Wednesday 3 Feb 2021, 10:00 → 12:10 Europe/Copenhagen

Relevance and Importance of splitting META-DATA from DATA

Erik Schultes, PhD

International Science Coordinator GO FAIR Foundation Leiden Center for Data Science



erik.schultes@go-fair.org https://www.go-fair.org http://orcid.org/0000-0001-8888-635X

FAIRification STEP 2 – FAIR principle F3

Wednesday 3 Feb 2021, 10:00 → 12:10 Europe/Copenhagen

10:00 → 10:10	Welcome and Introduction Speaker: Mr Bert Meerman (GFF)	③10m
10:10 → 10:25	FAIR assessment results Speaker: Dr Andreas Jaunsen (NeIC)	③ 15m
10:25 → 10:55	Relevance and Importance of splitting META-DATA from DATA Speaker: Dr Erik Schultes (GO-FAIR)	() 30m
10:55 → 11:05	BREAK	() 10m
11:05 → 11:20	Some recommendations for the practical, machine-friendly implementation of the FAIR F3 principle Speaker: Dr Robert Huber (Univ of Bremen)	③ 15m
11:20 → 11:35	File level identification support in Dataverse/DataverseNO Speaker: Mr Philipp Conzett (UiT / The Artic University of Norway)	③ 15m
11:35 → 11:55	Q & A Speaker: Mrs Josefine Nordling (CSC)	() 20m
11:55 → 12:00	Call to Action and Close Speaker: Mr Bert Meerman (GFF)	() 5m

FAIRification STEP 2 – FAIR principle F3

Wednesday 3 Feb 2021, 10:00 → 12:10 Europe/Copenhagen

Acknowledgments



Nikola Vasilijevic Technical University of Denmark



Mark Wilkinson Universidad Politécnica de Madrid

2016

nature > scientific data > comment > article

SCIENTIFIC DATA

Comment | OPEN | Published: 15 March 2016

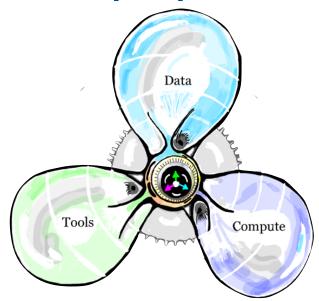
The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons Song - Show fewer authors

Scientific Data 3, Article number: 160018 (2016) Download Citation 🚽

https://www.nature.com/articles/sdata201618

Data and services that are findable, accessible, interoperable, re-usable both for machines and for people.



Relevance and Importance of splitting META-DATA from DATA

Box 2 | The FAIR Guiding Principles

https://www.nature.com/articles/sdata201618

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- 13. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

Relevance and Importance of splitting META-DATA from DATA

Box 2 | The FAIR Guiding Principles https://www.nature.com/articles/sdata201618 To be Findable: F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource To be Accessible: A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure where necessary A2. metadata are accessible, even when the data are no longer available To be Interoperable: I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. 12. (meta)data use vocabularies that follow FAIR principles 13. (meta)data include gualified references to other (meta)data To be Reusable: R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

self-documenting data

https://www.unidata.ucar.edu/software/netcdf/documentation/NUG/_best_practices.html

NetCDF files, example:

01

long-range-WindScanner-measurements.nc

21 MB Binary

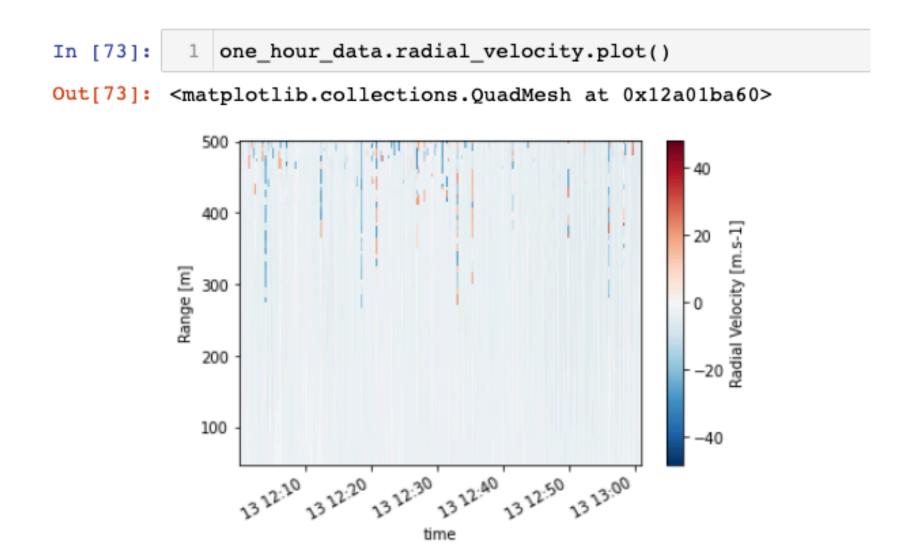
- 1 hour of 5 Hz measurements
- 1.5 milions of samples per variable

