ONTOFORCE

Data Source Updates

2023-01-31

MAIN OFFICE
Moutstraat 108
9000 Ghent
Belgium

Ontoforce.com
+32 9 396 80 07
info@ontoforce.com

ONTOFORCE

Table of contents

Introduction

This document shows the current update state of the different data sources maintained by ONTOFORCE as part of our federated data offering.

After downloading the newest data files these files need to be processed through our integrated

ETL process. The data made available along this document was created by:

□ ETL run started on 2023-01-25.

□ ETL results published on 2023-01-31.

Data Source Update Table

EOL = End Of Life, These sources are finished projects that do not expect new data updates.

Manual = Not in update automation framework but manually refreshed periodically.

N/A = Update automation framework is in development. Manual updates may occur.

Data Source	Completeness	Last Modified	Last Refresh Check	Refresh Schedule
1000 Genomes Project	Partial	2019-02-27	EOL	EOL
7th Framework Programme	Full	2021-03-19	EOL	EOL
Anatomical Therapeutic Chemical	Full	2018-10-01	N/A	N/A
Antibody Registry	Full	2022-04-09	2023-01-23	weekly
BRCA Exchange	Full	2022-09-01	2023-01-24	weekly
Cellosaurus	Full	2022-12-15	2023-01-19	weekly
ChEBI	Full	2023-01-01	2023-01-18	weekly
ChEMBL	Full	2022-08-18	2023-01-22	weekly
ClinicalTrials.gov	Full	2023-01-20	2023-01-23	weekly
ClinVar Record	Full	2021-04-01	Manual	Manual
Cooperative Patent Classification	Full	2022-01-25*	2023-01-22	weekly
CORDIS	Full	2021-03-17	N/A	N/A
DailyMed	Full	2021-12-14*	2023-01-24	weekly
dbSNP	Partial	2020-01-28	2023-01-24	weekly
dbVar	Full	2022-12-08	2023-01-20	weekly
DrugCentral	Full	2022-09-13	2023-01-24	weekly
Eagle-i	Full	2021-12-14	EOL	EOL
ENZYME	Full	2022-12-14	2023-01-22	weekly
EPO	Partial	2022-08-17	2023-01-19	weekly
EudraCT	Full	2023-01-25	2023-01-24	weekly
Evidence Ontology	Full	2022-12-04	2023-01-18	bi-weekly

Experimental Factor Ontology	Full	2023-01-16	2023-01-20	weekly
ExPORTER	Full	2021-10-03*	2023-01-20	weekly
Federal Information Processing Series (FIPS)	Full	2021-09-28*	2023-01-24	weekly
FRIS	Full	2021-03-19*	2023-01-20	weekly
Gene Ontology	Full	2023-01-18	2023-01-19	weekly
GeneRIF	Full	2023-01-18	2023-01-19	weekly
GHR	Full	2018-10-01	N/A	N/A
gnomAD	Partial	2020-03-02	EOL	EOL
GRID	Full	2021-06-15	EOL	EOL
GUDID	Full	2023-01-01	2023-01-18	weekly
HMDB	Full	2021-11-17	2023-01-23	weekly
HomoloGene	Full	2014-05-06	2023-01-24	weekly
Horizon 2020	Full	2021-09-17	2021-10-06	weekly
Human Phenotype Ontology	Full	2022-06-11	2023-01-18	bi-weekly
HSDB	Full	2019-10-28	2023-01-24	weekly
HUGO Gene Nomenclature Committee	Full	2023-01-20	2023-01-23	weekly
Human Disease Ontology	Full	2022-09-29	2023-01-18	bi-weekly
ICD10-CM	Full	2022-06-09	2023-01-23	weekly
ICD9-CM	Full	2018-10-01	EOL	EOL
IMSR	Full	2023-01-17	2023-01-19	weekly
InterPro	Full	2022-08-03	2023-01-24	weekly
ISO Countries	Full	2020-01-01	Manual	Manual
IUPHAR Compendium	Full	2022-03-31*	2023-01-22	weekly
Journals	Full	2023-01-19	2023-01-20	weekly
MeSH	Full	2023-01-19	2023-01-20	weekly

