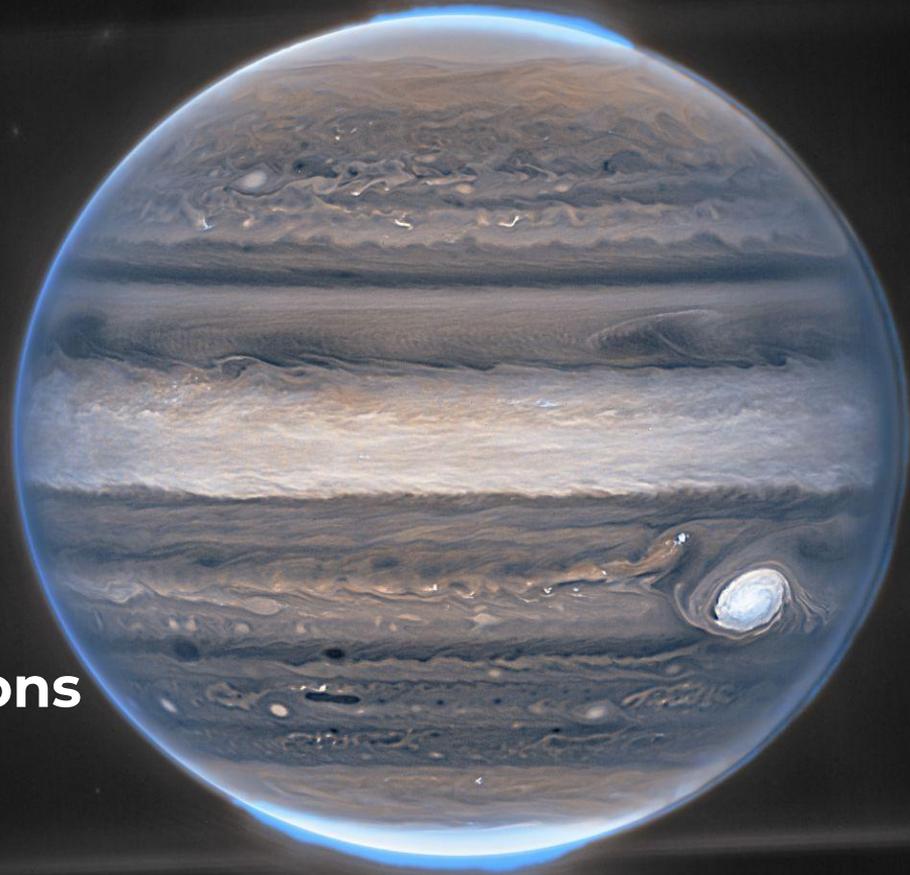




Spacemedia

Import automatisé d'images
spatiales sur Wikimedia Commons

Vincent Privat
Capitole du Libre 2023, Toulouse



NASA, ESA, Jupiter ERS Team; image processing
by Ricardo Hueso (UPV/EHU) and Judy Schmidt

CC by 4.0

À propos de ...



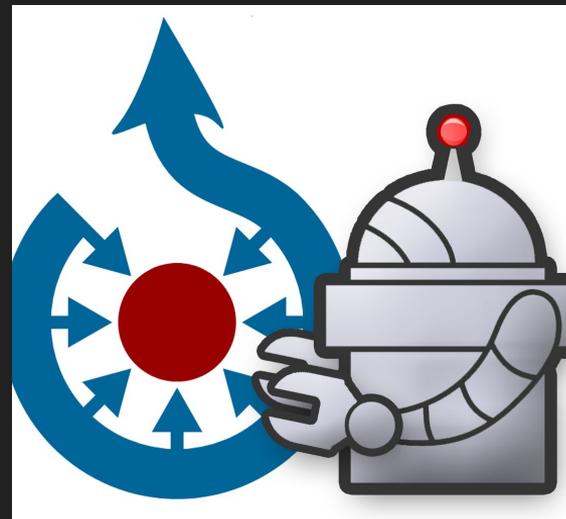
Wikimedia Commons

La médiathèque libre de tous les projets Wikimedia. En 20 ans, 100 millions d'images, vidéos et sons. Basée sur Mediawiki, tout comme Wikipédia et Wikidata.



