INVENIO RDM

Maximilian MOSER
Sotirios TSEPELAKIS

Reminder: What are the FAIR principles?

- **F**indable
  - The object should easily be found in searches
- **A**ccessible
  - The conditions for access to the object must be clear
- **I**nteroperable
  - The object's (meta-)data must be interpretable by others
- **R**eusable
  - The license for reusing the object (once accessed) must be clear

Reminder: What are research data repositories?

Basically just virtual places where researchers can deposit their data to share it with others, in a controlled fashion

For example

https://about.zenodo.org/

Why do we want a repository?

Imagine you're working on a paper for scheduling optimization and want to see how good your approach performs on data already used in the literature.

Get your datasets from e.g.
http://schedulingresearch.com/

What is the problem here?

“He’s dead, Jim.”

- Have you ever wanted to reproduce a paper’s results, but couldn’t find the input data anywhere?
- Have you ever followed a link in a paper just to find that it’s dead?
- Will you remember the path on the server where you put your data in half a year?

Isn’t this nicer?

The normalised Sentinel-1 Global Backscatter Model, mapping Earth’s land surface with C-band microwaves

https://doi.org/10.1038/s41597-021-01059-7

The referenced dataset

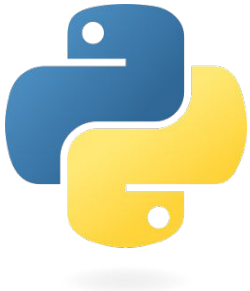
The Sentinel-1 Global Backscatter Model (S1GBM) - Mapping Earth's Land Surface with C-Band Microwaves

https://doi.org/10.48436/r9fn3-nyd51

What was that just now?

A customized instance of InvenioRDM, which is…

- a free and open-source research data repository
- based on Flask and Python
- currently under development by CERN and partners
- documented on https://inveniordm.docs.cern.ch/



https://www.python.org/community/logos/
https://github.com/inveniosoftware/docs-invenio-rdm

But does it help with FAIRness?

…

But does it help with FAIRness?

Yes.

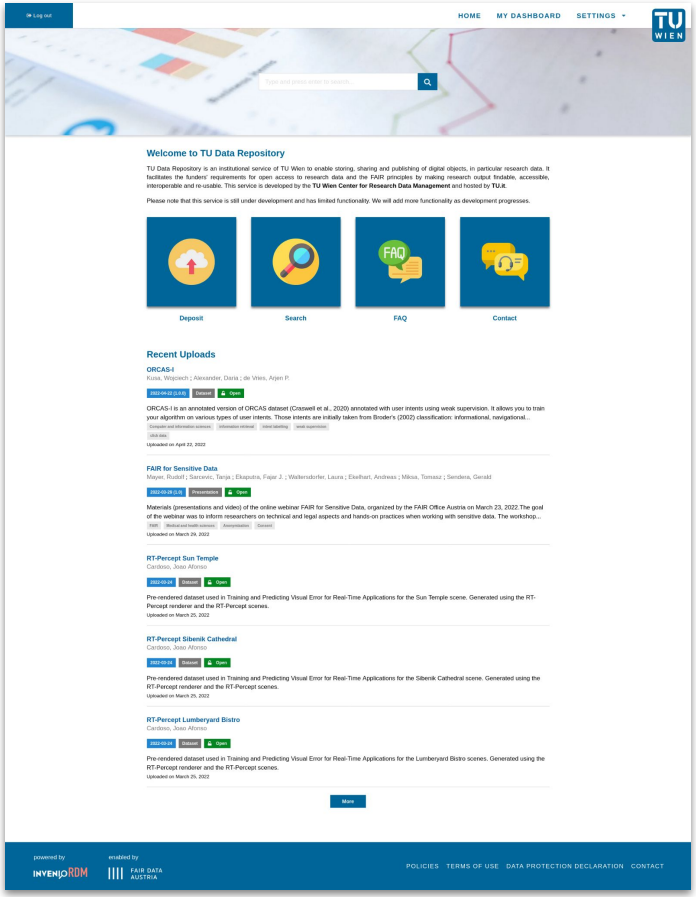
F	F1: (Meta) data are assigned globally unique and persistent identifiers	**A DOI is issued to every published record on InvenioRDM.**
	F2: Data are described with rich metadata	**InvenioRDM's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichments.**
	F3: Metadata clearly and explicitly include the identifier of the data they describe	**The DOI is a top-level and a mandatory field in the metadata of each record.**
	F4: (Meta)data are registered or indexed in a searchable resource	**(1) Metadata of each record is indexed and searchable directly in InvenioRDM's search engine immediately after publishing. (2) Metadata of each record is sent to DataCite servers during DOI registration and indexed there.**
A	A1: (meta)data are retrievable by their identifier using a standardized communications protocol	**Metadata for individual records as well as record collections are harvestable using the OAI-PMH protocol by the record identifier and the collection name. Metadata is also retrievable through the public REST API.**
	A1.1: The protocol is open, free and universally implementable	**See point A1. OAI-PMH and REST are open, free and universal protocols for information retrieval on the web.**
	A1.2: The protocol allows for an authentication and authorisation where necessary	**Metadata are publicly accessible and licensed under public domain. No authorization is ever necessary to retrieve it.**
	A2: Metadata should be accessible even when the data is no longer available	**(1) Data and metadata will be retained for the lifetime of the repository. (2) Metadata are stored in high-availability database servers which are separate to the data itself. (note: recommendations for local implementations should be communicated here)**
I	I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	**InvenioRDM uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARC-XML.**
	I2: (Meta)data use vocabularies that follow the FAIR principles	**For certain terms we refer to open, external vocabularies, e.g.: license (Open Definition), funders (FundRef) and grants (OpenAIRE).**
	I3: (Meta)data include qualified references to other (meta)data	**Each referenced external piece of metadata is qualified by a resolvable URL.**
R	R1: (Meta)data are richly described with a plurality of accurate and relevant attributes	**Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and InvenioRDM's enrichments.**
	R1.1: (Meta)data are released with a clear and accessible data usage license	**(1) License is one of the mandatory terms in InvenioRDM's metadata, and is referring to an Open Definition license. (2) Data downloaded by the users is subject to the license specified in the metadata by the uploader.**
	R1.2: (Meta)data are associated with detailed provenance	**(1) All data and metadata uploaded is traceable to a registered InvenioRDM user. (2) Metadata can optionally describe the original authors of the published work.**
	R1.3: (Meta)data meet domain-relevant community standards	**InvenioRDM is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.**

Gonzales, Sara, & Holmes, Kristi L. (2022, June 6). InvenioRDM: A Collaborative Community. InvenioRDM Workshop 2022 at Open Repositories 2022 (OR2022), Denver, US. Zenodo. https://doi.org/10.5281/zenodo.6643282

Demo Time!