Mondo Disease Ontology	Full	2023-01-04	2023-01-23	weekly
National Drug Code	Full	2018-10-01	N/A	N/A
NCBI Gene	Full	2023-01-20	2023-01-23	weekly
NCBI Taxonomy	Full	2023-01-20	2023-01-23	weekly
Nomenclature of territorial units for statistics (NUTS)	Full	2019-07-31	2023-01-23	weekly
ONTOFORCE	Full	2023-01-21	2023-01-23	weekly
Open Humans	Full	2018-10-01	N/A	N/A
Open Targets	Full	2021-02-28*	2021-10-31	weekly
OpenCage Geocoder	Full	2021-04-21	Manual	Manual
OpenFDA Drug Adverse Events	Full	2023-01-16	2023-01-24	weekly
ORCID	Full	2022-10-07	2022-10-11	weekly
Orphanet Rare Disease Ontology	Partial	2022-06-14	2023-01-23	weekly
PROVEAN	Full	2015-01-30*	2023-01-24	weekly
PubChem	Partial	2023-01-19	2023-01-20	weekly
PubMed	Partial*	2023-01-24	2023-01-24	daily
PubTator	Full	2022-12-26	2023-01-23	weekly
Reactome	Full	2021-07-08*	2022-06-17	weekly
RxIMAGE	Full	2018-10-01	N/A	N/A
RxNorm	Full	2021-03-01*	2021-09-08	weekly
SCImago Journal & Country Rank	Full	2022-06-30	2023-01-22	weekly
Semanticscience Integrated Ontology	Full	2021-07-06	2023-01-18	bi-weekly
SNOMED CT	Full	2022-09-01	2023-01-19	weekly
SureChEMBL	Partial	2021-08-09	2023-01-18	weekly
SwissVar	Full	2020-06-17*	2023-01-24	weekly
Uberon	Full	2021-11-28	2022-03-16	bi-weekly
UNII	Full	2022-12-15	2023-01-22	weekly

UniProt Knowledgebase	Partial	2022-12-14	2023-01-18	weekly
WHO: International Clinical Trials Registry Platform	Full	2023-01-17	2023-01-20	weekly
Wikidata	Partial	2022-10-13	2023-01-18	bi-weekly
WikiPathways	Full	2023-01-25	2023-01-24	weekly
Wikipedia	Partial	2022-10-27	2023-01-12	bi-weekly

* These sources were flagged in QC for either having changed the way they can be downloaded or the downloaded files no longer fit the expected data model. Therefore, they are not updated. Actions will be taken to restore the updateability of this data.

Data Source Completeness

Completeness level in this document is in regard to the integration of all entities provided by the given data source. This means that every entity should be findable in the federation endpoint. Completeness here does not mean all annotations on the entity provided by the source are in the federation endpoint. This means that certain search texts may not find an entity even though it is present in the platform. A detailed description of our data model of all annotations that can be expected can be found in our RDS package documentation here.

1000 genome project partial data

The DISQOVER federation endpoint includes all variations from 1000 genomes VCF file, with PASS filter and is covering a subset of 202 genes that are somatically mutated and causally implicated in human cancer, from the COSMIC database.

gnomAD partial data

The DISQOVER federation endpoint includes all variations from gnomAD VCF file, with PASS filter and is covering a subset of manually curated genes that are somatically mutated and causally implicated in human cancer from the COSMIC database.

dbSNP partial data

The DISQOVER federation endpoint includes all variants with a cross-reference from SwissVar and Provean. In addition, it covers all variants of a manually curated subset of genes that are somatically mutated and causally implicated in human cancer from the COSMIC database.

EPO partial data

The DISQOVER federation endpoint includes EPO patents that are associated with the CPC codes:

"A61K", "C12Q", "C07" and "Y10S514/00"

Orphanet Rare Disease Ontology partial data

Orphanet Rare Disease Ontology's open data availability has dwindled since the first time it was integrated within the federation platform. The federation endpoint contains the Orphanet Rare Disease Ontology entities still available as open data.

For customers who have a license to the full data available in Orphanet Rare Disease Ontology pipelines for local integration are made available in the knowledge base.

PubChem partial data

In the DISQOVER federation endpoint PubChem is used for chemical equivalency mapping between the chemicals exposed by other sources ingested in DISQOVER. Synonyms from PubChem are also added to allow broader semantic searching of Chemicals.

PubChem Chemicals that do not have equivalencies with Chemicals from other sources are not in the federation endpoint.

PubMed partial data*

Although we provide all information from the PubMed downloadable files there are some PMC articles that cannot be found in DISQOVER like PMID:27253008 or PMID:27170958. These articles are generated from the NIH-Books dataset and are not provided by NIH in the PubMed dataset.

SureChEMBL partial data

In the DISQOVER federation endpoint SureChEMBL is used to create links between patents and chemicals in ChEMBL. The relations for which the Patent and Chemical are available in the DISQOVER endpoint are being exposed

UniProt Knowledgebase partial data

In the DISQOVER federation endpoint the expert curated subset of proteins SwissProt is made available.