The embedded metadata...

```
<xarray.Dataset>
Dimensions:
                                (los_no: 180, range: 91, time: 14317)
Coordinates:
 * range
                                (range) int32 50 55 60 65 70 ... 485 490 495 500
  * time
                                (time) datetime64[ns] 2018-08-13T12:00:01.005000114 ... 2018-08-13T13:00:59.855000019
  * los no
                                (los no) int32 1 2 3 4 5 ... 176 177 178 179 180
                                (range, time) int64 1 1 1 1 1 1 1 ... 1 0 0 1 1 1
    mask
Data variables:
    radial velocity
                                (range, time) float32 ...
                                (range, time) float32 -17.398 ... -23.877
    CNR
    spectral broadening
                                (range, time) float32 ...
    los number
                                (time) int32 142 143 144 145 146 ... 55 56 57 58
    azimuth angle
                                (time) float32 117.541 121.105 ... 121.13 117.57
    elevation angle
                                (time) float32 26.86 26.548 ... 10.418 10.557
    azimuth angle calculated
                                (los no) float64 160.0 158.6 ... 115.8 112.1
                                (los_no) float64 2.047 2.113 ... 29.82 30.01
    elevation angle calculated
Attributes:
    title:
                          Long-range WindScanner multi-rotor wake measurements
    authors:
                          Nikola Vasiljević
    summary:
                          Data set containing 5 Hz radial velocity measuremen...
    data owner:
                          DTU Wind Energy
    DOI:
                          10.11583/DTU.9896459
                          CC BY 4.0
    licence:
    logbook:
    averaging period:
                          200 ms
    temporal resolution:
                          200 ms
    site:
                          Risoe campus
    instrument:
                          Long-range WindScanner system
    instrument id:
                          https://doi.org/10.3390/rs8110896
                          55.686014
    latitude:
    longitude:
                          12.097596
    height:
                          0
```

```
<xarray.DataArray 'radial velocity' (range: 91, time: 14317)>
[1302847 values with dtype=float32]
Coordinates:
 * range
           (range) int32 50 55 60 65 70 75 80 ... 470 475 480 485 490 495 500
 * time
           (time) datetime64[ns] 2018-08-13T12:00:01.005000114 ... 2018-08-13T13:00:59.855000019
           mask
Attributes:
   standard name: radial velocity of scatterers toward instrument
                 http://data.windenergy.dtu.dk/controlled-terminology/wind...
   concept id:
   short name:
                 Vrad
   long name:
                 Radial Velocity
   units:
                 m.s-1
```

... makes for easy processing



Why decouple data from metadata?

- Metadata can often be lightweight compared to data *Kb instead of Gb or Tb*
- Cost of maintaining metadata lower than data Metadata can be more persistant than data
 A2
- Convergence on FAIR Implementations

Can use same tech stack for building/serving metadata

Higher chance of converging to a fewer metadata standards (templates) vs jungle of "standards"

We can use the same metadata format **1** *RDF is a way to go in any of its serializations, such as JSON-LD, TTL, etc.*

Leads to machine-actionablity even though data will always be depended on the underlaying data standards (not all data is SPARQL friendly)

• Separation of concerns and division of responsibilitie Often researchers can only publish metadata (data access remains restricted).



https://data.dtu.dk/articles/online_resource/Perdig_o-2017_multi-lidar_flow_mapping_over_the_complex_terrain_site/7228544