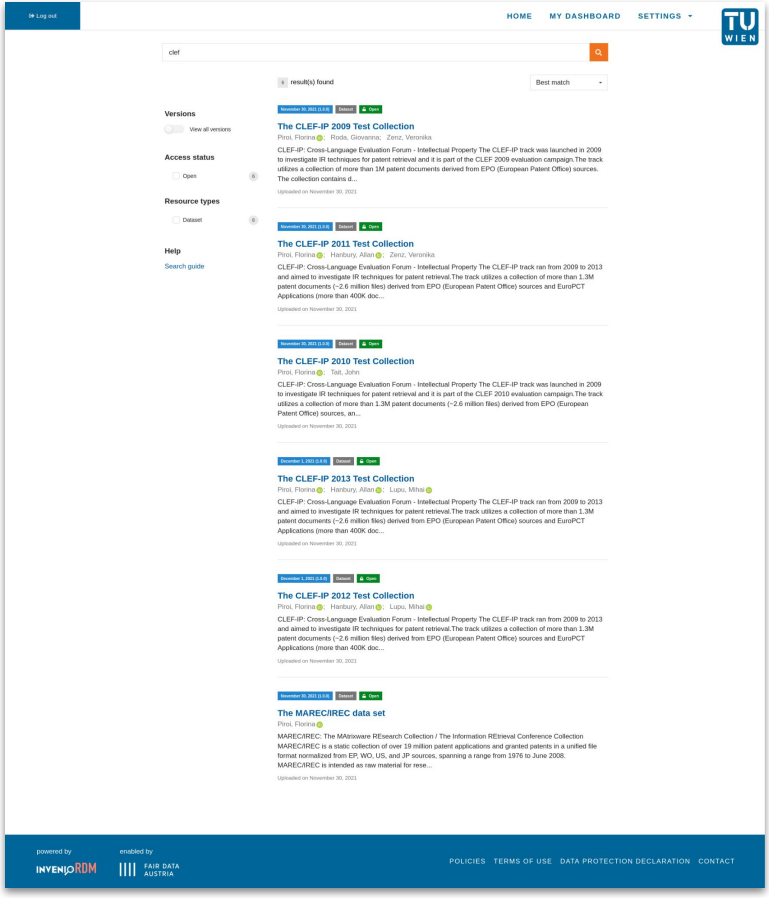
https://test.researchdata.tuwien.ac.at/

Frontpage

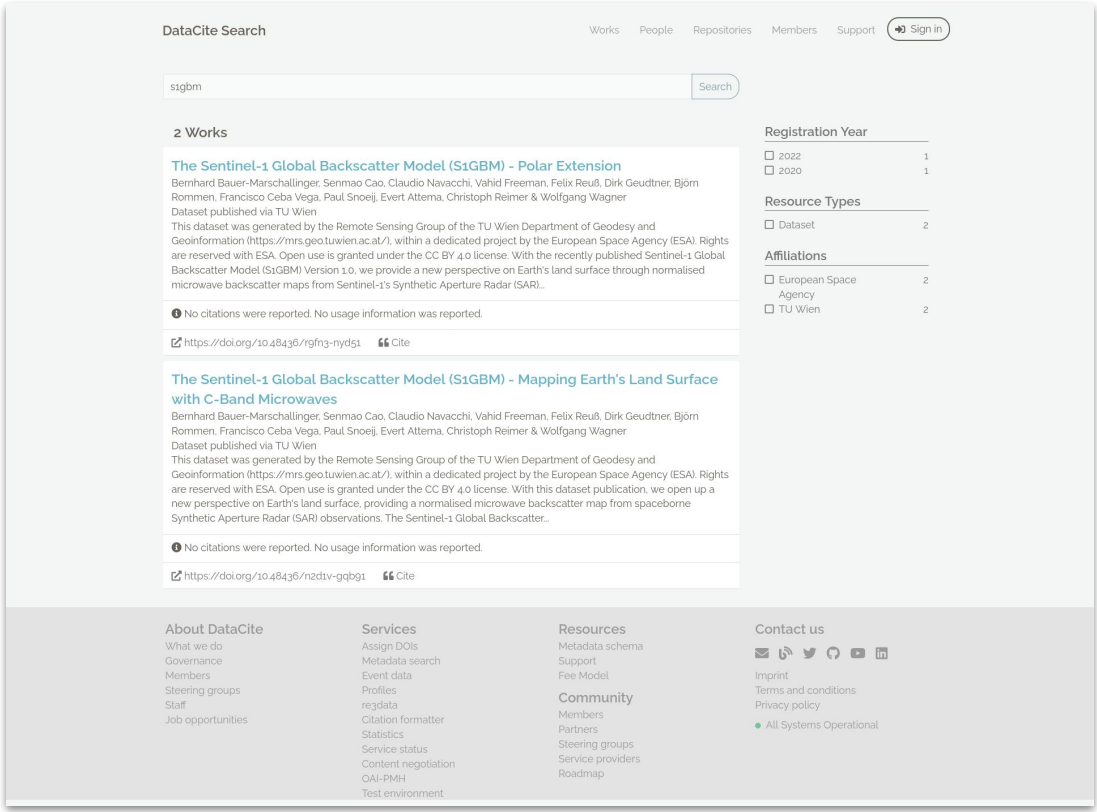


https://researchdata.tuwien.ac.at

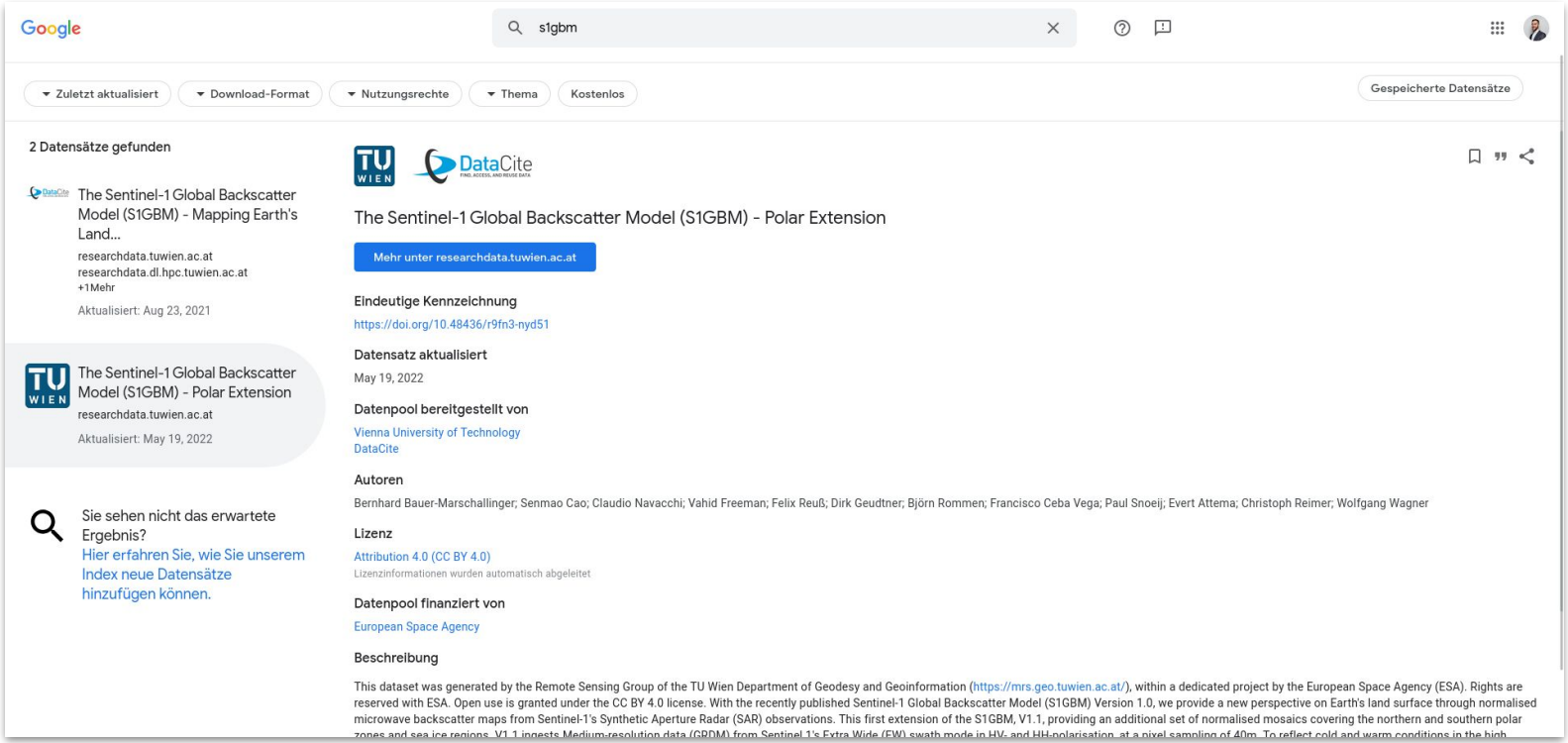
Search


Log out
HOME
MY DASHBOARD
SETTINGS
TU WIEN
clef
6 result(s) found
Best match
Versions
View all versions
Access status
Open 6
Resource types
Dataset 6
Help
Search guide
The CLEF-IP 2009 Test Collection
Piroi, Florina; Roda, Giovanna; Zenz, Veronika
CLEF-IP: Cross-Language Evaluation Forum - Intellectual Property The CLEF-IP track was launched in 2009 to investigate IR techniques for patent retrieval and it is part of the CLEF 2009 evaluation campaign.The track utilizes a collection of more than 1M patent documents derived from EPO (European Patent Office) sources. The collection contains d...
Uploaded on November 30, 2021
The CLEF-IP 2011 Test Collection
Piroi, Florina; Hanbury, Allan; Zenz, Veronika
CLEF-IP: Cross-Language Evaluation Forum - Intellectual Property The CLEF-IP track ran from 2009 to 2013 and aimed to investigate IR techniques for patent retrieval.The track utilizes a collection of more than 1.3M patent documents (~2.6 million files) derived from EPO (European Patent Office) sources and EuroPCT Applications (more than 400K doc...
Uploaded on November 30, 2021
The CLEF-IP 2010 Test Collection
Piroi, Florina; Tait, John
CLEF-IP: Cross-Language Evaluation Forum - Intellectual Property The CLEF-IP track was launched in 2009 to investigate IR techniques for patent retrieval and it is part of the CLEF 2010 evaluation campaign.The track utilizes a collection of more than 1.3M patent documents (~2.6 million files) derived from EPO (European Patent Office) sources, an...
