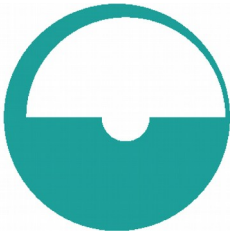
Trying to track the re-use of research data – open licensing, DOI usage and the Data Citation Index

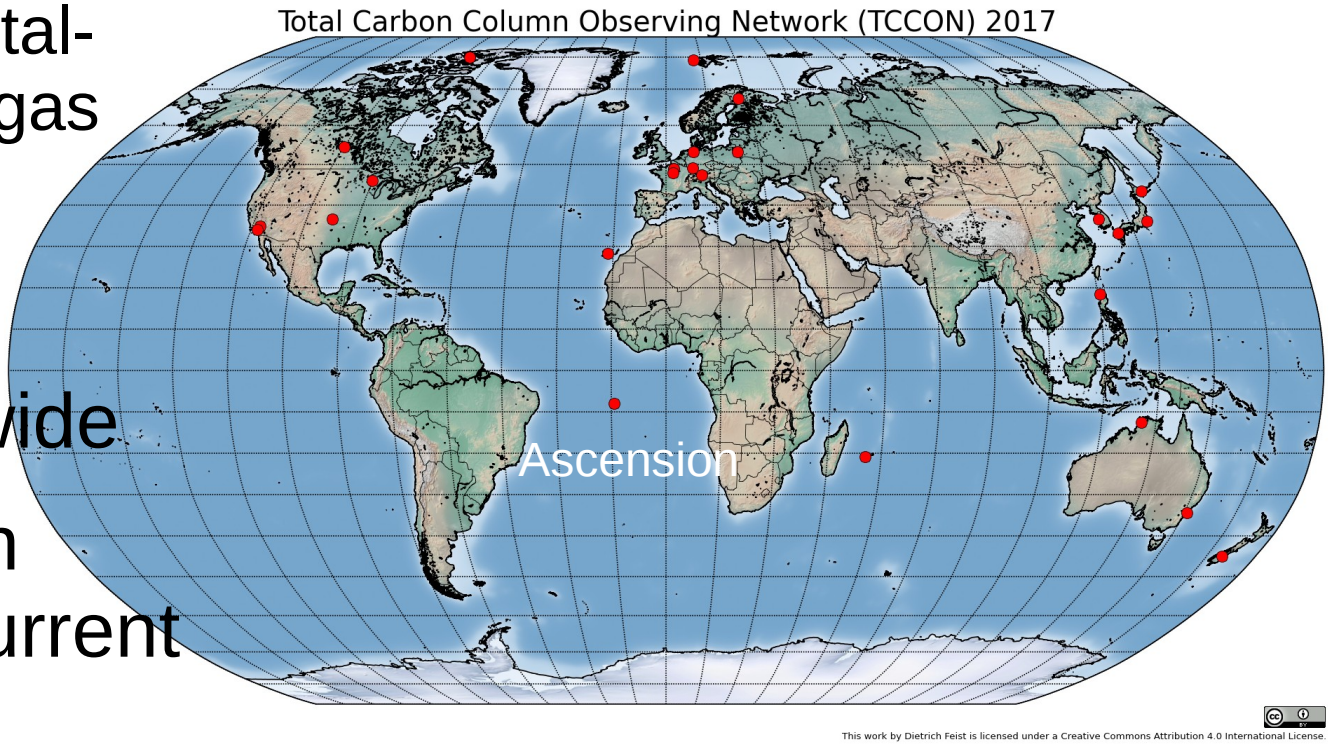
Dietrich Feist <dfeist@bgc-jena.mpg.de>
Max Planck Institute for Biogeochemistry
Jena, Germany

Max-Planck-Institut
für Biogeochemie



Total Carbon Column Observing Network (TCCON)

- Global network for total-column greenhouse gas (GHG) observations
- Established 2005
- 25 stations worldwide
- Primary calibration reference for all current and future GHG satellite observations