User:Don-vip (moi)

Contributeur Toulousain depuis 2012. Imports: MH, voyages, aéro, spatial... Fonds André Cros en 2018. Spacegeek. Photos WMFR/ESA/Airbus de JUICE :)



User:OptimusPrimeBot

Mon vaillant bot qui importe depuis 2020 les photos de ce projet. Facilite aussi depuis l'été dernier la vie des admins (détection de doublons)

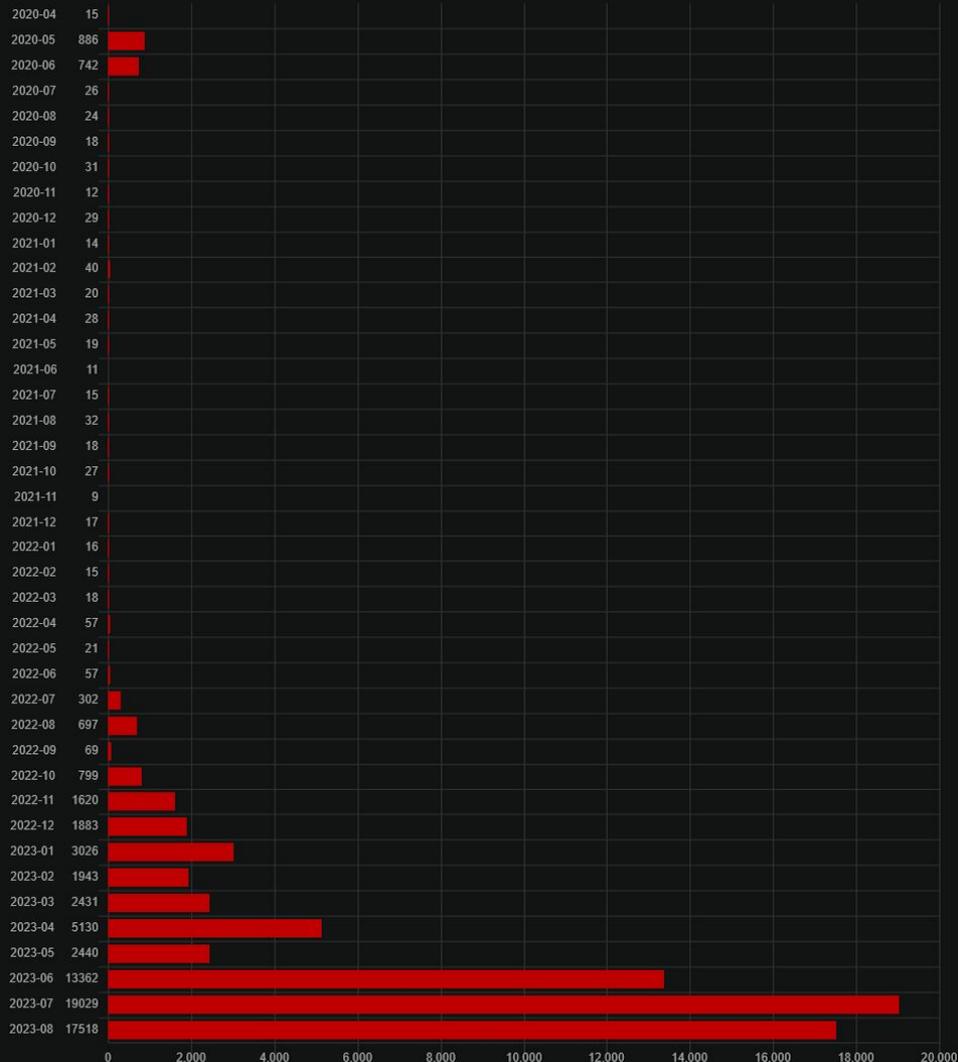
Constat fin 2019

- NASA, NASA partout
 - pourtant il en manque !
 - et on a ... plein de doublons
- L'ESA publie peu mais c'est beau →
 - User:Fæ/Project list/ESA (2017)
- D'autres sources existent :
 - agences, observatoires, USSF
 - commercial (SpaceX, Planet)
- Les imports en :
 - 1 fois pour des archives : cool !
 - n fois régulièrement : pénible !