Uploaded on November 30, 2021
The CLEF-IP 2013 Test Collection
Piroi, Florina; Hanbury, Allan; Lupu, Mihai
CLEF-IP: Cross-Language Evaluation Forum - Intellectual Property The CLEF-IP track ran from 2009 to 2013 and aimed to investigate IR techniques for patent retrieval.The track utilizes a collection of more than 1.3M patent documents (~2.6 million files) derived from EPO (European Patent Office) sources and EuroPCT Applications (more than 400K doc...
Uploaded on November 30, 2021
The CLEF-IP 2012 Test Collection
Piroi, Florina; Hanbury, Allan; Lupu, Mihai
CLEF-IP: Cross-Language Evaluation Forum - Intellectual Property The CLEF-IP track ran from 2009 to 2013 and aimed to investigate IR techniques for patent retrieval.The track utilizes a collection of more than 1.3M patent documents (~2.6 million files) derived from EPO (European Patent Office) sources and EuroPCT Applications (more than 400K doc...
Uploaded on November 30, 2021
The MAREC/IREC data set
Piroi, Florina
MAREC/IREC: The MAtrixware REsearch Collection / The Information REtrieval Conference Collection MAREC/IREC is a static collection of over 19 million patent applications and granted patents in a unified file format normalized from EP, WO, US, and JP sources, spanning a range from 1976 to June 2008. MAREC/IREC is intended as raw material for rese...
Uploaded on November 30, 2021
powered by
INVENIORDM
enabled by
FAIR DATA AUSTRIA
POLICIES
TERMS OF USE
DATA PROTECTION DECLARATION
CONTACT

