



Scientific
Information
Service

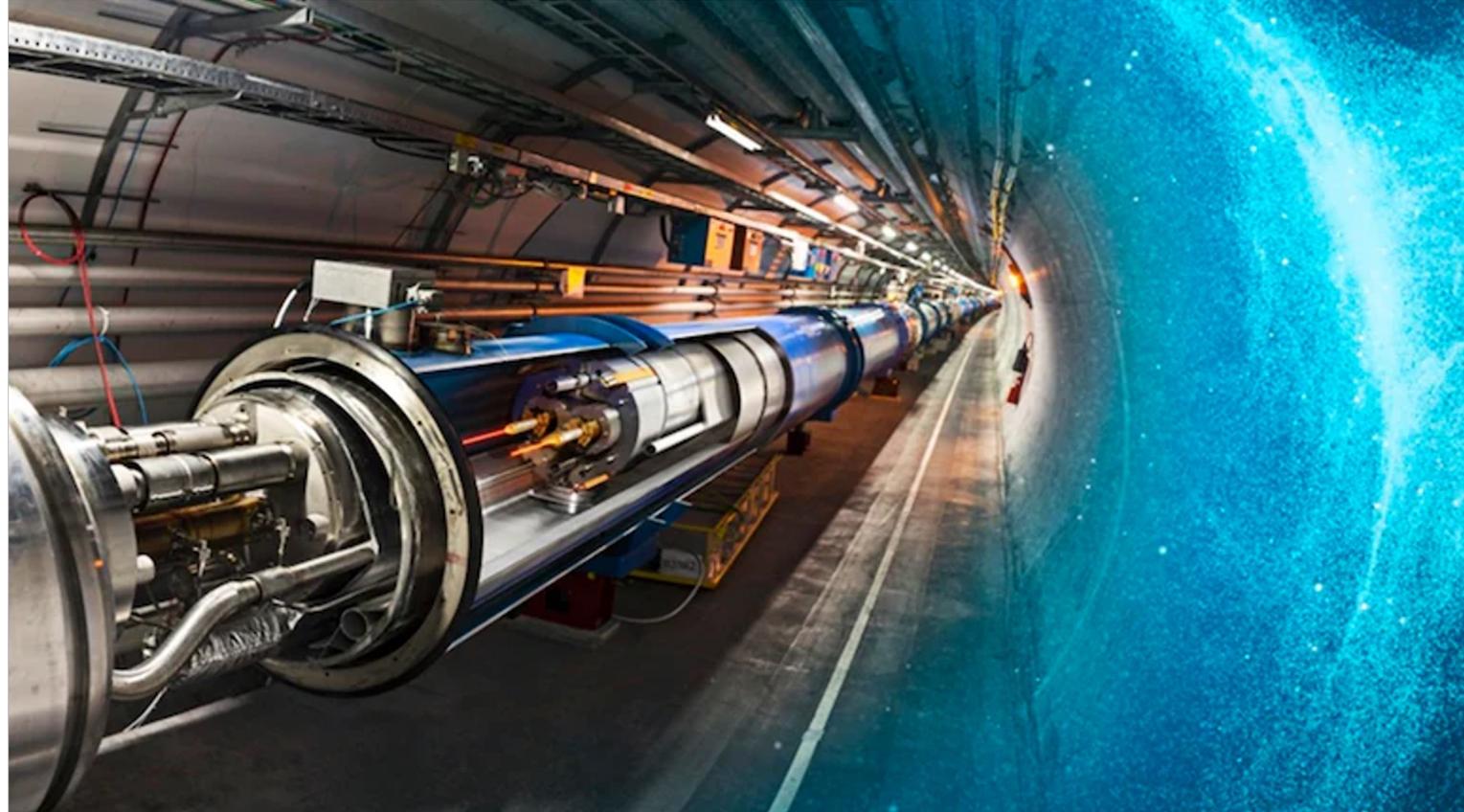
Sauvegarder et ouvrir la recherche pour tous

Leçons tirées du développement de services globaux
pour les sciences ouvertes et la gestion des données de recherche

Dr. Sünje Dallmeier-Tiessen, CERN
Cotonou, Octobre 2022

Mission du CERN

Fournir une gamme unique d'accélérateurs de particules qui permettent la recherche à la pointe des connaissances humaines, effectuent des recherches de classe mondiale en physique fondamentale, unissent des personnes du monde entier pour repousser les frontières de la science et de la technologie, au profit de tous



Collaboration ouverte et inclusive créée en 1954 avec 12 états membres européens

23 États membres

Autriche – Belgique – Bulgarie – République tchèque – Danemark – Finlande – France – Allemagne – Grèce – Hongrie – Israël – Italie – Pays-Bas – Norvège – Pologne – Portugal – Roumanie – Serbie – Slovaquie Espagne – Suède – Suisse – Royaume-Uni

3 États membres associés dans la phase préalable à l'adhésion
Chypre – Estonie – Slovaquie

7 États membres associés
Croatie – Inde – Lettonie – Lituanie – Pakistan Turquie – Ukraine

6 Observateurs
Japon – Russie – États-Unis
Union européenne – JINR – UNESCO



Le budget annuel du CERN est de 1200 MCHF

~ 2'600 Collaborateurs
~ 2'000 Contractors
~ 13'000 Physiciens (utilisateurs)

Plus de 50 accords de coopération avec des États et territoires non membres

Albanie – Algérie – Argentine – Arménie – Australie – Azerbaïdjan – Bangladesh – Biélorussie – Bolivie Bosnie-Herzégovine – Brésil – Canada – Chili – Colombie – Costa Rica – Équateur – Égypte – Géorgie – Islande
Iran – Jordanie – Kazakhstan – Liban – Malte – Mexique – Mongolie – Monténégro – Maroc – Népal
Nouvelle-Zélande – Macédoine du Nord – Palestine – Paraguay – République populaire de Chine – Pérou – Philippines – Qatar
République de Corée – Arabie saoudite – Sri Lanka – Afrique du Sud – Thaïlande – Tunisie – Émirats arabes unis – Vietnam

“... et les résultats de ces travaux expérimentaux et théoriques sont publiés ou de toute autre façon rendus généralement accessibles.”