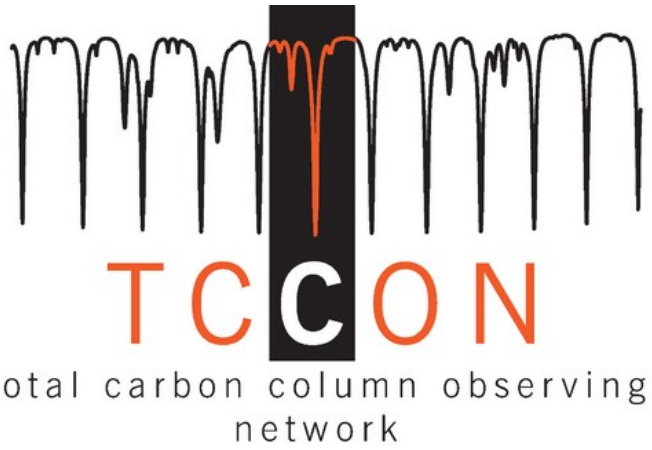


Ascension Island TCCON station

- Established 2012
- Only equatorial station and only station sampling air from Africa and South America
- Operating in extremely harsh environment
- Data used in ~14 publications in 2016 and 2017.

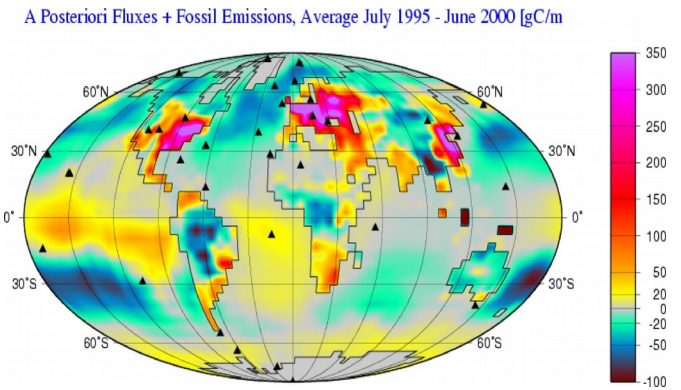
TCCON data policy and licence

- TCCON uses self-defined open data license (no acceptable standard licence).
- Data may be freely downloaded and examined without registration.
- Use in scientific pulications requires contacting PIs before submission. Proper acknowledgement required, co-authorship might be requested.
- Data hosted by CaltechDATA since Oct 2017.



TCCON data users

- Providers of GHG satellite observations: NASA, JAXA, ESA, ...
- Users of GHG satellite observations
- Carbon cycle inverse modelling community
- Other similar networks



Rödenbeck et al., Atmos. Chem. Phys., 9, 5331–5342, 2009

Introduction to DOIs

- A DOI provides a reference to an online document, dataset or other resource through a landing page.
- URLs may change but DOIs should live much longer.
- A DOI like 10.14291/tccon.archive/1348407 contains a prefix (10.14291) and a suffix (rest).
- DOIs should have metadata about the resource attached.
- DOIs do not prescribe any data format for the resource itself.

TCCON DOI usage

- TCCON switched from URLs to DOIs in 2014 after the data archive was moved to a new hosting institution.
- Main goals:
  - Provide an exact reference to the dataset that was used in a publication (reproducibility).
  - Allow tracking of TCCON data use in scientific publications (traceability).

Advantages of DOIs vs URLs

- URLs change often for a variety of reason. Therefore, URLs cited in the scientific literature may become invalid very quickly.
- TCCON has already had to move its archive twice in 10 years, rendering all dataset URLs invalid. Due to the DOIs, the forced move in 2017 was accomplished without major disruptions for the data users or the existing citations.

DOI granularity

- Granularity describes how specific a DOI is for the linked data resource.
- There might be just one DOI for the whole archive or individual DOIs for subsets of data.
- TCCON dataset DOIs uniquely identify the processing software revision, the station and the data release but not the subset of data that was used, e.g.: 10.14291/tccon.ggg2014.ascension01.R0/1149285

DOI metadata

- DOIs have metadata about the resource attached.
- The type of metadata and the format is not fixed. Several accepted standards exist: e.g. Dublin Core, DataCite Metadata Schemata, ...
- DOI metadata is provided in XML format.
- XML metadata file format definition is also provided in XML.

XML metadata for DOIs: Datacite Metadata Schema 4.0

- Describes a set of required and optional metadata elements for a resource:
  - Creators & contributors
  - Description
  - Data Versions
  - Data formats information
  - Time information
  - Geographic information
  - Funding information
  - ...