Idée et objectifs

- Import des images manquantes
- Entièrement automatique
- Réduire au minimum le travail de vérification / catégorisation
- Extensible avec l'ajout régulier de nouvelles sources / missions
- Gérer toutes les informations :
 - Image, titre & description
 - Légende & SDC
 - Géo-localisation
 - Catégories



Sources

Commons:Spacemedia#Monitored_repositories_and_status



Sujets



500/Pa. 183 2023-11-18 00:07:17 UT

Archi



Frontend



Backend
(cron x35)

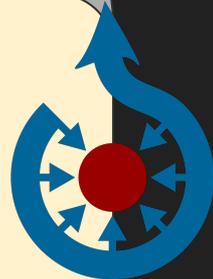
Cloud VPS

Wikimedia Cloud Services

{api}



MediaWiki



Toolforge



Space
media



Replica

Réplication 95%



Prod



MariaDB



gitlab.wikimedia.org



Front

Sources

Name	Free Media	Uploaded	Images to upload	Videos to upload	Ignored
Arianespace (YouTube)	114	74	0	18	22
DLR (Flickr)	3206	3181	0	0	48
ESA	3454	3440	0	0	106
ESA (Flickr)	2826	2744	0	6	230
IAU	2157	1611	0	0	596
NASA (Box)	1344	931	0	236	2
Potential Flickr accounts	0	0	0	0	0
SpaceX	766	766	0	0	1
U.S. Space Force/Command (DVIDS)	15371	7454	0	583	8891
U.S. Space Force/Command (Flickr)	193	192	0	0	65
KARI	1157	1156	1	0	0
NASA (ASTER)	544	542	2	0	0
Webb (ESA)	263	261	2	0	0
Webb (NASA)	396	394	3	0	0
NOIRLab	4639	3975	11	0	0
Copernicus	1156	866	290	0	0
NASA (SIRS)	661	191	470	0	0
Capella Space	857	2	855	0	0
Umbra	1250	3	1243	0	0
Hubble (NASA)	4477	992	3880	0	0
Individuals (Flickr)	12468	7334	4610	167	0
Hubble (ESA)	5208	1262	4614	0	0
NASA (MODIS)	5746	495	5230	0	0
ESO	15066	1702	12650	0	0
NASA (Photojournal)	25825	4260	25380	137	0
NASA (SDO)	76655	7998	35957	32700	0
NASA (Flickr)	64517	22055	37140	726	6997
NASA	220792	32097	158050	6453	28555

NASA

Free Media (220792) Uploaded (32097) Images to upload (158050)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Media	Title	Date
	iss065e242201 	2021-08-13 

Workflow de mise à jour

- Fetch (api / scrap)
- Pour chaque fichier manquant :
 - Calcul des hashes (sha1 / phash)
 - Recherche du fichier sur Commons (par sha1 ou id+mime+phash)
 - Nettoyage titre et descriptions
 - Contrôle licences / blocklist de photographes / entreprises
 - Contrôle blocklist texte (contenu de faible valeur / sans grand intérêt)
 - Upload fichier + wikiCode (titre, descriptions, catégories)
 - Edition SDC (légende, métadonnées)
- Pouet / Txweet. Suivant



Identification des images à vérifier



via une API

Le must !

- Flickr, S3, Box, YouTube
- DVIDS, Photojournal (Solr)
- (images|svs.gsfc).nasa.gov

Documentation, règles claires, clients, paramètres de recherche

En scrappant une galerie

Le cas le plus courant.

Récupération paginée d'une galerie HTML, puis récupération du détail des images une par une. Du plus récent jusqu'à une date antérieure (1 an, 1 mois, 1 semaine suivant les cas)

En testant si des URLs existent

Rare, mais parfois il n'y a pas de liste.