Convention de fondation du CERN (1953)

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CONVENTION

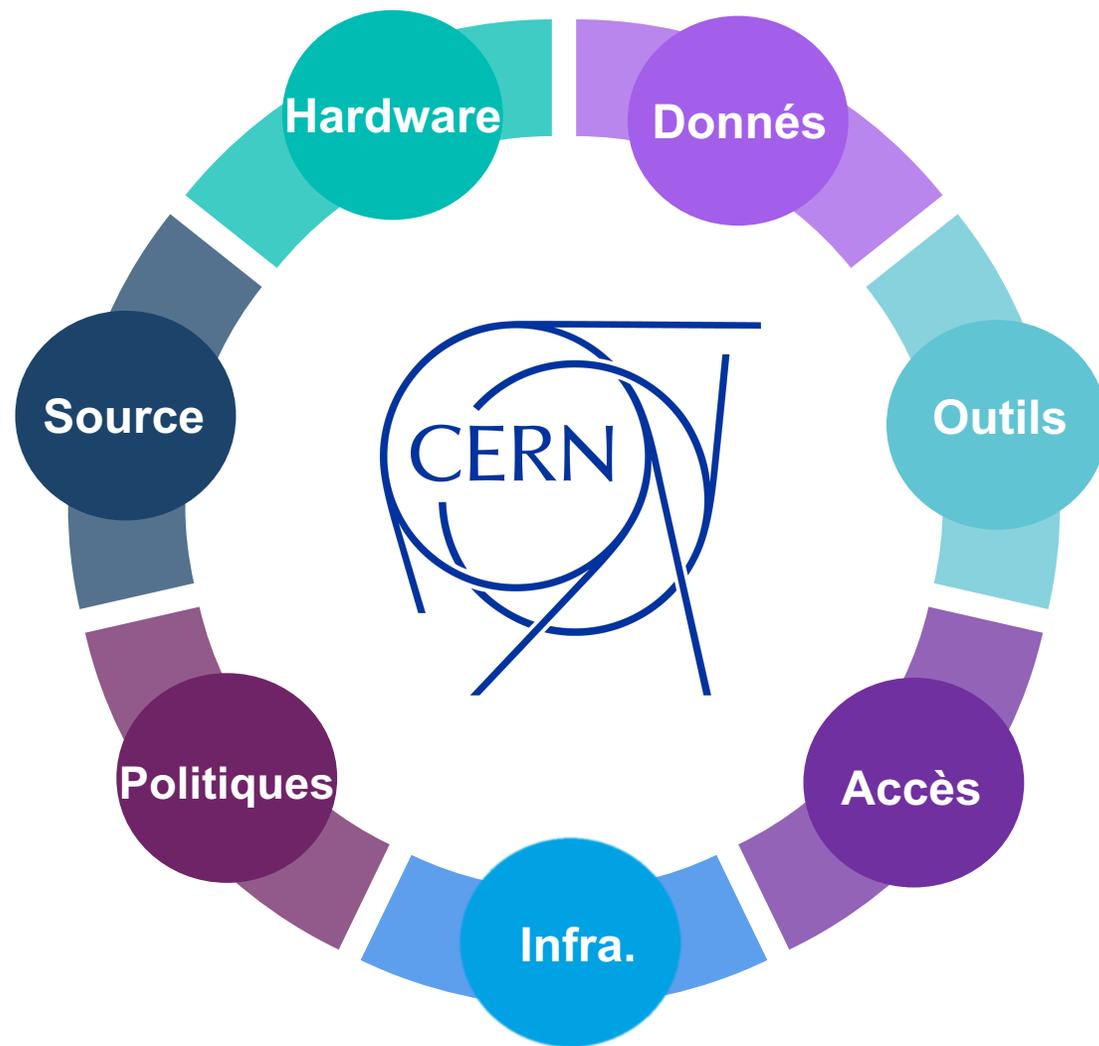
FOR THE ESTABLISHMENT OF A EUROPEAN ORGANIZATION
FOR NUCLEAR RESEARCH

PARIS, 1st JULY, 1953



Illustration par Stephanie van de Sandt

La science ouverte au CERN





Le CERN stocke plus de 100 PB de données par an

Collection de données



- 1 PB/sec de données de collisions impossibles à stocker
- Systèmes de déclenchement de chacun des 4 détecteurs principaux :
 - Niveau 1 : réduction presque en temps réel (< 2,5 microsecondes) à 100 000 événements/seconde
 - Niveau 2 : dans les 200 microsecondes, réduction supplémentaire à 1 000 événements/seconde pour le stockage
- 90 PB/an du LHC + 25 PB d'autres expériences

Stockage de données



- Le centre de données du CERN dispose de 411 PB sur bandes + 365 PB sur disques (y compris les doublons)
- EOS (le système de stockage propriétaire du CERN) héberge désormais plus de 5 milliards de fichiers et a servi 2,5 exaoctets de données (en 2020)
- Latence élevée des bandes problématique pour les services de données ouvertes (1-3 minutes ou même plus)

Au début: Les trois piliers de la réussite des données ouvertes

Préservation et accès à long terme

Services

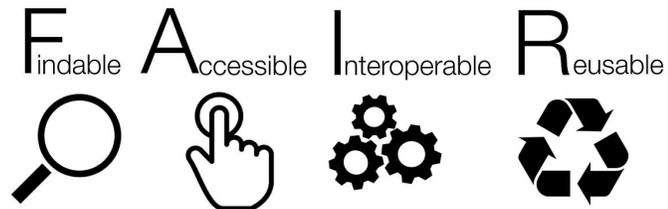
Plateformes d'archivage et publication

Publier et partager

Plateformes numériques
(+l'assurance qualité)