Perdigão-2017: multi-lidar flow mapping over the complex	terrain site	
Cite Share + Collect	1507 11 5 views downloads citations	
Version 2 ✓ Online Resource posted on 11.10.2019, 11:05 by Robert Menke, Jakob Mann, Nikola Vasiljevic		
This dataset has been recorded by eight long-range WindScanners that were deployed during the Perdigão 2017 campaign. For the campaign which took place in central Portugal near the village of Perdigão eight scanners were located on two mountain ridges that run in parallel for about 2 km. The scanners carried out three synchronized scanning scenarios: transect scans at three positions perpendicular to the ridge with two scanners at two transects and one transect with four scanners; a scan following transects 80m above the ridges; and virtual mast scans at four locations.	CATEGORIES Atmospheric Sciences Meteorology Renewable Power and Energy Systems Engineering (excl. Solar Cells)	
This is an original dataset. It represents Level 2.3 data product in the FAIR lidar data schematics, that is geo-located radial velocities stored in NetCDF files with dimensions of time, range and line-of-sight number.	KEYWORDS scanning lidar wind lidar multi-lidars WindScanner	Datasets t Filter files or fol / Perdigão Catalog / U
Consult a list of references for more details about : (1) WindScanner (https://doi.org/10.3390/rs8110896) (2) Perdigão site and prequel experiment from 2015 (https://doi.org/10.5194/amt-10-3463-2017) (3) FAIR lidar data standard (http://e-	long-range WindScanner recirculation zone	Name ^ 20170327165000.nc 20170327172000.nc 20170327175000.nc
windlidar.windenergy.dtu.dk/documents/report.pdf) (4) Perdigão data repository (https://perdigao.fe.up.pt/)	CC BY 4.0	20170327182000.nc
Consult link in Related publications which holds information on the paper that presents a part of the Perdigão-2017 dataset.	EXPORTS Select an option 🔻	20170327192000.nc
The data can be downloaded from the webpage:		20170327202000.nc

https://bit.ly/2APFISJ

5 citations ⊭			
ergy Solar			
		Search equipment and p	people. Q
	Datasets		
lar	t Filter files or folders		٩
ner	/ Perdigão Catalog / Upper Air / Lidar / DTU Scanning Lidar Data / netcdf / ridge / WS5		Description
	Name ^		The datasets were produced, from 1st N to 15th June, by the following institution
	20170327165000.nc		DLR, DTU, ENERCON, INEGI, UCAR (AI UC, EOL, IPMA, ISFS, NCAS, ND, OU),
	■ 20170327172000.nc		WindsforS Institution: U.PORTO Data Catalog for
	20170327175000.nc		Perdigão
	20170327182000.nc		Details
	■ 20170327192000.nc		Size: 1.538 Mbytes
	■ 20170327195000.nc		Created at: 2019-07-26 10:08
	₽ 20170327202000.nc		Access
	₽ 20170327205000.nc		нттр
	₽ 20170327212000.nc		OPENDAP
	2,863 total	2 3 4 5 > ▶	NCML

<resource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://datacite.org/</pre> schema/kernel-4" xsi:schemaLocation="http://datacite.org/schema/kernel-4 http:// schema.datacite.org/meta/kernel-4.3/metadata.xsd"> <identifier identifierType="DOI">10.11583/DTU.7228544.v2</identifier> <creators> <creator> <creatorName nameType="Personal">Robert Menke</creatorName> <givenName>Robert</givenName> <familyName>Menke</familyName> </creator> <creator> <creatorName nameType="Personal">Jakob Mann</creatorName> <qivenName>Jakob</qivenName> <familyName>Mann</familyName> </creator> <creator> <creatorName nameType="Personal">Nikola Vasiljevic</creatorName> <givenName>Nikola</givenName> <familyName>Vasiljevic</familyName> </creator> </creators> <titles><title>Perdigão-2017: multi-lidar flow mapping over the complex terrain site</title></</pre> titles> <publisher>Technical University of Denmark</publisher> <publicationYear>2019</publicationYear> <subjects> <subject>scanning lidar</subject> <subject>wind lidar</subject> <subject>multi-lidars</subject> <subject>WindScanner</subject> <subject>long-range WindScanner</subject>

F1 DOI: <u>10.1038/sdata.2016.18</u>

NIH	National Library of National Center for Biotec	Medicine Chnology Information	Log in
Pub	Med.gov		
Search	PubMed		Search

Pub Med.gov	
10.1038/sdata.2016.18	X Search
Advanced Create alert Create RSS	User Guide
Found 1 result for <i>10.1038/sdata.2016.18</i>	Save Email Send to Display options

> Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18.