DataCite

DataCite - International Data Citation

Note that this version of the schema is not backward compatible with previous schema versions. DataCite will provide ongoing support for the use of previous schema versions for a minimum of one year after the release of this version.

DataCite Metadata Schema Documentation for the Publication and Citation of Research Data

Citation:

DataCite Metadata Working Group. (2016). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.0. DataCite e.V. http://doi.org/10.5438/0012.

Members of the Metadata Working Group

Madeleine de Smaele, TU Delft (co-chair of working group)

Joan Starr, California Digital Library (co-chair of working group)

Jan Ashton, British Library

Amy Barton, Purdue University Library

Tina Bradford, NRC/CISTI (New)

Anne Ciolek-Figiel, Inist-CNRS

Stefanie Dietiker, ETH Zurich (New)

Jannean Elliott, DOE/OSTI

Berrit Genat, TIB

Karoline Harzenetter, GESIS

Barbara Hirschmann, ETH Zurich (Departing)

Stefan Jakobsson, SND (New)

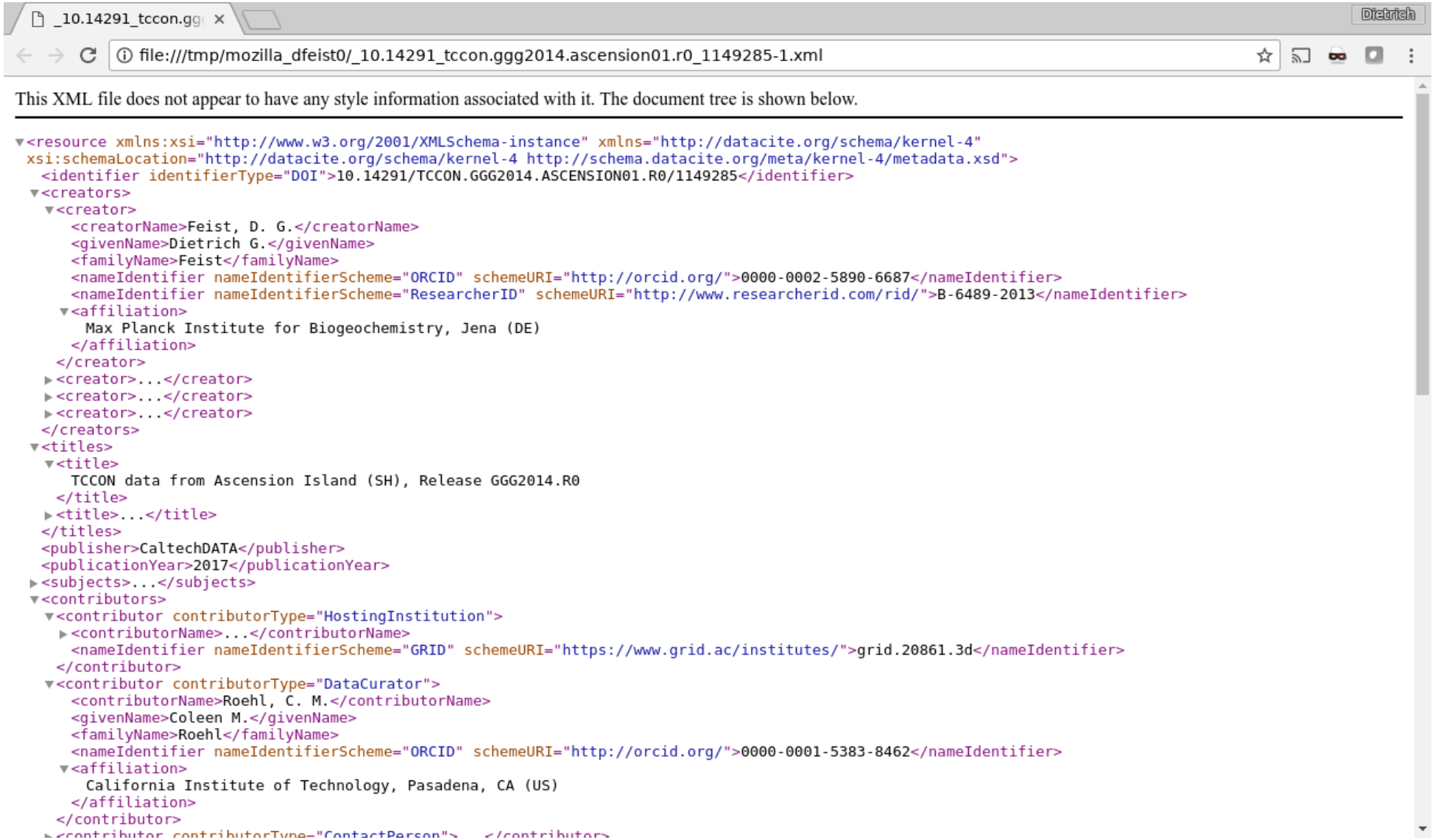
Jean-Yves Mailloux, NRC/CISTI (Departing)