Réutilisation

Plateformes d'analyse de données de recherche reproductibles

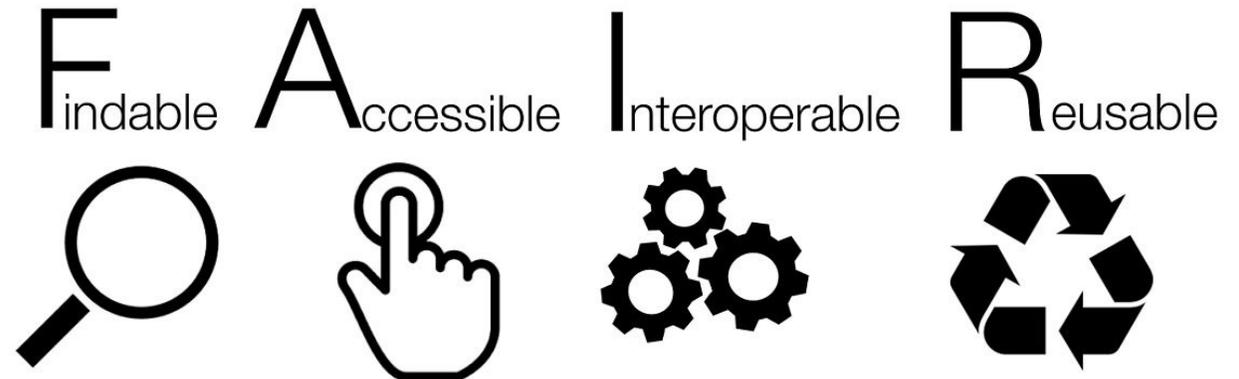


FAIR est le nouveau standard

Bien que les données ouvertes soient cruciales, **l'intégrité des résultats** de la recherche est la base.

La préservation des résultats tout au long du processus de recherche est nécessaire afin d'assurer...

- La découvrabilité, trouvabilité
- L'accessibilité à long terme
- La standardisation, interopérabilité
- L'assurance qualité
- La (ré)utilisation



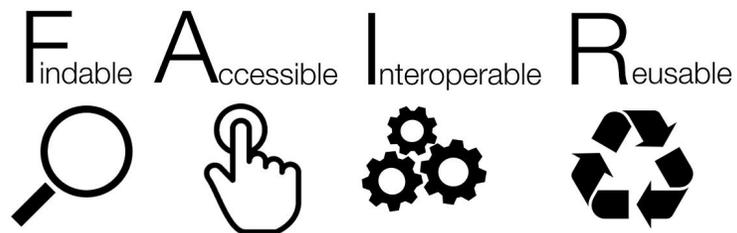
Le CERN et ses partenaires ont créé une suite d'outils de gestion des données de la recherche



- Le portail Open Data du CERN permet aux expériences HEP de partager leurs données (collisions, simulations, logiciels associés, etc.)
- Contient déjà > 2 PB de données

- Zenodo est un serveur de recherche ouvert multidisciplinaire commandé par la EC
- Inclut 80 % des DOI de logiciels du monde

- HEPData est la plateforme principale pour les tableaux et les ensembles de données liés à env. 10'000 articles.
- Géré par l'Université de Durham en collaboration avec le CERN



ZENODO

un exemple de réalisation

The screenshot displays the Zenodo website interface. At the top, there is a blue header with the Zenodo logo, a search bar, and navigation links for 'Upload' and 'Communities'. The user profile 'jose.benito.gonzalez@cern.ch' is visible in the top right.

Featured communities

- Transform to Open Science**: A community focused on the TOPS mission. It includes a 'Browse' button and a 'New upload' button. The description states: 'Transform to Open Science (TOPS) is a \$40 million, 5-year mission, led by NASA's Science Mission Directorate's Open-Source Science initiative. Within the TOPS mission, NASA is designating 2023 as the Year Of Open Science, a community initiative to spark change and inspire open science...'. It is curated by 'cgentemann'.

Recent uploads

- Transformers: State-of-the-Art Natural Language Processing** (October 1, 2020, v4.20.1): Software, Open Access. Authors: Wolf, Thomas; Debut, Lysandre; Sanh, Victor; Chaumond, Julien; Delangue, Clement; Moi, Anthony; Cistac, Perrick; Ma, Clara; Jernite, Yacine; Plu, Julien; Xu, Canwen; Le Scao, Teven; Gugger, Sylvain; Drame, Mariama; Lhoest, Quentin; Rush, Alexander M. Description: 'This patch releases fixes a bug in the OPT models and makes Transformers compatible with huggingface_hub version 0.8.1. Add final_layer_norm to OPT model #17785 Prepare transformers for v0.8.0 huggingface-hub release #17716'. Uploaded on June 21, 2022. 83 more version(s) exist for this record.
- A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration** (June 19, 2022, v119): Dataset, Open Access. Authors: Banda, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo. Description: 'Version 119 of the dataset. MAJOR CHANGE NOTE: The dataset files: full_dataset.tsv.gz and full_dataset_clean.tsv.gz have been split in 1 GB parts using the Linux utility called Split. So make sure to join the parts before unzipping. We had to make this change as we had huge issues uploading...'. Uploaded on June 21, 2022. 119 more version(s) exist for this record.
- geosapi: GeoServer REST API R Interface** (June 17, 2022, v0.6-3): Software, Open Access. Author: Emmanuel Blondel. URL: <https://github.com/eblondel/geosapi/blob/master/NEWS.md#geosapi-06-3>. Uploaded on June 17, 2022. 10 more version(s) exist for this record.

Need help?

Contact us

Zenodo prioritizes all requested related to the COVID-19 outbreak.

We can help with:

- Uploading your research data, software, preprints, etc.
- One-on-one with Zenodo supporters.
- Quota increases beyond our default policy.
- Scripts for automated uploading of larger datasets.

Why use Zenodo?

