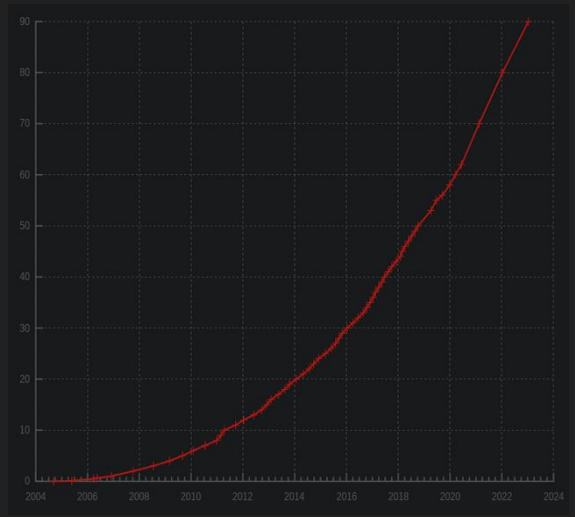
Spacemedia

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

À propos de …

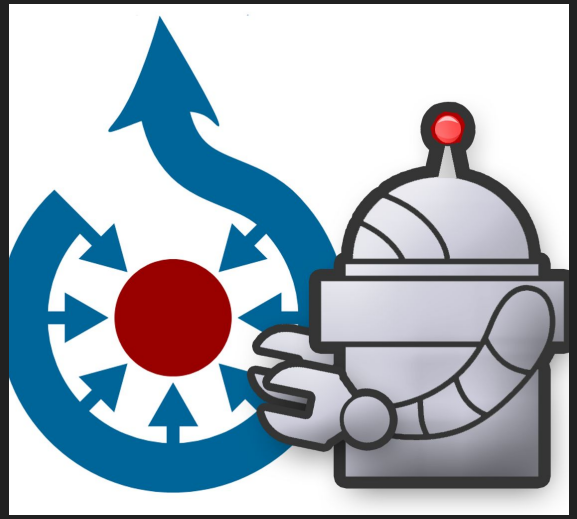


Wikimedia Commons

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 :)