Wikidata partial data

In the DISQOVER federation endpoint organizations and their related location data from Wikidata are being exposed.

The subset of organizations exposed was determined by querying the source for every organization listed as one of the following types:

- company
- business enterprise
- public company
- private company
- corporation
- holding company
- limited liability company
- private limited liability company
- privately held company
- limited company
- public limited company
- benefit corporation
- joint-stock company
- association
- public university
- university
- private not-for-profit educational institution
- college
- campus
- high school
- hospital
- government agency

- junior college
- facility
- School

Wikipedia partial data

In the DISQOVER federation endpoint Organizations and their related location data from Wikipedia are being exposed.

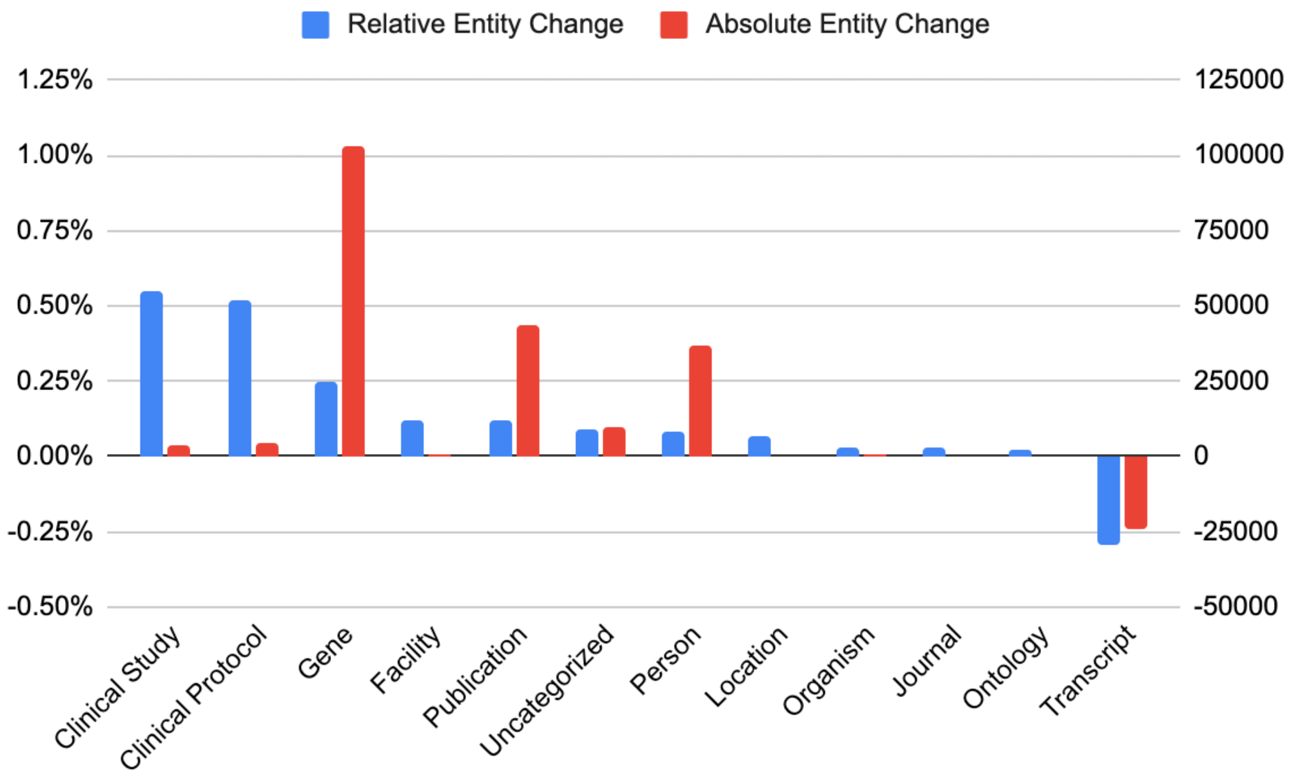
The subset of organizations exposed was determined by the listings on the following Wikipedia pages:

- List of largest private companies in the United Kingdom
- List of largest biotechnology & pharmaceutical companies
- Fortune Global 500
- Euro Stoxx 50
- AEX index
- List of largest chemical producers
- List of largest manufacturing companies by revenue
- Dow Jones Global Titans 50
- S&P Latin America 40
- S&P/ASX 50
- Austrian Traded Index
- BEL 20
- S&P/TSX 60
- S&P/TSX Composite Index
- Índice de Precio Selectivo de Acciones
- CSI 300 Index
- OMX Copenhagen 20
- OMX Helsinki 25
- CAC 40
- DAX
- FTSE/Athex Large Cap
- BSE SENSEX
- ISEQ 20
- TA-125 Index
- FTSE MIB
- LuxX Index
- NZX 50 Index