The FAIR Guiding Principles for scientific data management and stewardship

Mark D Wilkinson ¹, Michel Dumontier ², I Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton ³, Arie Baak ⁴, Niklas Blomberg ⁵, Jan-Willem Boiten ⁶, Luiz Bonino da Silva Santos ⁷, Philip E Bourne ⁸, Jildau Bouwman ⁹, Anthony J Brookes ¹⁰, Tim Clark ¹¹, Mercè Crosas ¹², Ingrid Dillo ¹³, Olivier Dumon, Scott Edmunds ¹⁴, Chris T Evelo ¹⁵, Richard Finkers ¹⁶, Alejandra Gonzalez-Beltran ¹⁷, Alasdair J G Gray ¹⁸, Paul Groth, Carole Goble ¹⁹, Jeffrey S Grethe ²⁰, Jaap Heringa ²¹, Peter A C 't Hoen ²², Rob Hooft ²³, Tobias Kuhn ²⁴, Ruben Kok ²¹, Joost Kok ²⁵, Scott J Lusher ²⁶, Maryann E Martone ²⁷, Albert Mons ²⁸, Abel L Packer ²⁹, Bengt Persson ³⁰, Philippe Rocca-Serra ¹⁷, Marco Roos ³¹, Rene van Schaik ³², Susanna-Assunta Sansone ¹⁷, Erik Schultes ³³, Thierry Sengstag ³⁴, Ted Slater ³⁵, George Strawn, Morris A Swertz ³⁶, Mark Thompson ³¹, Johan van der Lei ³⁷, Erik van Mulligen ³⁷, Jan Velterop ³⁸, Andra Waagmeester ³⁹, Peter Wittenburg ⁴⁰, Katherine Wolstencroft ⁴¹, Jun Zhao ⁴², Barend Mons ⁴³ 26 37





SHARE

PAGE NAVIGATION

Affiliations + expand

PMID: 26978244 PMCID: PMC4792175 DOI: 10.1038/sdata.2016.18

Title & authors

DataCite Schema Support DataCite Metadata Schema is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions.

Metadata Schema 4.3

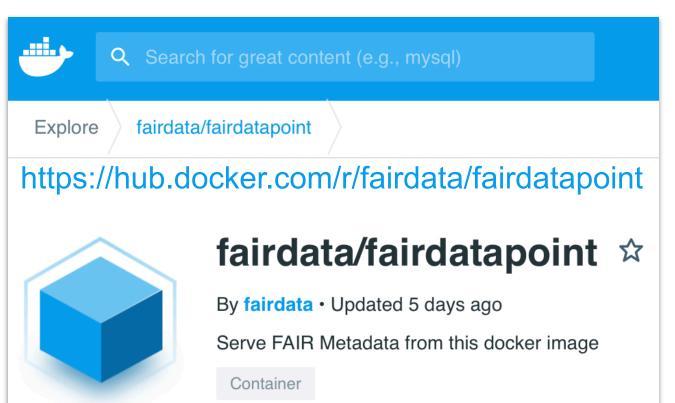
Released 16 Aug 2019. Changes in this version include:

- Addition of optional "affiliationIdentifier", "affiliationIdentifierScheme", and "schemeURI" for affiliation
- Addition of optional "schemeURI" for funderIdentifier
- Addition of "ROR" to allowed values for funderIdentifierType

Yes please feel free to share our plans. [F3] will be implemented through our metadata working group who are currently getting our schema 4.4 released and then they will move onto these discussions.

- Matt Buys, Executive Director of Datacite, February 2 2021

What does F3 look like?



FDP:

- Metadata publication platform
- LOD
- Follows DCAT
- Adheres to F3
- Adheres to I3 (qualified links)

	Repository	Catalog	Data	Distribution
rdf:type	x	x	х	x
dct:title	x	x	x	x
dct:hasVersion		x	х	x
dct:description	x	x	х	x
dct:publisher	x	x	х	
dct:language	x	x	х	
dct:license	x	x	х	x
dct:issued		x	х	x
dct:modified		x	х	x
dct:conformsTo	x	x	х	x
dct:rights	x	x	x	x
dct:accessRights	x	x	x	x
dct:hasPart		x		
dct:isPartOf		x	x	x
dcat:themeTaxonom Y		x		
dcat:contactPoint	x			
dcat:keyword	x			
dcat:theme	x			
dcat:endPointURL	x			
dcat:endPointDescri ption	x			
fdp:metadataldentifi er	x	x	x	x
fdp:metadatalssued	x	x		
fdp:metadataModifi ed	x	x	x	
fdp-o:startDate	x			
fdp-o:endDate	x			
fdp-o:repositoryLang uage	x			
fdp-o:hasSoftwareVe rsion	x			
fdp-o:conformsToFd pSpec	x			
ldp:DirectContainer	x			
foaf:homepage		х		
rdfs:label			х	
dcat:distribution			х	
dcat:theme			х	
dcat:contactPoint			х	
dcat:keyword	F	3	х	
dcat:landingPage			х	
dcat:accessURL				x
dcat:downloadURL				x
dcat:mediaType				x
dcat:format				x
dcat:byteSize				x