Lars Holm Nielsen, CERN (Departing)

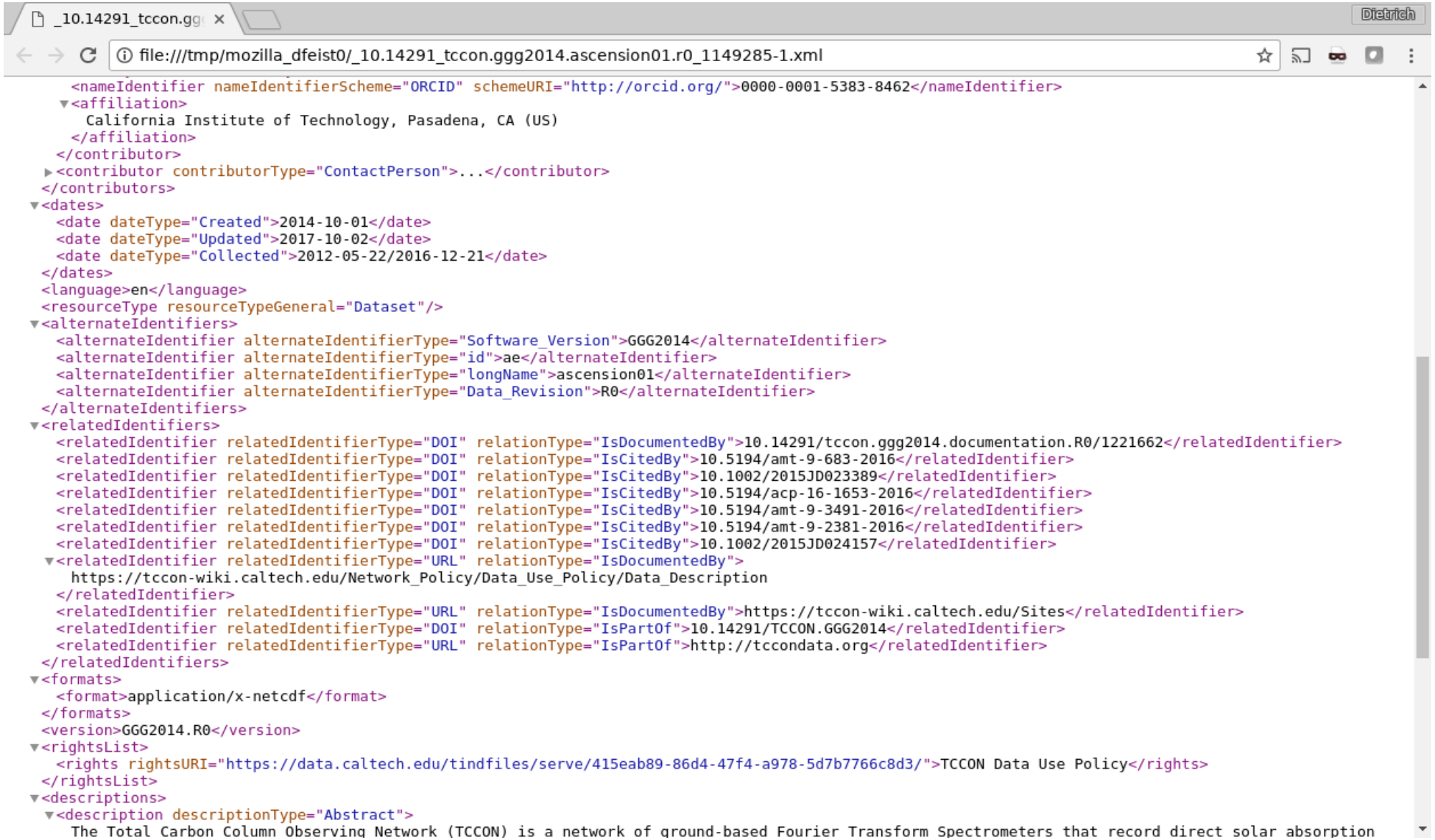
Mohamed Yahia, Inist-CNRS

Frauke Ziedorn, TIB (On leave, Metadata Supervisor)