Tests d'URLs avec une partie variable :

- identifiant numérique (KARI)
- date (SDO, "images du jour")

Calcul des hashes SHA1 / phash

- Mediawiki: Hashes SHA1 (!= avec modifs EXIF, résolution, effets...)
- https://en.wikipedia.org/wiki/Perceptual_hashing
- <https://github.com/KilianB/JImageHash>



Réduction
taille
→



- Réduction couleurs
- Calcul différences (moyenne, pixels adjacents...)
- Assignment de bits => calcul du hash



= 8f373714acfcf4d0



Sorties

[https://commons.wikimedia.org/wiki/File:Chausey,_French_Channel_Islands_\(PIA25800_fig1\).jpg](https://commons.wikimedia.org/wiki/File:Chausey,_French_Channel_Islands_(PIA25800_fig1).jpg)

wikiCode

SDC

Informations sur le fichier | Données structurées

Légendes Modifier

français Ajustez en une ligne la description de ce qui représente ce fichier

English Chausey is a group of small islands and islets off the coast of Normandy and is part of the French Channel Islands.

Description modifier | modifier le wikicode

Description **English:**

Chausey is a group of small islands and islets off the coast of Normandy and is part of the French Channel Islands. Chausey bounced back and forth between England and France for 800 years before finally officially belonging to France in the 19th century. The archipelago comprises 365 islands at low tide (2019 image), compared to only 52 islands at high tide (2010 image). The images were acquired July 7, 2018 and September 10, 2018, cover an area of 6.7 by 10.5 km, and are located at 48.9 degrees north, 1.8 degrees west.

With its 14 spectral bands from the visible to the thermal infrared wavelength region and its high spatial resolution of about 50 to 300 feet (15 to 90 meters), ASTER images Earth to map and monitor the changing surface of our planet. ASTER is one of five Earth-observing instruments launched Dec. 18, 1999, on Terra. The instrument was built by Japan's Ministry of Economy, Trade and Industry. A joint U.S./Japan science team is responsible for validation and calibration of the instrument and data products.

The broad spectral coverage and high spectral resolution of ASTER provides scientists in numerous disciplines with critical information for surface mapping and monitoring of dynamic conditions and temporal change. Example applications are monitoring glacial advances and retreats, monitoring potentially active volcanoes, identifying crop stress, determining cloud morphology and physical properties, wetlands evaluation, thermal pollution monitoring, coral reef degradation, surface temperature mapping of soils and geology, and measuring surface heat balance.

The U.S. science team is located at NASA's Jet Propulsion Laboratory in Pasadena, Calif. The Terra mission is part of NASA's Science Mission Directorate, Washington.

More information about ASTER is available at <https://asterweb.jpl.nasa.gov/>.

Date 7 juillet 2018 (published 2023-03-16T17:27:44Z)

Source [Catalog page](#) • [Full-res \(JPEG\) • \[Full-res \\(TIF\\) • \\[Full-res \\\(MP4\\\) • \\\[Full-res \\\\(GIF\\\\)\\\]\\\(#\\\)\\]\\(#\\)\]\(#\)](#)

Auteur NASA-NETWAST/Japan Space Systems, and U.S./Japan ASTER Science Team

Autres versions

 TIF version ?

 JPEG version ?

Lieu de la prise de vue 48° 54′ 00″ N, 1° 48′ 00″ O Voir cet endroit et d'autres images sur [OpenStreetMap](#)

 **Autres langues** :

 This media is a product of the **Terra mission**.
Credit and attribution belongs to the mission team, if not already specified in the "author" row

Conditions d'utilisation modifier | modifier le wikicode

 Le détenteur des droits d'auteur de ce fichier, NASA/JPL-Caltech, autorise n'importe qui à l'utiliser pour n'importe quelle utilisation, pourvu que le détenteur des droits d'auteur soit correctement attribué. La redistribution, les œuvres dérivées, l'utilisation commerciale et toutes les autres utilisations sont autorisées.

Attribution: Courtesy NASA/JPL-Caltech

According to JPL's image use policy • [additional restriction](#) is that no endorsement of any product or service by Caltech, JPL, or NASA is claimed or implied.

Caltech's disclaimer: Caltech makes no representations or warranties with respect to ownership of copyrights in the images, and does not represent others who may claim to be authors or owners of copyright of any of the images, and makes no warranties as to the quality of the images. Caltech shall not be responsible for any loss or expenses resulting from the use of the images, and you release and hold Caltech harmless from all liability arising from such use.