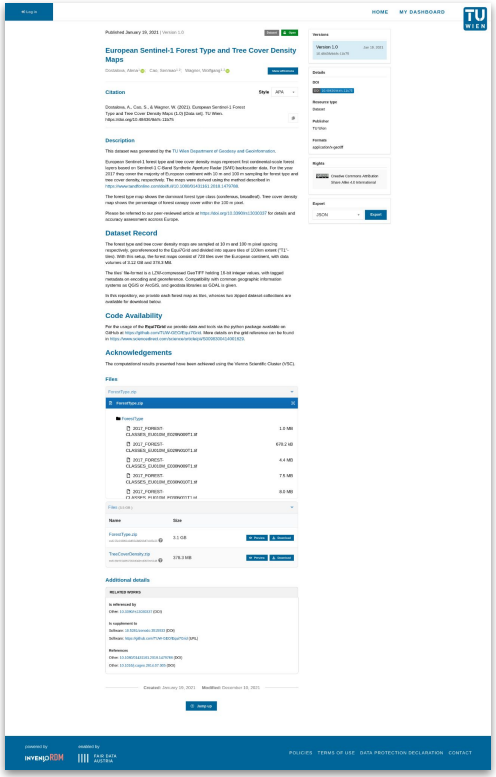
Search (External: DataCite)



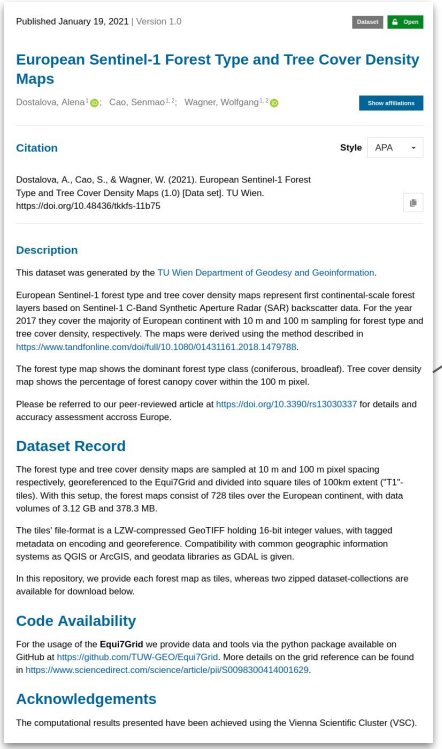
Search (External: Google)



Record Landing Page



Record Landing Page



Published January 19, 2021 | Version 1.0

Dataset | Open

European Sentinel-1 Forest Type and Tree Cover Density Maps

Dostalova, Alena[1]; Cao, Senmao[1, 2]; Wagner, Wolfgang[1, 2]

Show affiliations

Citation

Style APA

Dostalova, A., Cao, S., & Wagner, W. (2021). European Sentinel-1 Forest Type and Tree Cover Density Maps (1.0) [Data set]. TU Wien. https://doi.org/10.48436/tkkfs-11b75

Description

This dataset was generated by the TU Wien Department of Geodesy and Geoinformation.

European Sentinel-1 forest type and tree cover density maps represent first continental-scale forest layers based on Sentinel-1 C-Band Synthetic Aperture Radar (SAR) backscatter data. For the year 2017 they cover the majority of European continent with 10 m and 100 m sampling for forest type and tree cover density, respectively. The maps were derived using the method described in https://www.tandfonline.com/doi/full/10.1080/01431161.2018.1479788.

The forest type map shows the dominant forest type class (coniferous, broadleaf). Tree cover density map shows the percentage of forest canopy cover within the 100 m pixel.

Please be referred to our peer-reviewed article at https://doi.org/10.3390/rs13030337 for details and accuracy assessment accross Europe.

Dataset Record

The forest type and tree cover density maps are sampled at 10 m and 100 m pixel spacing respectively, georeferenced to the Equi7Grid and divided into square tiles of 100km extent ("T1"-tiles). With this setup, the forest maps consist of 728 tiles over the European continent, with data volumes of 3.12 GB and 378.3 MB.

The tiles' file-format is a LZW-compressed GeoTIFF holding 16-bit integer values, with tagged metadata on encoding and georeference. Compatibility with common geographic information systems as QGIS or ArcGIS, and geodata libraries as GDAL is given.

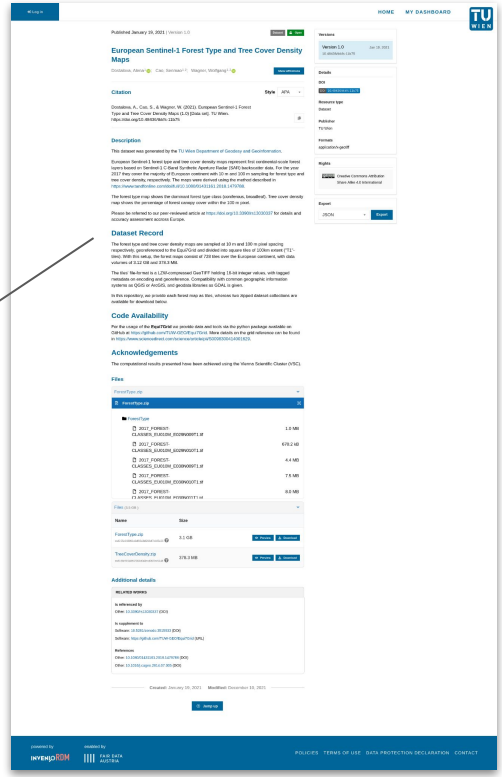
In this repository, we provide each forest map as tiles, whereas two zipped dataset-collections are available for download below.