F/IR FAIR Data Point

KIU FAIR Data Point

This FAIR Data Point contains the metadata of datasets and other artefacts related to the COVID-19 pandemic and other health data in Kampala International University and the rest of Uganda.

Catalogs

Covid-19 Case Report Form

Covid-19 case report forms following WHO standard

Issued 13-07-2020 Modified 13-07-2020

Migrant Media Reports

These are national and international media reports published during the covid-19 crisis.

Issued 13-07-2020 Modified 13-07-2020

Metadata Issued 01-07-2020	Metadata Modified 13-07-2020
Version	
1.0	
Language	
<u>en</u>	
License	
<u>cc-by-nc-nd3.0</u>	
Institution	
<u>kiu.ac.ug</u>	
Start date	
03-06-2020	
Last update	
13-07-2020	
Institution country	
<u>Q1036</u>	

Log in

FDP F/IR FAIR Data Point

KIU FAIR Data Point

This FAIR Data Point contains the metadata of datasets and other artefacts related to the COVID-19 pandemic and other health data in Kampala International University and the rest of Uganda.

Catalogs

Covid-19 Case Report Form

Covid-19 case report forms following WHO standard

Issued 13-07-2020 Modified 13-07-2020

Migrant Media Reports

These are national and international media reports published during the covid-19 crisis.

Issued 13-07-2020 **Modified** 13-07-2020

Versien	
Version	
1.0	
Language	
<u>en</u>	
License	
<u>cc-by-nc-nd3.0</u>	
Institution	
<u>kiu.ac.ug</u>	
Start date	
03-06-2020	
Last update	

FAIR Data Point



Migrant Media Reports

These are national and international media reports published during the covidcrisis.

Datasets

Media Report Dataset

These datasets are media reports concerning events related to the impact of covid-19 crisis or immigrants in Tunisia.

Q57979909 Q38926

Issued 13-07-2020 Modified 13-07-2020

19	Metadata Issued 13-07-2020	Metadata Modified 13-07-2020
	Version 1.0	
r	Language <u>en</u>	
	License	
	lssued 13-07-2020	
	Modified 13-07-2020	
	Theme taxonomy • <u>Q38926</u>	

Q57979909

FAIR FAIR Data Point

Dataset <u>Data Point</u> / <u>Migrant Media Reports</u> / Media Report Dataset

Media Report Dataset

These datasets are media reports concerning events related to the impact of covid-19 crisis on immigrants in Tunisia.

Distributions

Media Reports listed in a google document

These are internet links to media reports listed in a google document.

Issued 13-07-2020 Modified 13-07-2020 Media Type https://w3id.org/mediatype/text/html

Metadata Issued 13-07-2020	Metadata Modified 13-07-2020
Version	
1.0	
Language	
<u>en</u>	
License	
Issued	
13-07-2020	
Modified	
13-07-2020	
Theme	
Q38926	