- **Safe** — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
- **Trusted** — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
- **Citeable** — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.
- **No waiting time** — Uploads are made available online as soon as you hit publish, and your DOI is registered within seconds.
- **Open or closed** — Share e.g. anonymized clinical trial data with only medical professionals via our restricted access mode.
- **Versioning** — Easily update your dataset with our versioning feature.
- **GitHub integration** — Easily preserve your GitHub repository in Zenodo.
- **Usage statistics** — All uploads display...

<https://zenodo.org/>



Faciliter la tâche des utilisateurs (*Zenodo)

Télécharger

50 GB* pour chaque set de données
Tous les formats de fichiers sont acceptés

Delete Save Publish

New upload

Instructions: (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

Files

Choose files Start upload

Filename (4 files)	Size	Progress	Delete
2020-06-report-3.json.gz	5 Mb		
broker-2019-01-17T13:48:21.mk	200 Mb		
pubmed-events.json	957 Kb		
storage_growth.py	703 B		

Note: File addition, removal or modification are not allowed after you have published your upload. This is because a Digital Object Identifier (DOI) is registered with DataCite for each upload.

(minimum 1 file required, max 50 GB per dataset - contact us for larger datasets)

Décrire

Métadonnées riches mais flexibles
Basées sur le schéma DataCite
Réserver le DOI avant la publication

Upload type required

Publication Poster Presentation Dataset Image Video/Audio Software Lesson Other

Basic information required

Digital Object Identifier 10.5272/zenodo.682186

Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.

Reserve DOI

Publication date 2020-10-15

Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

Title Data and descriptions for XYZ research project

Required.

Authors Alice Smith Alex Ioannidis

ORCID (e.g. 0000-0002-18)

Description This is the datasets and their description/documentation for the work published on project XYZ.

Publication

DOI citable
Formats d'exportation

August 12, 2020 Dataset Open Access

OpenAIRE Covid-19 publications, datasets, software and projects metadata.

4,620 views 438 downloads

Indexed in OpenAIRE

Publication date: August 12, 2020

DOI: 10.5281/zenodo.3980491

Communities: OpenAIRE, OpenAIRE Research Graph, Zenodo

License (for files): Creative Commons Zero v1.0 Universal

Versions: Version 1.0 Aug 12, 2020

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.3980490. This DOI represents all versions, and will always resolve to the latest one. Read more.

Files (52.8 MB)

Name	Size
COVID-19.json.gz	52.8 MB

Citations

Show only: Literature (0) Dataset (0) Software (0) Unknown (0)

Citations to this version

Donner le pouvoir aux utilisateurs !

Communities

The image shows three overlapping screenshots of the Zenodo website. The top screenshot is for 'LORY - Lucerne Open Repository', showing a search bar and 'Recent uploads'. The middle screenshot is for 'TWISTx Proceedings', also showing a search bar and 'Recent uploads'. The bottom screenshot is for 'Knowledge Junction', showing a search bar and a list of recent uploads, including 'Chemical Monitoring Reporting Guidelines (SSD2)' and 'Concept paper on the future of data in EFSA'. A yellow arrow points from the EFSA news article in the bottom screenshot to the EFSA logo in the middle screenshot.

Projects, Subjects, Institutes, Nations, Conferences, ...

The screenshot shows the 'Software Carpentry' community page. It features a header with the community name and a description: 'Materials published by Software Carpentry'. Below this, it lists 'The most recent upload:' which is a document titled 'Software Carpentry: Using Databases and SQL' by Allen, James; Andrea, Paula; Banaszkiewicz, Piotr; Barry, Pauline; et al. The document is dated 15 May 2018 and is available under a Creative Commons license. A 'View' button is present next to the document title.

Want your own community? [Sign Up](#)

It's easy. Just sign-up and create a new community.

- **Curate** – accept/reject what goes in your community collection.
- **Export** – your community collection is automatically exported via OAI-PMH
- **Inbound** – not custom inbound link to send to

Accept Reject

The screenshot shows the EFSA website with a news article titled 'EFSA to share data on open-access platform' dated 17 January 2019. The article features a large image of a green keyboard key with a padlock icon and the word 'Access' written on it. Below the image, the text states: 'EFSA has taken a major step towards becoming a fully open data organisation by committing to publish the scientific data it uses for EU-wide monitoring programmes and surveys and many of its risk assessments. In a report published today, EFSA lays out how it intends to share data collected in areas such as: food consumption habits; pesticide residues in food; chemical contaminants and additives in food; foodborne disease outbreaks; and antimicrobial resistance.'



Standardisation et flexibilité

zenodo

Search

Upload Communities

Log in Sign up

December 31, 2018

Figure Open Access

Fig. 7 in A mountain of millipedes VI. New records, new species, a new genus and a general discussion of Odontopygidae from the Udzungwa Mts, Tanzania (Diplopoda, Spirostreptida, Odontopygidae)

Enghoff, Henrik

Fig. 7. *Hoffmanides dissutus* (Hoffman, 1963), ♂, from Udzungwa Mts National Park. Photograph by A. Illum. Scale bar = 5 mm.

Preview



Files (2.8 MB)

Name	Size	
figure.png	2.8 MB	Preview Download
md5:9191af35e44c7ba2659ed5a41b2b722b		

Citations

Show only: Literature (0) Dataset (0) Software (0) Unknown (0) Citations to this version

Part of



Indexed in



Publication date: December 31, 2018

DOI: 10.5281/zenodo.1146170

Keyword(s):

Biodiversity Taxonomy Animalia Arthropoda Diplopoda Spirostreptida Odontopygidae Hoffmanides

Published in: European Journal of Taxonomy: 394 pp. 1-29.