Code Availability

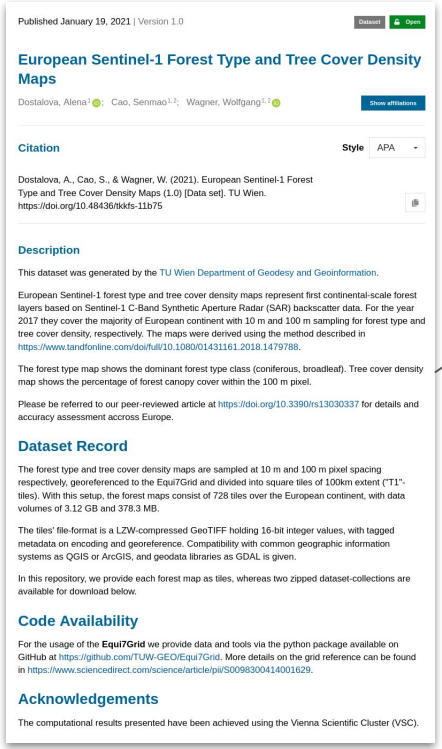
For the usage of the **Equi7Grid** we provide data and tools via the python package available on GitHub at https://github.com/TUW-GEO/Equi7Grid. More details on the grid reference can be found in https://www.sciencedirect.com/science/article/pii/S0098300414001629.

Acknowledgements

The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).



Record Landing Page



Published January 19, 2021 | Version 1.0

Dataset | Open

European Sentinel-1 Forest Type and Tree Cover Density Maps

Dostalova, Alena[1]; Cao, Senmao[1,2]; Wagner, Wolfgang[1,2]

Show affiliations

Citation

Style APA

Dostalova, A., Cao, S., & Wagner, W. (2021). European Sentinel-1 Forest Type and Tree Cover Density Maps (1.0) [Data set]. TU Wien. https://doi.org/10.48436/tkkfs-11b75

Description

This dataset was generated by the TU Wien Department of Geodesy and Geoinformation.

European Sentinel-1 forest type and tree cover density maps represent first continental-scale forest layers based on Sentinel-1 C-Band Synthetic Aperture Radar (SAR) backscatter data. For the year 2017 they cover the majority of European continent with 10 m and 100 m sampling for forest type and tree cover density, respectively. The maps were derived using the method described in https://www.tandfonline.com/doi/full/10.1080/01431161.2018.1479788.

The forest type map shows the dominant forest type class (coniferous, broadleaf). Tree cover density map shows the percentage of forest canopy cover within the 100 m pixel.

Please be referred to our peer-reviewed article at https://doi.org/10.3390/rs13030337 for details and accuracy assessment accross Europe.

Dataset Record

The forest type and tree cover density maps are sampled at 10 m and 100 m pixel spacing respectively, georeferenced to the Equi7Grid and divided into square tiles of 100km extent ("T1"-tiles). With this setup, the forest maps consist of 728 tiles over the European continent, with data volumes of 3.12 GB and 378.3 MB.

The tiles' file-format is a LZW-compressed GeoTIFF holding 16-bit integer values, with tagged metadata on encoding and georeference. Compatibility with common geographic information systems as QGIS or ArcGIS, and geodata libraries as GDAL is given.

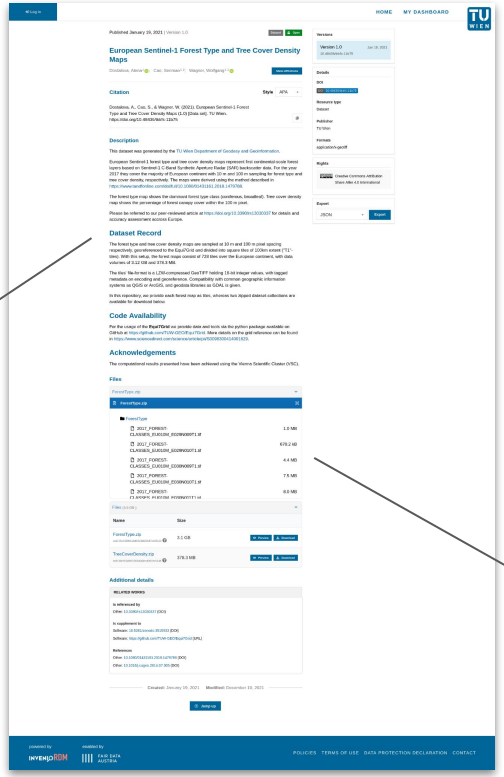
In this repository, we provide each forest map as tiles, whereas two zipped dataset-collections are available for download below.

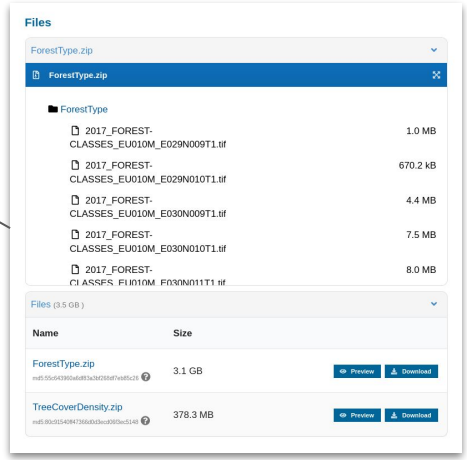
Code Availability

For the usage of the **Equi7Grid** we provide data and tools via the python package available on GitHub at https://github.com/TUW-GEO/Equi7Grid. More details on the grid reference can be found in https://www.sciencedirect.com/science/article/pii/S0098300414001629.

Acknowledgements

The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).





