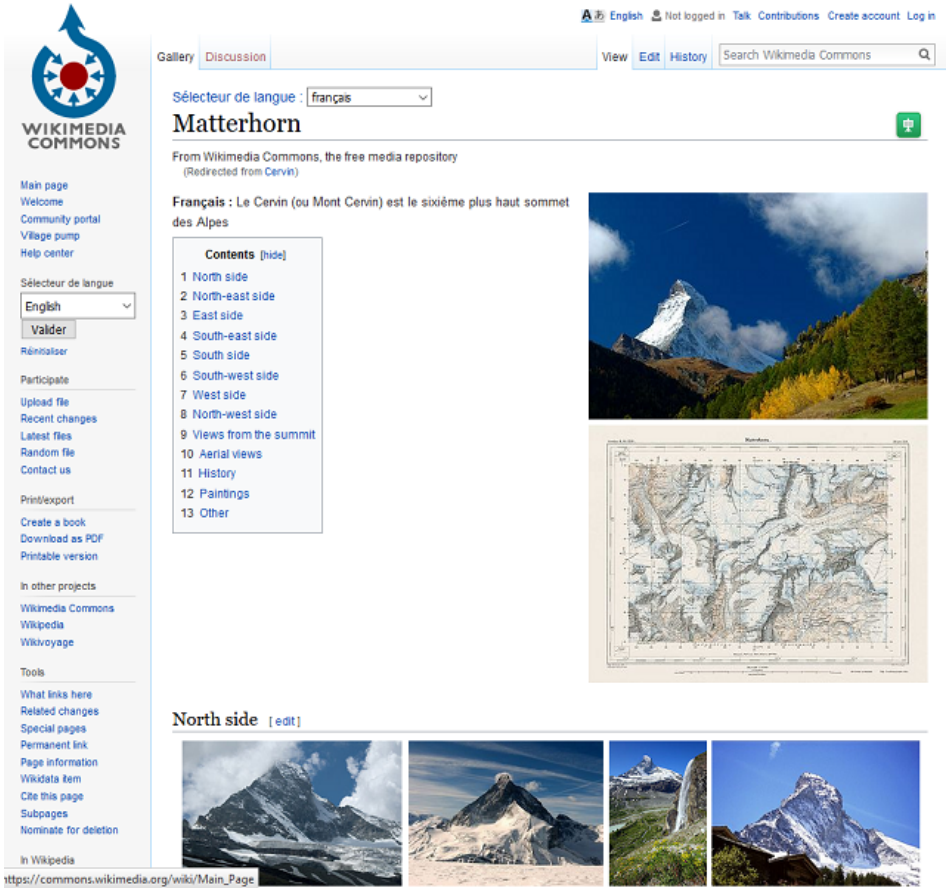
L'organisation actuelle de Commons est informelle et non structurée, ce qui rend difficile d'y chercher des fichiers, et donc d'en améliorer et en utiliser le contenu. Le projet Données Structurées de Wikimedia Commons est une initiative qui vise à y déployer Wikibase, le moteur logiciel de Wikidata. Chaque fichier de Wikimedia Commons constitue en effet un objet susceptible de description, avec des propriétés renseignables à la façon de Wikidata: par exemple une photographie donnée a un sujet, que l'on pourrait renseigner par une référence vers un autre objet Wikidata; elle a un temps de pause (nombre de secondes), une licence (objet Wikidata), etc.

Différents types de propriétés

Beaucoup de ces propriétés sont déjà fournies automatiquement: une photographie d'un appareil numérique moderne contient des métadonnées EXIF qui renseignent le moment de la prise de vue, les paramètres photographiques (longueur focale, temps de pause, sensibilité du capteur, ouverture du diaphragme...), la localisation du boitier pour les appareils dotés d'un GPS, le modèle de l'objectif, etc. D'autres de ces propriétés doivent être fournies par un humain, voire spécifiquement par l'«ayant droit» (personne détentrice des droits d'auteur): le sujet de l'image demande une intervention humaine pour être renseigné; et la licence ne peut être apposée que par l'ayant droit, puisqu'elle a un caractère légal.