Related identifiers:

Cited by: <http://treatment.plazi.org/id/038D2864FFA3FFA2FDA4FE88B9FC09>

Part of: 10.5852/ejt.2018.394, <http://treatment.plazi.org/pub/FFB4501CFFB2FFB1FF94FFB0892EFF8E> (LSID), <https://zenodo.org/record/1146158>, <https://zenodo.org/record/1146158>

Communities: Biodiversity Literature Repository

License (for files): License Not Specified

Versions

Version 1 10.5281/zenodo.1146170 Dec 31, 2018

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.1146169. This DOI represents all versions, and will always resolve to the latest one. Read more.

Share



Keyword(s):

PID Graph, research.fi, persistent identifier, interoperability, legal interoperability, research output

Subject(s):

research work research activity academic writing research data identifiers (information) interoperability

Meeting:

Pidapalooza 2021, 27-28.01.2021 (Session Block 2, track 2)

Related identifiers:

Cites 10.11646/zootaxa.4193.3.7 (Publication) 10.1653/0015-4040(2005)88[502:KTTFOC]2.0.CO;2 (Publication) 10.11646/zootaxa.4747.2.10 (Publication)

Has part

<http://treatment.plazi.org/id/03DF87B7FF86FFBF6F68FF340F5BFE10> (Taxonomic treatment) <http://treatment.plazi.org/id/03DF87B7FF86FFBF6F68FDB00FE8FC3A> (Taxonomic treatment) <http://treatment.plazi.org/id/03DF87B7FF86FFBF6F68FB9D0B1CF82B> (Taxonomic treatment) <http://treatment.plazi.org/id/03DF87B7FF83FFBAFF68FF3408D7FAD2> (Taxonomic treatment) 10.5281/zenodo.5497141 (Figure) 10.5281/zenodo.5497144 (Figure)

Custom keywords:

Genus Pliolestes Species venetus Kingdom Animalia Order Paucituberculata Scientific name authorship Goin Phylum Chordata Taxon rank species Family Caenolestidae

Locations:

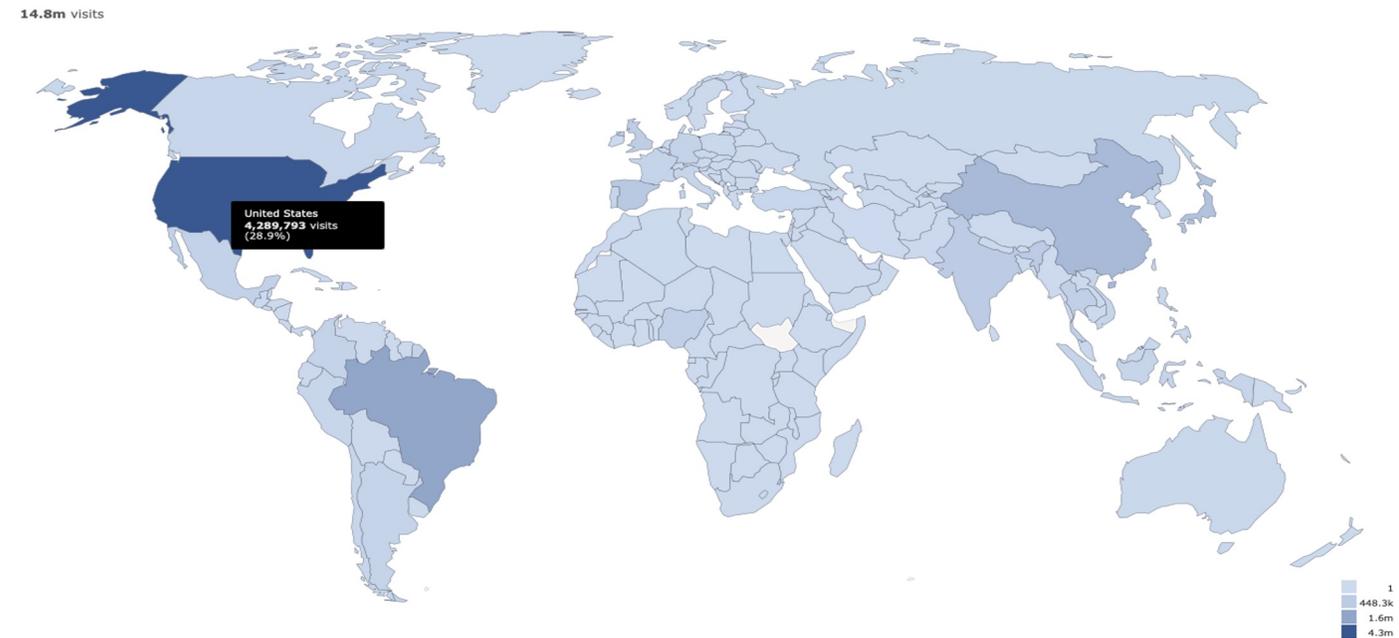
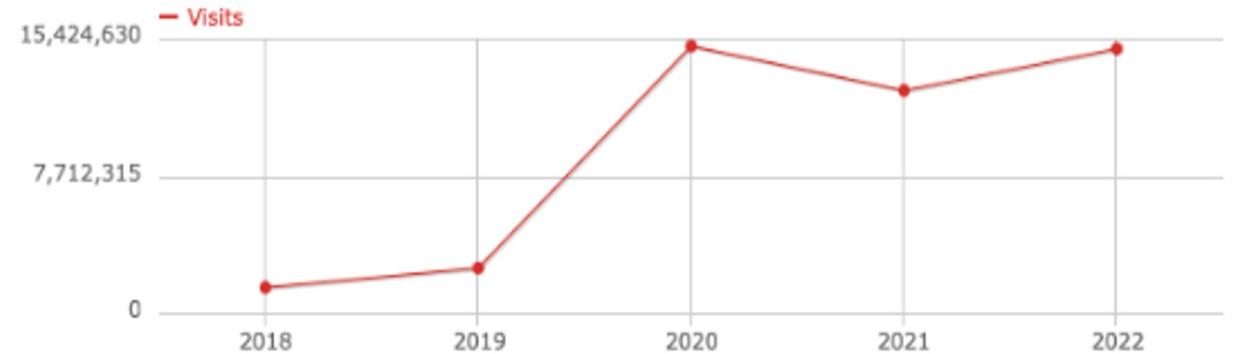
R / V Alis, EBISCO Expedition, st. DW 2613, Plateau des Chesterfield (-19.616667, 158.7) NEW CALEDONIA: 42.5 x 19.4 mm, R/V Alis, EBISCO Expedition, st. DW 2613, Plateau des Chesterfield, 19°37' S, 158°42' E, 519-522 m (MNHN IM-2007-30854; Fig. 9A; GenBank accession number (cox1 sequence): KJ550158). R / V Alis, EBISCO Expedition, st. DW 2610, Plateau des Chesterfield (-19.566668, 158.68333) 16.3 x 13.0 mm, R/V Alis, EBISCO Expedition, st. DW 2610, Plateau des Chesterfield, 19°34' S, 158°41' E, 486-494 m (MNHN IM-2000-30789; paratype 2; Fig. 9D; fragment of the spire, well preserved). Coral Sea (-19.616667, 158.7) NEW CALEDONIA: Coral Sea, Plateau des Chesterfield, 19°37' S, 158°42' E, 519-522 m (EBISCO st. DW 2613).