Q57979909

Log in

F/IR FAIR Data Point

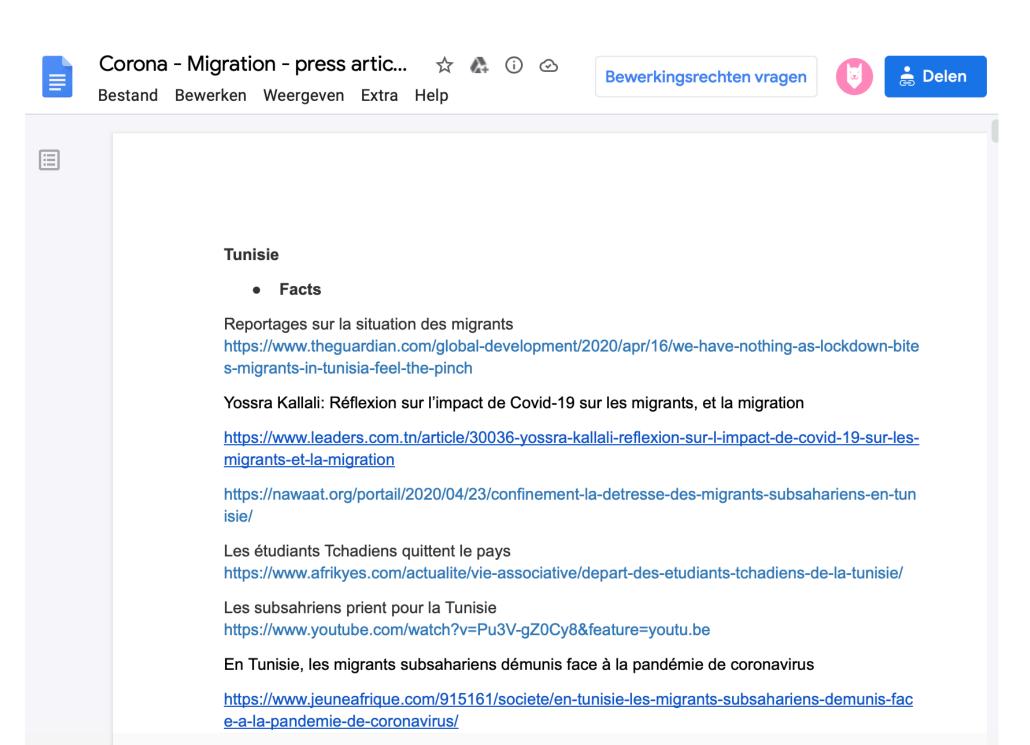
KIU FAIR Data Point / Migrant Media Reports / Media Report Dataset / Media Reports listed in a google do...

Media Reports listed in a google document

These are internet links to media reports listed in a google document.

Access online	Acce to the data
Metadata Issued 13-07-2020	Metadata Modified 13-07-2020
/ersion I.O	
anguage <u>en</u>	
icense	
ssued	

Log in



https://home.fairdatapoint.org

FAIR Data Point index

Filter:	tive 15 Inactive 3 Unreachabl	e 32 Invalid 6 Unknown	0
Endpoint 🔺 🔻	Registration 🔺 🔻	Modification 🔺 🔻 St	tatus
https://fdp.uc.rnu.tn	27/07/2020, 12:29:33	07/10/2020, 08:35:47 Ac	tive
https://fdp.lumc.nl	26/08/2020, 14:58:14	07/10/2020, 07:53:49 Ac	tive
https://staging.fairdatapoint.org	29/04/2020, 15:23:20	06/10/2020, 17:50:17 Ac	tive
https://fdp.aau.edu.et	09/08/2020, 20:06:46	06/10/2020, 11:31:39 Ac	tive
https://fdp.vodan.fairdatapoint.org	12/06/2020, 13:06:57	06/10/2020, 10:12:20	tive
http://ejprd.fair-dtls.surf-hosted.nl:8082	19/08/2020, 11:57:17	05/10/2020, 10:21:51	tive
http://lumc-beat-covid.fair-dtls.surf-hosted.nl	03/06/2020, 16:33:03	05/10/2020, 09:36:56 Ac	tive
https://fdp.tangaza.ac.ke	03/09/2020, 11:16:08	03/10/2020, 16:16:48 Ac	tive
http://ejprd.fair-dtls.surf-hosted.nl:8084	11/09/2020, 14:03:18	02/10/2020, 23:38:40 Ac	tive
https://fdp.ibbu.edu.ng	11/08/2020, 14:24:33	02/10/2020, 00:21:12	tive
https://fdp.sdsc.edu	01/05/2020, 23:44:58	01/10/2020, 22:42:27	tive
https://fdp.test.fairdatapoint.org	16/07/2020, 13:47:48	01/10/2020, 16:43:50 Ac	tive
https://fairsfair.fair-dtls.surf-hosted.nl	29/07/2020, 14:35:30	01/10/2020, 09:46:53	tive
https://fdps.kiu.ac.ug	22/07/2020, 13:52:08	30/09/2020, 16:05:55 Ac	tive
https://app.fairdatapoint.org	29/04/2020, 16:37:21	30/09/2020, 14:37:06 Ac	tive

Thank you & Questions?