Files

ForestType.zip

ForestType.zip

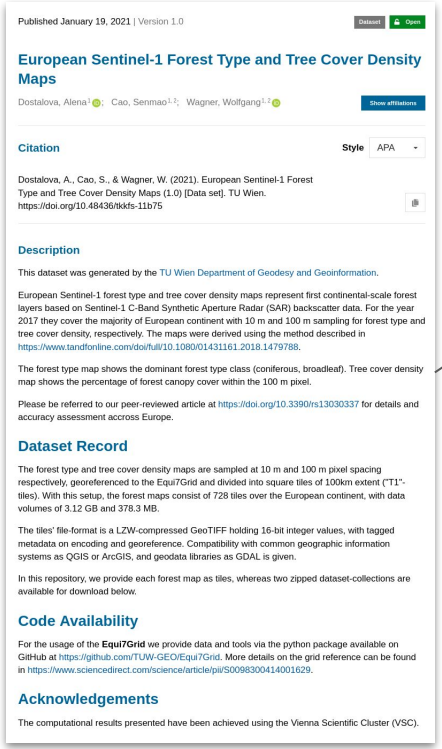
ForestType

2017_FOREST-CLASSES_EU010M_E029N009T1.tif	1.0 MB
2017_FOREST-CLASSES_EU010M_E029N010T1.tif	670.2 kB
2017_FOREST-CLASSES_EU010M_E030N009T1.tif	4.4 MB
2017_FOREST-CLASSES_EU010M_E030N010T1.tif	7.5 MB
2017_FOREST-CLASSES_EU010M_E030N011T1.tif	8.0 MB

Files (3.5 GB)

Name	Size	
ForestType.zip md5:55c643960a6d8f3a3bf26fdf7eb85c26	3.1 GB	Preview Download
TreeCoverDensity.zip md5:80c0154084736fab0d3ecd065ec5148	378.3 MB	Preview Download

Record Landing Page



Published January 19, 2021 | Version 1.0

Dataset · Open

European Sentinel-1 Forest Type and Tree Cover Density Maps

Dostalova, Alena[1]; Cao, Senmao[1,2]; Wagner, Wolfgang[1,2]

Show affiliations

Citation

Style APA

Dostalova, A., Cao, S., & Wagner, W. (2021). European Sentinel-1 Forest Type and Tree Cover Density Maps (1.0) [Data set]. TU Wien. https://doi.org/10.48436/tkkfs-11b75

Description

This dataset was generated by the TU Wien Department of Geodesy and Geoinformation.

European Sentinel-1 forest type and tree cover density maps represent first continental-scale forest layers based on Sentinel-1 C-Band Synthetic Aperture Radar (SAR) backscatter data. For the year 2017 they cover the majority of European continent with 10 m and 100 m sampling for forest type and tree cover density, respectively. The maps were derived using the method described in https://www.tandfonline.com/doi/full/10.1080/01431161.2018.1479788.

The forest type map shows the dominant forest type class (coniferous, broadleaf). Tree cover density map shows the percentage of forest canopy cover within the 100 m pixel.

Please be referred to our peer-reviewed article at https://doi.org/10.3390/rs13030337 for details and accuracy assessment accross Europe.

Dataset Record

The forest type and tree cover density maps are sampled at 10 m and 100 m pixel spacing respectively, georeferenced to the Equi7Grid and divided into square tiles of 100km extent ("T1"-tiles). With this setup, the forest maps consist of 728 tiles over the European continent, with data volumes of 3.12 GB and 378.3 MB.

The tiles' file-format is a LZW-compressed GeoTIFF holding 16-bit integer values, with tagged metadata on encoding and georeference. Compatibility with common geographic information systems as QGIS or ArcGIS, and geodata libraries as GDAL is given.

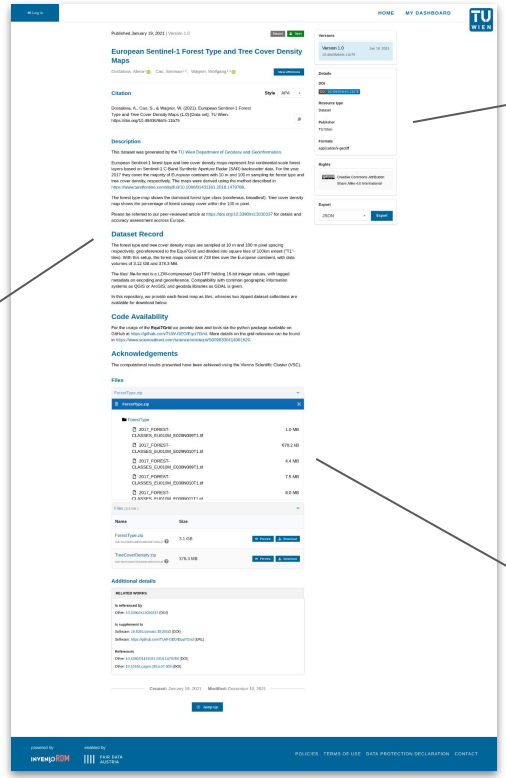
In this repository, we provide each forest map as tiles, whereas two zipped dataset-collections are available for download below.

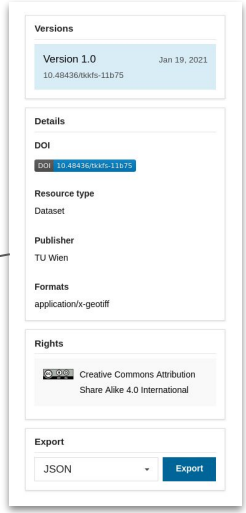
Code Availability

For the usage of the **Equi7Grid** we provide data and tools via the python package available on GitHub at https://github.com/TUW-GEO/Equi7Grid. More details on the grid reference can be found in https://www.sciencedirect.com/science/article/pii/S0098300414001629.

Acknowledgements

The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).





Versions

Version 1.0 — Jan 19, 2021
10.48436/tkkfs-11b75

Details

DOI
10.48436/tkkfs-11b75

Resource type
Dataset

Publisher
TU Wien

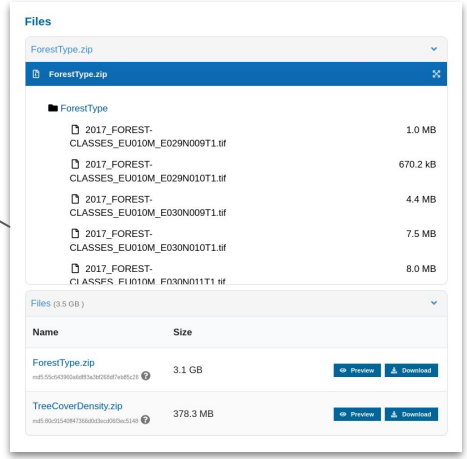
Formats
application/x-geotiff

Rights

Creative Commons Attribution Share Alike 4.0 International

Export

JSON — Export



Files

ForestType.zip

ForestType.zip

ForestType

2017_FOREST-CLASSES_EU010M_E029N009T1.tif	1.0 MB
2017_FOREST-CLASSES_EU010M_E029N010T1.tif	670.2 kB
2017_FOREST-CLASSES_EU010M_E030N009T1.tif	4.4 MB
2017_FOREST-CLASSES_EU010M_E030N010T1.tif	7.5 MB
2017_FOREST-CLASSES_EU010M_E030N011T1.tif	8.0 MB