Scientific Information Service

Zenodo en chiffres

- **~2,8 millions d'entrées**
 - 1.5m textes
 - 750k images
 - 200k logiciels
 - 200k sets de données
- **1000 TB de données,**
- **~8 millions de fichiers**
- **~10 000 communautés**
- **15 millions de visiteurs/an**
- **Déjà 14,8 millions en juin 2022**



ZENODO A OUVERT LA VOIE AUX "CLONES" À SUIVRE

aperta

Son yüklenen çalışmalar

TERMAL GÜNEŞ ENERJİSİ UYGULAMALARI

TEKNİK RAPORLARDA KARŞILAŞILAN BAZI DEYİMLER

AKSAY ÜNİTESİ(ARKEOLOJİK ESERLERİN SPEKTROSKOPİK VE ANALİTİK YÖNTEMLERLE İNCELENMESİ)BİLMİSEL TOPLANTI BİLDİRİLERİ 1,(23-25 KASIM 1988)

ARKEOMETRİ ÜNİTESİ BİLİMSEL TOPLANTI BİLDİRİLERİ 6,(15-17 MAYIS 1988,İSTANBUL)

zenodo

Recent uploads

Aligned ISNI and Ringgold identifiers for institutions

POPC with 0, 10, 20, and 30 mol-% of cholesterol at 310 K. Chamm36 force field.

Gene Ontology Data Archive

DADA2 formatted 16S rRNA gene sequences for both bacteria & archaea

Zenodo now supports usage statistics

Using GitHub?

Zenodo in a nutshell

RODARE

Recent uploads

Is It Here/There Yet? - Real Life Experiences of Generating/Evaluating Extreme Data Sets Around the World

The Official "Green HPCG"

C++ & Python API for Scientific I/O with openPMD

RODARE Docs

RODARE now offers usage statistics!

RODARE

hasdai

about roadmap corpora providers users

hasdai is a pilot infrastructure for preservation of scientific annotation. It was developed by Data Futures with partners Bodleian Libraries and CERN after HDA Planary #11, and supports redelivery of existing research data from legacy technologies, as well as new projects, with long-term support using Core Trust Seal repositories. Annotation trials from additional research domains discussed at HDA Planary #12 are now being developed, and a Preservation of Scientific Annotation Working Group (PSA) is being established under the HDA Preservation Tools, Techniques and Policies IG. Brief information about the roadmap and current collaboration partners is available via the buttons above, or click the corpora button or the image below to browse.

ZENTRUM FÜR NACHHALTIGES FORSCHUNGSDATENMANAGEMENT

HOME UPLOAD LOGIN

Recent uploads

3D-Aufnahme einer Grabplatte, Museum für Kunst und Gewerbe, Hamburg

Viele viele schöne Käferchen...

Recent activity of the FDM-Center

ZfDM Repository terms of service

California Institute of Technology Research Data Repository

Search 895 records

Data Sets Software Submit



Scientific
Information
Service

INVENIO RDM

Le référentiel clé en main de gestion des données de recherche

<https://inveniosoftware.org/products/rdm/>

Développé en partenariat avec plus de 25 organisations



Principes pour les infrastructures ouvertes (une sélection)

- **Open source**
- **Conduit par la communauté (=écouter les utilisateurs)**
- **Gouverné de manière transparente**
- **Prestation de services/décisions transparentes**
- **Financé durablement**

Avantage pour vous:

Vous et la communauté le pilotez ensemble, vous avez votre mot à dire, et vous contrôlez le budget durablement.

Comment créer une plateforme de données de recherche

- Gardez le processus ouvert et participatif
- Intégrez les commentaires et les besoins des utilisateurs (cela semble plus simple que ça ne l'est !).

La plateforme elle-même

- Rendez-la utilisable et focusses sur des interface et processus simples
- Utilisez une infrastructure ouverte et fiable (extensible, modifiable, personnalisable aux sous-groupes), par ex. Dataverse, DSpace, InvenioRDM, etc.
- **Contenu : aussi ouvert que possible, aussi fermé que nécessaire**

Frais:

- Coûts de mise en œuvre, par ex. frais de personnel ou par l'intermédiaire d'un service/d'une société de logiciels. InvenioRDM, Dataverse etc sont gratuits et open source.
- Coûts opérationnels à long terme : fonctionnement, développement d'améliorations et de personnalisation, coûts de conservation ou de modération (selon la configuration). Bénéfice : vous renforcez vous-même vos capacités.