Vers des recherches multilingues

L'emploi de Wikibase sur Commons permettra également de bénéficier de la souplesse des objets Wikidata dans les formulaires de téléversement de fichier: il serait ainsi possible à des personnes ne parlant pas du tout anglais de s'y retrouver plus facilement. Des menus déroulants pourraient ainsi à terme y apparaître dans leur langue pour les champs avec un nombre limité d'options (comme les licences par exemple). Pour des propriétés plus complexes, comme la description de l'image, un champs identique à ceux de Wikidata pourrait permettre d'écrire une valeur pour que l'autocomplétion propose dynamiquement un petit menu en fonction de la saisie de l'utilisateur: écrire «Cervin» proposerait une liste comprenant un nom de famille «Cervin», une colline en Antarctique, et le Cervin italo-suisse; en sélectionnant ce dernier, on lierait à l'objet Q1374 de Commons, et la légende apparaîtrait automatiquement comme «Matterhorn» aux germanophones et anglophones, «Cervino» aux italophones, mais aussi «??????????» en russe ou «???????» en japonais.



Recherche de «Cervin» dans Wikimedia Commons redirigée vers les images venant de «Matterhorn» https://commons.wikimedia.org/wiki/Matterhorn

Vers des recherches multicritères

Le projet ouvre en outre des perspectives considérables pour la recherche documentaire sur Commons. Elle est actuellement limitée à des chaînes de caractères dans les titres et les descriptions des fichiers. Avec Données Structurées, il devient possible d'effectuer des recherches selon plusieurs critères, comme les images ayant Q12495 comme sujet et ordonnées par date, pour suivre les étapes de la construction du Burj Khalifa à Dubai; ou de localiser les images d'un même navire sur une carte du monde pour en suivre les croisières. On peut envisager des requêtes complexes fonction des coordonnées géographiques du sujet et de celles de l'appareil photographique, pour déterminer l'angle de la prise de vue.

Vers une meilleure catégorisation

Un autre point où Données Structurées peut apporter une amélioration considérable: la catégorisation. Sur Commons, les images sont groupées dans des catégories et sous-catégories, qui se raffinent au besoin selon le nombre d'images disponibles. Ainsi, un sujet peu couru, comme la commune italienne de Cervino, de 5000 habitants, a-t-il une unique catégorie dotée de quelques images; à l'inverse, un sujet beaucoup photographié, comme la commune d'Esino Lario, de seulement 760 habitants mais qui a accueilli Wikimania en 2016, voit les images associées subdivisées en sous-catégories largement arbitraires, disponibles seulement en anglais, et qui rendent difficile la recherche d'une image en particulier. La difficulté de subdiviser adéquatement les catégories de Commons est brocardée dans la communauté par l'expression «looking left» («regardant à gauche»), en référence à des sous-catégories comme «women looking left». Avec Wikibase, ces catégories arbitraires et uniquement anglophones laissent place à des ensembles de propriétés en nombre illimité, qui n'interfèrent pas les unes avec les autres, et que l'on peut choisir de prendre en compte ou non en fonction des besoins.

Ce projet très ambitieux pourrait donner ses premiers fruits vers 2018 ou 2019, avec le déploiement du moteur et le début du traitement des fichiers actuels de Commons.

L'auteur remercie Sandra Fauconnier pour sa relecture et ses remarques.

[Rama] [Neko]

Administrateur et Oversight de Wikimedia Commons.

Abstract

Français
Parmi les projets de la Wikimedia Foundation, un nouveau projet cherche à uniformiser les métadonnées pour structurer la «jungle» de Wikimedia Commons de façon similaire à la base de connaissance structurée en web sémantique Wikidata.