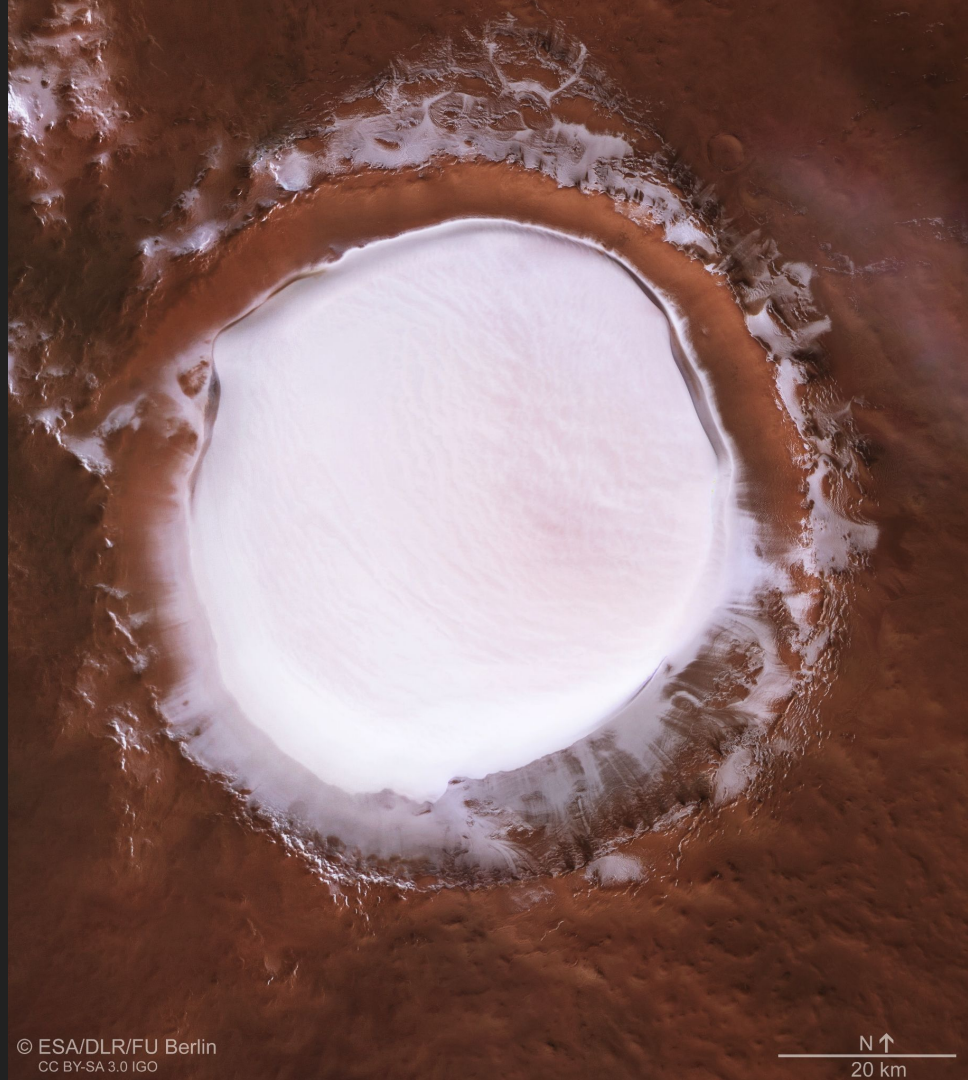


User:OptimusPrimeBot

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)