Un peu de motivation : les chercheurs du monde entier utilisent les données ouvertes du LHC@CERN

HEP

PUBLISHED FOR SISSA BY SPRINGER

RECEIVED: May 17, 2019
REVISED: November 26, 2019
ACCEPTED: December 2, 2019
PUBLISHED: December 16, 2019

Testing non-standard sources of parity violation in jets at the LHC, trialled with CMS Open Data

Christopher G. Lester^a and Matthias Schott^{b,c}*

^a Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge, United Kingdom
^b Massachusetts Institute of Technology, Cambridge, U.S.A.
^c Institute of Physics, Johannes Gutenberg University, Staudingerweg, Mainz, Germany

E-mail: lester@hep.phy.cam.ac.uk, matthias.schott@cern.ch

ABSTRACT: The Standard Model violates parity, but only by mechanisms which are subtle to Large Hadron Collider (LHC) experiments (on account of the lack of initial polarisation or spin-sensitivity in the detectors). Nonetheless, new physical processes potentially violate parity in ways which are detectable by those same experiments. If sources of new physics occur only at LHC energies, they are instead by direct search. We probe the feasibility of such measurements using approximately 0.2fb^{-1} of data recorded in 2012 by the CMS collaboration and made public within the CMS Data Initiative. In particular, we test an inclusive three-jet event selection which is highly sensitive to non-standard parity violating effects in quark/gluon interactions. Our measurements, no significant deviation from the Standard Model is seen and various experimental limitations have been found. We discuss other ways that search non-standard parity violation could be performed, noting that these would be sensitive to very different sorts of models to those which our method would constrain. We hope our initial studies provide a valuable starting point for rigorous future analyses using full LHC datasets at 13 TeV with a careful and less conservative estimate of experimental uncertainties.

KEYWORDS: Exotics, Hadron-Hadron scattering (experiments), proton-proton scattering, CP violation, Jet physics

ARXIV EPRINT: [1904.11195](https://arxiv.org/abs/1904.11195)

OPEN ACCESS, © The Authors.
Article funded by SCOAP³.
[https://doi.org/10.1007/JHEP12\(2020\)17](https://doi.org/10.1007/JHEP12(2020)17)

MIT-CTP 4891

Exposing the QCD Splitting Function with CMS Open Data

Andrew Larkoski,^{1,2} Simone Marzani,² Jesse Thaler,^{2,3} Aashish Tripathi,^{2,3} and Wei Xue^{2,3}

¹ Physics Department, Reed College, Portland, OR 97026, USA
² University at Buffalo, The State University of New York, Buffalo, NY 14260-1500, USA
³ Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

The splitting function is a universal property of quantum chromodynamics (QCD) which describes how energy is shared between partons. Despite its ubiquitous appearance in many QCD calculations, the splitting function cannot be measured directly, since it is always always multiplied by a collinear singularity factor. Recently, however, a new jet substructure observable was introduced which asymptotes to the splitting function for sufficiently high jet energies. This provides a way to expose the splitting function through jet substructure measurements at the Large Hadron Collider. In this letter, we use public data released by the CMS experiment to study the 2-prong substructure of jets and test the $1 \rightarrow 2$ splitting function of QCD. To our knowledge, this is the first ever physics analysis based on the CMS Open Data.

Computing and Software for Big Science (2021) 5:15
<https://doi.org/10.1007/s41781-021-00060-4>

ORIGINAL ARTICLE

Analysis-Specific Fast Simulation at the LHC with Deep Learning

C. Chen¹ · O. Cerri² · T. Q. Nguyen² · J. R. Vilmar² · M. Pierini³

Received: 12 October 2020 / Accepted: 12 May 2021 / Published online: 9 June 2021
© The Author(s) 2021

Abstract We present a fast simulation application based on a deep neural network, designed to create large analysis-specific datasets. Taking as an example the generation of $W + \text{jet}$ events produced in $\sqrt{s} = 13\text{ TeV}$ proton-proton collisions, we train a neural network to model detector resolution effects as a transfer function acting on an analysis-specific set of relevant features, computed at generation level, i.e., in absence of detector effects. Based on this model, we propose a novel fast-simulation workflow that starts from a large amount of generator-level events to deliver large analysis-specific samples. The adoption of this approach would result in about an order-of-magnitude reduction in computing and storage requirements for the collision simulation workflow. This strategy could help the high energy physics community to face the computing challenges of the future High-Luminosity LHC.

Keywords Hadron Collider Physics · Fast Simulation · Deep Learning · High Energy Physics computing

Introduction At the CERN Large Hadron Collider (LHC), high-energy proton-proton (pp) collisions are studied to consolidate our understanding of physics at the energy frontier and possibly to search for new phenomena. While these studies are typically conducted according to a data driven methodology, synthetic data from simulated pp collisions are a key ingredient to a robust analysis development. Particle physicists

rely extensively on an accurate simulation of the physics processes under study, including a detailed description of the response of their detector to a given set of incoming particles. These large sets of synthetic data are typically generated with experiment-specific simulation software, based on the GEANT4 [1] library. Through Monte Carlo techniques, GEANT4 provides the state of the art in terms of simulation accuracy. The first two runs of the LHC highlighted the remarkable agreement between data and simulation, with discrepancies observed at the level of a few percent. On the other hand, running GEANT4 is demanding in terms of resources. As a consequence of this, delivering synthetic data at the pace at which the LHC delivers real data is one of the most challenging tasks for the computing infrastructures of the LHC experiments. It is then more and more common for LHC physics analyses to be affected by large systematic uncertainties due to the limited amount of simulated data. This is particularly true for precise measurements of Standard Model processes for which large datasets are already available today. In the future, with the high-luminosity LHC upgrade, this will become a serious problem for most of the LHC data analyses [2]. Our community is called to reduce the computing resources needed for central simulation workflows by at least one order of magnitude, not to jeopardize the accuracy gain expected when operating the LHC at a high luminosity.