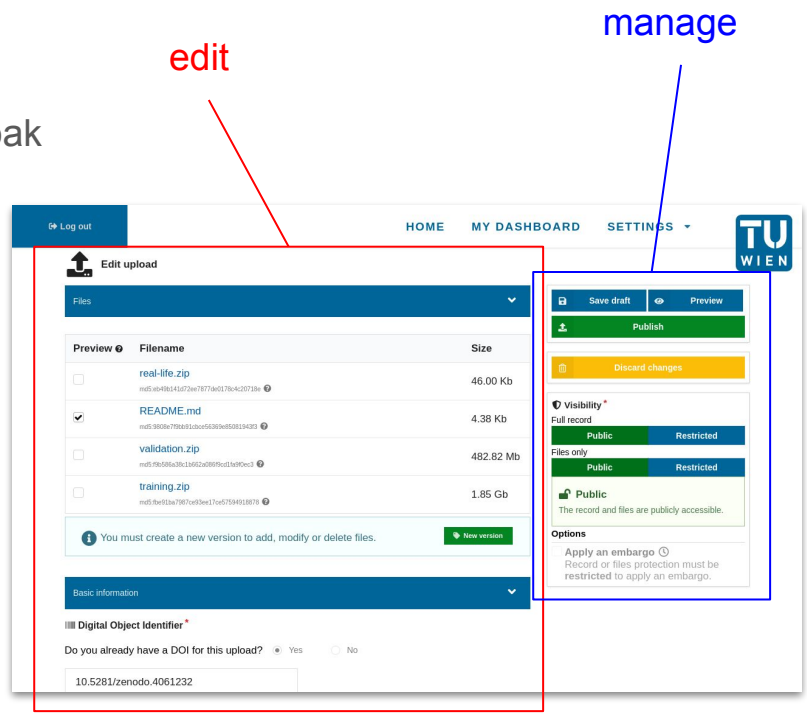
Files (3.5 GB)

Name	Size	
ForestType.zip md5:55c643960a6d8f3a2b926fdf7eb85c26	3.1 GB	Preview · Download
TreeCoverDensity.zip md5:80c9154084736fab0d3ecd065ec5148	378.3 MB	Preview · Download

Access Restrictions

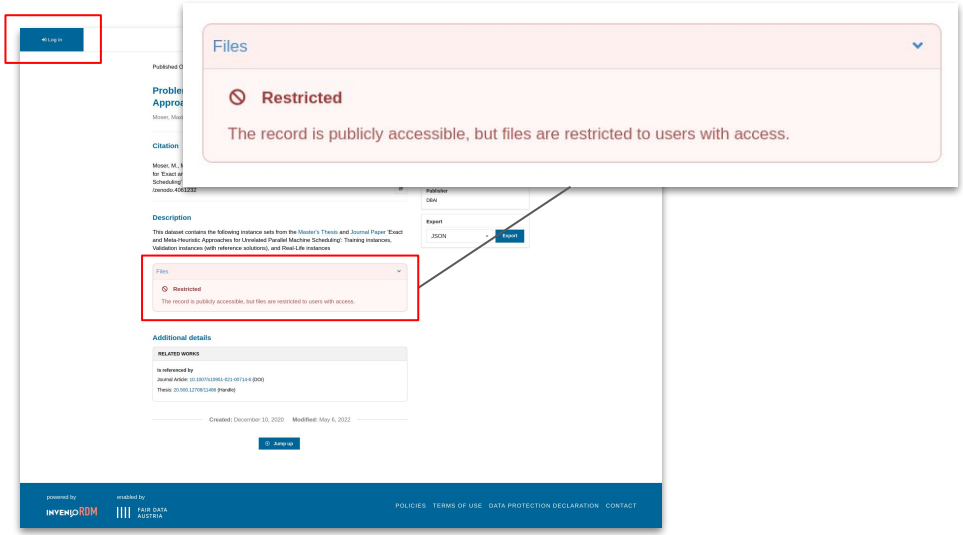
- Authentication
  - Local login vs OAuth2 (resp. OpenID Connect)
    - Get other services integrated, e.g. Keycloak
      => integrates TU SSO
  - Sessions for browsers, tokens for cURL

- Authorization
  - Different permissions for different actions
    - read, read_files, create, edit, manage, ...
  - Role-based access (e.g. role “trusted-user”)
  - Share-by-Link (like Google Docs)

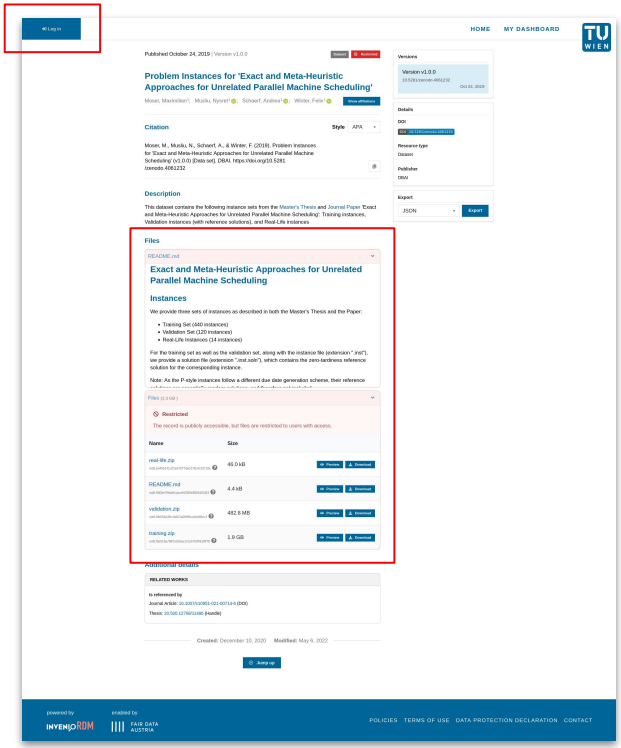


Share by Link

Can be used to share access to restricted datasets:



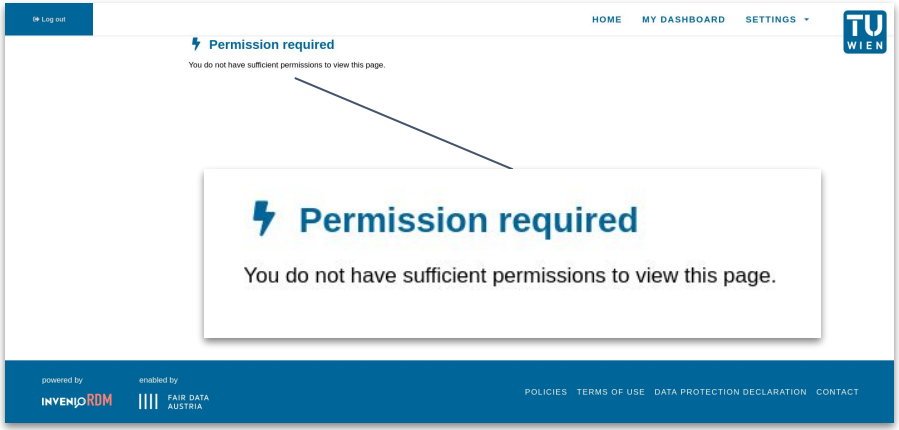
Normal Access



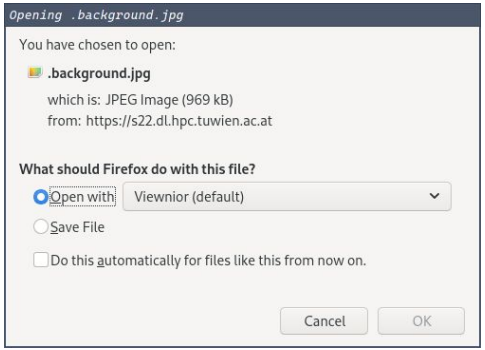
Access via Secret Link

Share by Link

Can also be used to enable double-blind peer reviews:



Metadata cannot be accessed...



… but files can be downloaded

TUW Research Data vs InvenioRDM

Customized with our own modules: https://gitlab.tuwien.ac.at/fairdata

- Invenio-Theme-TUW
  - TU Wien corporate design, and extra pages
- Invenio-Config-TUW
  - TU Wien SSO for authentication
  - customized permissions
  - tweaked inner workings
- Invenio-Utilities-TUW
  - additional CLI commands for administration

Summary: Core InvenioRDM features

- Record Management
  - Metadata based on DataCite kernel 4.3
  - File deposit
  - DOI minting on publication
  - Record Versioning
  - Export of metadata in other formats (e.g. Dublin Core, OAI-PMH, …)
  - Access sharing (e.g. via secret link, similar to Google Docs)
- Communities
- Modularity and customizability
- Growing developer community (on Discord)

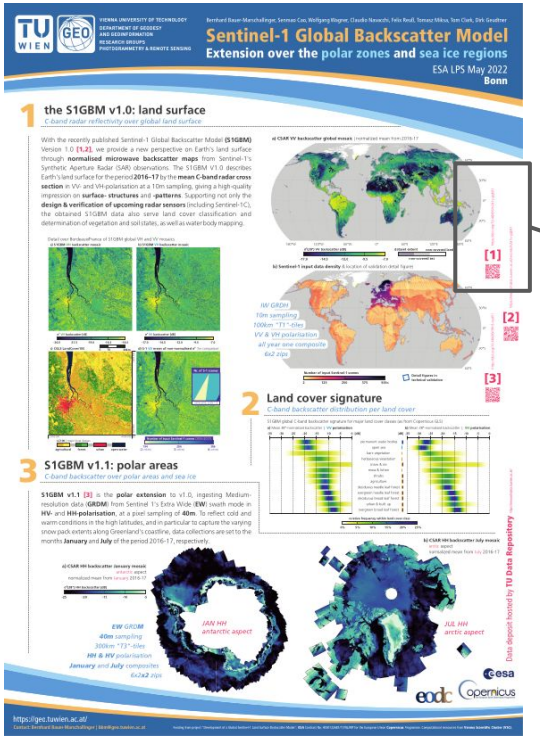
Bonus tips for depositing data

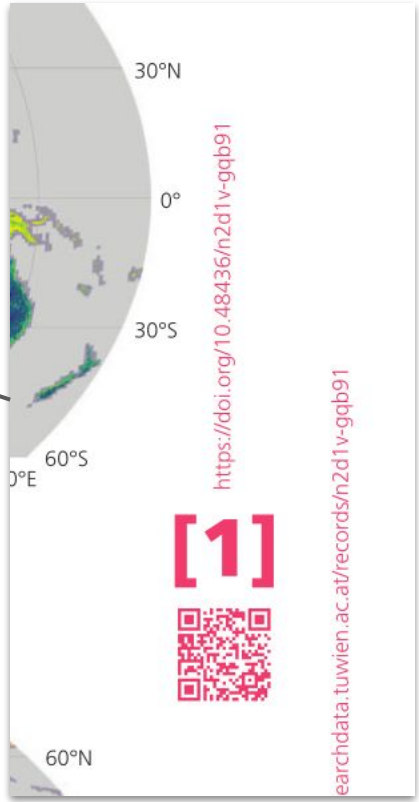
- **Rights**: Make sure you have the rights to upload the data
  - check back with all involved creators
  - agree on that before even creating the dataset

- **Longevity**: Choose proper file formats for your uploads
  - simple and open standards over proprietary niche formats
  - c.f. Library of Congress recommendations

- **FAIRness**: Test your record's FAIRness with tools
  - e.g. F-UJI

And an idea for posters!





Future work

- Integration with our other services, e.g.
  - the DBRepo
  - our CRIS system
  - our maDMP tool: https://damap.org/
- Account creation workflow for externals
- Virus scanning on upload
- Preservation/archival of suitable uploaded files
- Factor in the uploaded files' contents for search
- Etc. – see also https://inveniosoftware.org/products/rdm/roadmap/

Thank you for your attention!

Questions?