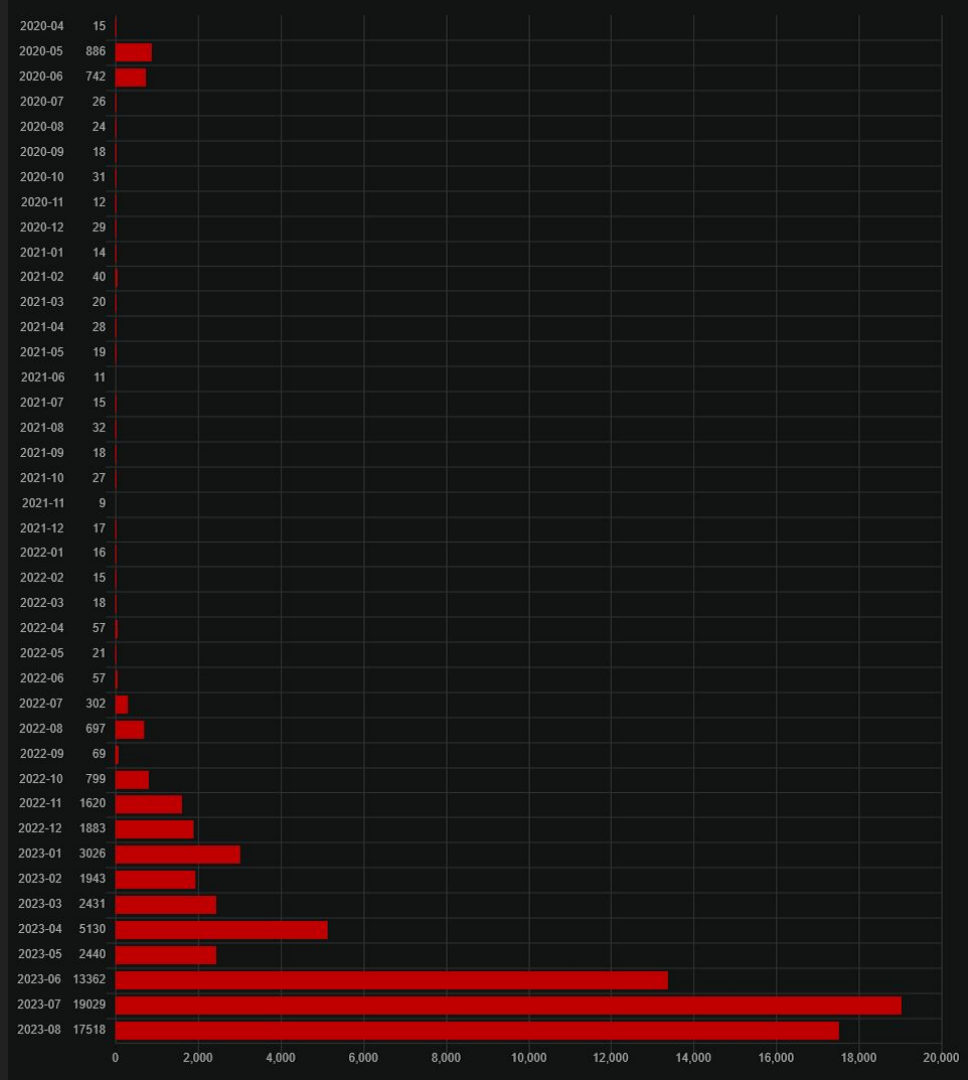
Constat fin 2019

- 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

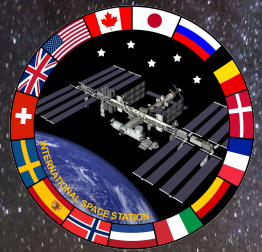
- Import des images manquantes
- Entièrement automatique
- Réduire au minimum le travail de vérification / catégorisation
- Extensible avec l'ajout régulier de nouvelles sources / missions
- Gérer toutes les informations :
  - Image, titre & description
  - Légende & SDC
  - Géo-localisation
  - Catégories



Sources

Commons:Spacemedia#Monitored_repositories_and_status

Sujets


INTERNATIONAL SPACE STATION


SDO/AIA 193 2023-11-16 00:07:17 UT

Archi

gitlab.wikimedia.org

Wikimedia Cloud Services

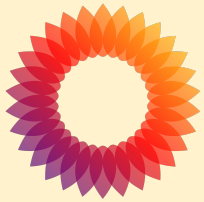
Frontend

Toolforge

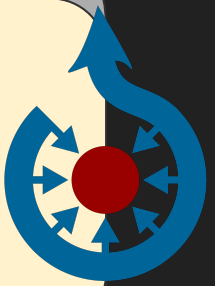
Backend
(cron x35)

Cloud VPS

{api}



MediaWiki



Space
media

Replica

Réplication 95%

Prod

Front

Spacemedia Home (current) Documentation Sources Issues Repository

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	1245	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

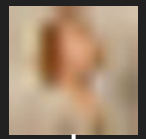
Rare, mais parfois il n'y a pas de liste.
Tests d'URLs avec une partie variable :

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

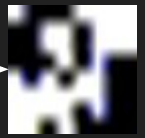
Calcul des hashes SHA1 / phash

- 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…)
- Assignation de bits => calcul du hash