* Springer

MIT-CTP 5129

Exploring the Space of Jets with CMS Open Data

Patrick T. Komiske,^{1,2,3} Radha Mastandrea,¹ Eric M. Metodiev,^{1,2,3} Preisha Nair,^{1,2,3} and Jesse Thaler^{1,2,3}*

¹ Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
² Department of Physics, Harvard University, Cambridge, MA 02138, USA

We explore the metric space of jets using public collider data from the CMS experiment. Starting from 2.3fb^{-1} of proton-proton collisions at $\sqrt{s} = 7\text{ TeV}$ collected at the Large Hadron Collider in 2011, we isolate a sample of 1,600,084 central jets with transverse momentum above 375 GeV. To validate the performance of the CMS detector in reconstructing the energy flow of jets, we compare the CMS Open Data to corresponding simulated data samples for a variety of jet kinematic and substructure observables. Even without detector unfolding, we find very good agreement for track-based observables after using charged hadron subtraction to mitigate the impact of pileup. We perform a range of novel analyses, using the “energy mover’s distance” (EMD) to measure the pairwise difference between jet energy flows. The EMD allows us to quantify the impact of detector effects, visualize the metric space of jets, extract correlation dimensions, and identify the most and least typical jet datasets and hundred giga-

PHYSICAL REVIEW D **100**, 015021 (2019)

COO

Searching in CMS open data for dimuon resonances with substantial transverse momentum

Cari Cesarani,^{1,2} Yotam Soreq,^{3,4} Matthew J. Strassler,^{1,2} Jesse Thaler,^{1,2,3,4} and Wei Xue^{2,3}

¹ Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA
² Theoretical Physics Department, CERN, 1211 Geneva 23, Switzerland
³ Department of Physics, Technion, Haifa 32000, Israel
⁴ Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Received 8 March 2019; published 16 July 2019)

We study dimuon events in 2.11fb^{-1} of 7 TeV pp collisions, using CMS Open Data, and search for a narrow dimuon resonance with moderate mass (14–66 GeV) and substantial transverse momentum (p_T). Applying dimuon p_T cuts of 25 and 60 GeV, we explore two overlapping samples: one with isolated muons, and one with prompt muons without an isolation requirement. Using the latter sample requires information about detector effects and QCD backgrounds, which we obtain directly from the CMS Open Data. We present model-independent limits on the product of cross section, branching fraction, acceptance, and efficiencies. These limits are stronger, relative to a corresponding inclusive search without a p_T cut, by factors of as much as 9. Our “ p_T -enhanced” dimuon search strategy provides improved sensitivity to models in which a new particle is produced mainly in the decay of something heavier, as could occur, for example, in decays of the Higgs boson or of a TeV-scale top partner. An implementation of this method with the current 13 TeV data should improve the sensitivity to such signals further by roughly an order of magnitude.

DOI: [10.1103/PhysRevD.100.015021](https://doi.org/10.1103/PhysRevD.100.015021)

I. INTRODUCTION The CERN Open Data portal [1] aims to make data from the Large Hadron Collider (LHC) publicly available as a long-term archive, with the first research-grade data from the CMS experiment released in 2014 [2]. In order to identify any issues that might interfere with their use by physicists of the future, it is important that open data frameworks be tested today. There are good scientific motivations to make use of this resource [3]. Open data makes it possible for scientists outside of the LHC collaborations to study standard model (SM) processes that are not well modeled by Monte Carlo (MC) generators, such as rare QCD backgrounds. Together with detector-simulated samples, open data also makes it possible to test event analysis strategies that rely on a detailed understanding of detector effects. The value of the CMS Open Data for exploratory studies of QCD has been demonstrated in Refs. [4,5]; see Refs. [6–8] for machine-learning studies on detector-simulated CMS samples. Refs. [9–11] for QCD studies on archival ALEPH data, and Ref. [12] for a diphoton analysis with public LHC data. In this paper, we report the first utilization of the CMS Open Data in a search for beyond the standard model (BSM) phenomena. We seek a new particle Y that decays promptly to dimuon pairs ($\mu^+ \mu^-$) and is typically produced with substantial transverse momentum (p_T). Our analysis is based on 2.11fb^{-1} of 7 TeV center-of-mass pp collision events recorded by the CMS experiment during the first part of 2011 and made public through the CERN Open Data portal [13]. We perform a narrow resonance search in the dimuon mass range $m_{\mu\mu} \in [14, 66]\text{ GeV}$ and study the effect of modest cuts on p_T , namely, $p_T^{\mu\mu} > 25\text{ GeV}$ and 60 GeV ; this approach (which we refer to as p_T -enhanced) could be applied to larger p_T values as well, or alternatively to a cut on the Y boost factor p_T^Y/m_Y . This type of search strategy was suggested some time ago [14], as one of several unconventional approaches for finding low-mass dilepton and diphoton resonances [15], but to our knowledge has never been carried out as a public analysis by the LHC collaborations. For this reason, the mass and p_T regime we cover is relatively unexplored. Moreover, our

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

2470-0010/2019/100(1)/015021(20) 015021-1 Published by the American Physical Society

MIT-CTP 5185

The Hidden Geometry of Particle Collisions

omiske, Eric M. Metodiev, and Jesse Thaler

Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
komiske@mit.edu, metodiev@mit.edu, jthaler@mit.edu

We establish that many fundamental concepts and techniques in quantum field collider physics can be naturally understood and unified through a simple new language. The idea is to equip the space of collider events with a metric, from geometric objects can be rigorously defined. Our analysis is based on the energy flow, which quantifies the “work” required to rearrange one event into another, which operates purely at the level of observable energy flow information, allows definition of infrared and collinear safety and related concepts. A number of collider observables can be exactly cast as the minimum distance between various manifolds in this space. Jet definitions, such as exclusive cone and combination algorithms, can be directly derived by finding the closest few-particle to the event. Several area- and constituent-based pileup mitigation strategies expressed in this formalism as well. Finally, we lift our reasoning to develop relations between theories, which are treated as collections of events weighted by \mathcal{L} . In all of these various cases, a better understanding of existing methods in our language suggests interesting new ideas and generalizations.

Merci beaucoup!

sunje.dallmeier-tiessen@cern.ch



Scientific
Information
Service

scientific-info.cern