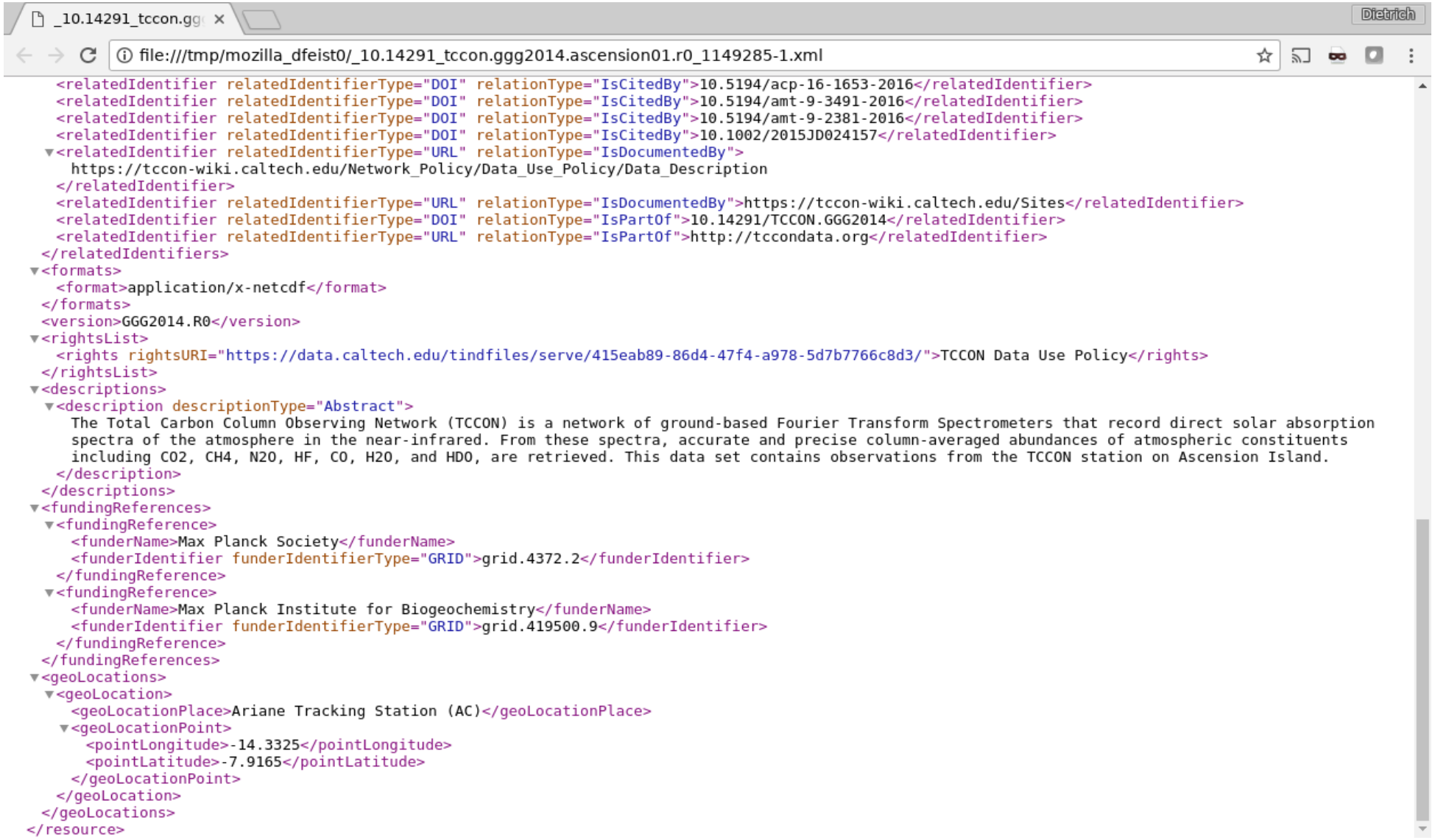
XML metadata file example



XML metadata