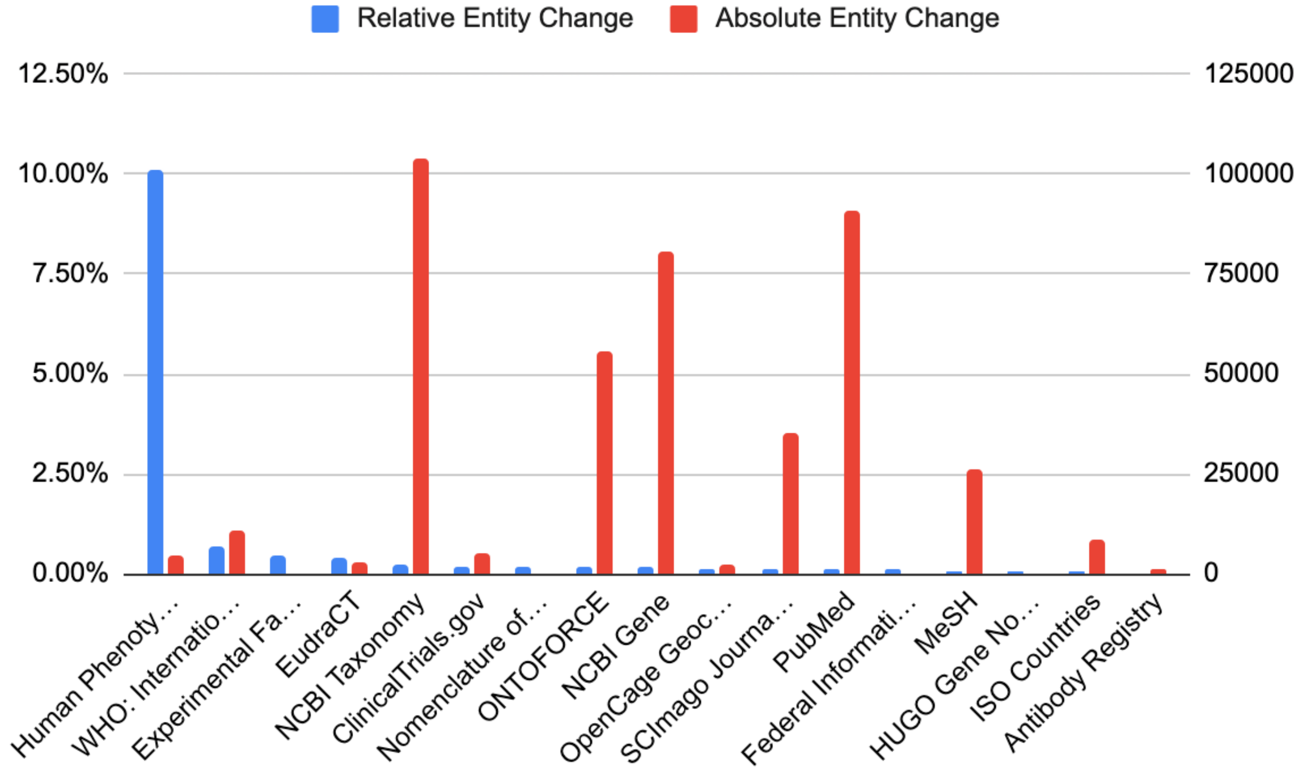
- FTSE 100 Index

Data Change Graphs

The below chart shows the relative and absolute amount of added or removed entities per data type in this latest data update in blue and red respectively.



The below chart shows the relative and absolute amount of added or removed entities per source in this latest data update in blue and red respectively.



Data Sources or Data Types not present in the above charts had no significantly changed counts.

This results in the following total Data Type counts

Active Substance	10.7k
Adverse Event	15.6M
Antibody	1.98M
Assay	19.8M
Biospecimen	4.17k
Cell Line	383k
Chemical	2.97M
Clinical Protocol	818k
Clinical Study	727k
Disease	171k
Drug Treatment	31.4M
Enzyme	12.2k
Facility	875k
Gene	42.0M
Homology	44.2k
Journal	43.2k
Location	112k
Medical Device	4.59M
Medicine	703k
Model Organism	84.9k
Ontology	951k
Organism	2.49M
Organization	377k
Patent	6.33M
Pathway	23.6k
Person	44.3M
Plasmid	29.6k
Project	3.22M
Protein	612k
Publication	35.8M
Transcript	8.12M
Variant	22.7M