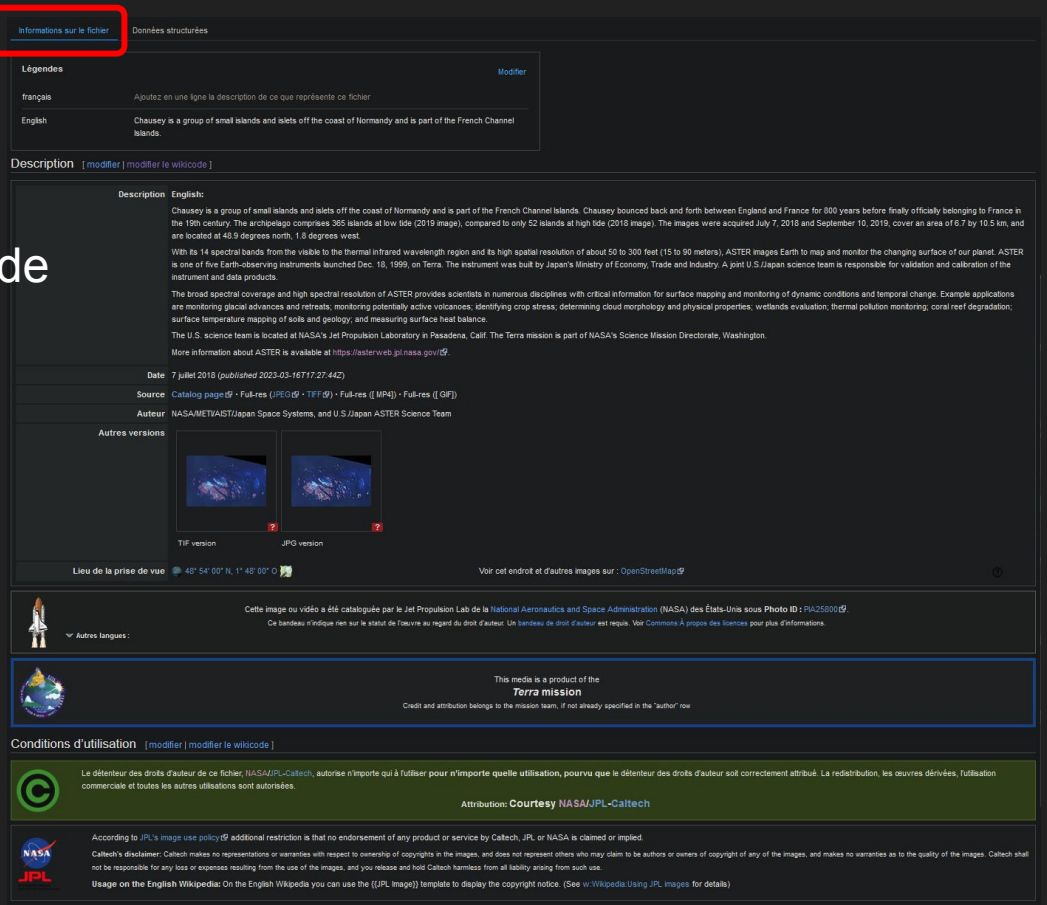
= 8f373714acfcf4d0

Sorties

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

wikiCode

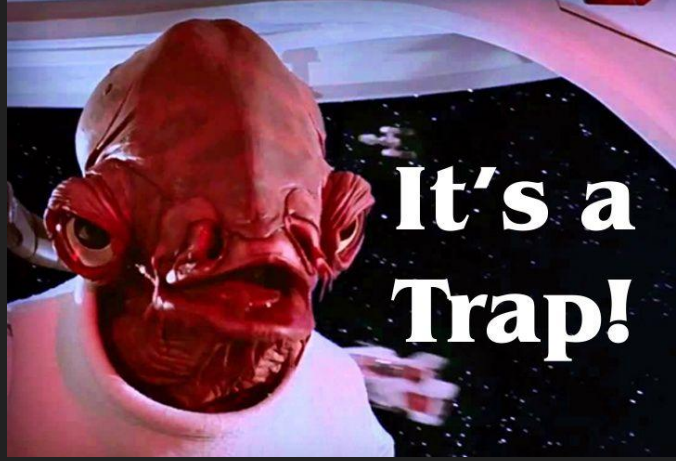
SDC



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

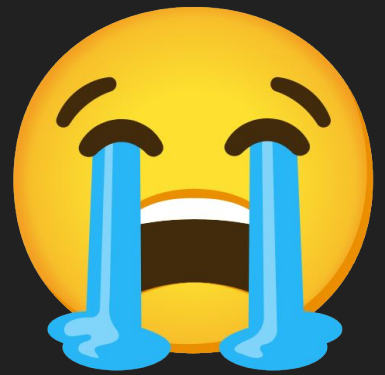
Best-of de la NASA



- 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 Trovatello + WMDE)
- 2014 : CC-BY-SA 3.0 IGO à l'ESA (HRSC/Rosetta) (Marco Trovatello + 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