Usage on the English Wikipedia: On the English Wikipedia you can use the {{JPL Image}} template to display the copyright notice. (See [Wikipedia Using JPL Images](#) for details)

Informations sur le fichier | **Données structurées**

Éléments décrits dans ce fichier Modifier

dispeim

de Wikidata ...

archipel de Chausey Marquer comme préintentionné

créé par Modifier

Terra Marquer comme préintentionné

statut des droits d'auteur Modifier

sous copyright Marquer comme préintentionné

licence Modifier

attribution only license *anglais* Marquer comme préintentionné

type MIME Modifier

image/jpeg Marquer comme préintentionné

hauteur Modifier

450 pixel Marquer comme préintentionné

largeur Modifier

700 pixel Marquer comme préintentionné

taille des données Modifier

92 995 octet Marquer comme préintentionné

pris avec Modifier

Advanced Spaceborne Thermal Emission and Reflection Radiometer Marquer comme préintentionné

somme de contrôle Modifier

2b9985f0c47567ba4ed16148dc9d5a370ad03ed Marquer comme préintentionné

méthode de détermination: SNA-1



Volume, doublons, ressources, infra

- NASA : ~500 000 images analysées (galeries, Flickr, Photojournal, missions)
- Encore beaucoup d'autres à identifier et traiter
- Images astronomiques énormes (plusieurs Go) mettent à mal les serveurs pour
 - les calculs de hashes qui demandent d'avoir les données en mémoire
 - les générations des miniatures
- Recherche de doublons par phash ne peut s'effectuer sur base complète (100M)
 - astuce pour filtrer le périmètre : recherche identifiant ou première phrase
- Toolforge est avant tout prévu pour héberger des petits scripts PHP/Python
 - passage à Cloud VPS pour avoir une grosse VM backend
 - maintien de Toolforge pour le frontend



Best-of de la NASA

- Se tromper dans les noms, les lieux, les dates
- Indiquer qu'une image vient du futur
- Publier des images sous copyright sans l'indiquer
- Publier des images en domaine public avec une licence non libre
- Se m♦ langer les pinceaux dans les jeux de caractères
- Utiliser tous les moyens de communication possible
- Décomissionner des sites sans se soucier de la disponibilité des images
- Faire en sorte que chaque site soit techniquement différent
- Faire de la pub pour les réseaux sociaux dans les descriptions



Vie et mort de la libre diffusion

- En dehors de la NASA, peu de culture de libre diffusion à long terme
- 2010-2012 : Mise en place au DLR de la CC-BY 3.0 (Marco Trovatiello + WMDE)
- 2014 : CC-BY-SA 3.0 IGO à l'ESA (HRSC/Rosetta) (Marco Trovatiello + FU Berlin)
- 2014 : Planet Labs diffuse sa galerie en CC-BY-SA 4.0
- 2015 : Creative Commons crée le CC0 pour SpaceX
- 2018 : Arianespace diffuse des vidéos YouTube en CC-BY
- 2019 : SpaceX et Planet passent en CC-BY-NC ☠️
- 2020 : Arianespace repasse en licence YouTube standard ☠️
- 2021 : le KARI arrête de diffuser des images en KOGL Type 1 ☠️
- 2022 : le DLR passe en CC-BY-NC-ND ☠️



Perspectives

- CI/CD quand phab:T194332 sera terminé (Toolforge Build Service)
- Mettre à jour les métadonnées SDC des images existantes
- Créer automatiquement les catégories annuelles / journalières
- Créer automatiquement les items Wikidata des satellites et lancements
- Isoler la détection de doublons dans un projet autonome
 - Sauf si entre temps la WMF propose une solution officielle
- Créer de nouvelles instances thématiques, par exemple:
 - Sciences : Ifremer, CERN, Institut Alfred Wegener, etc.
 - Militaire : USA, Pays-Bas, autres pays ?
 - GLAM ?





Merci !

@VincentPrivat
@vincentprivat.bsky.social
@VincentPrivat@mastodon.social