file example

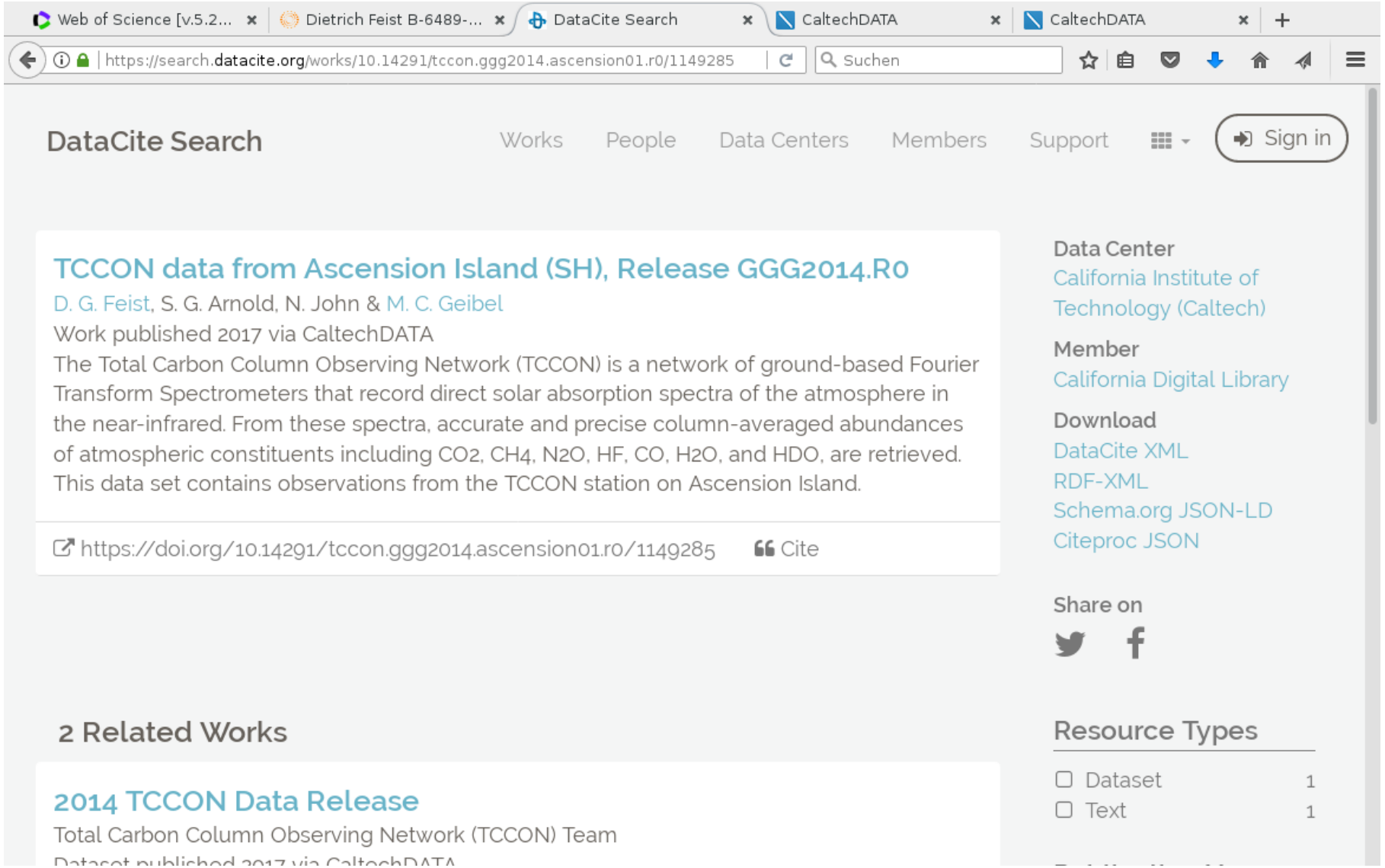


XML metadata file example

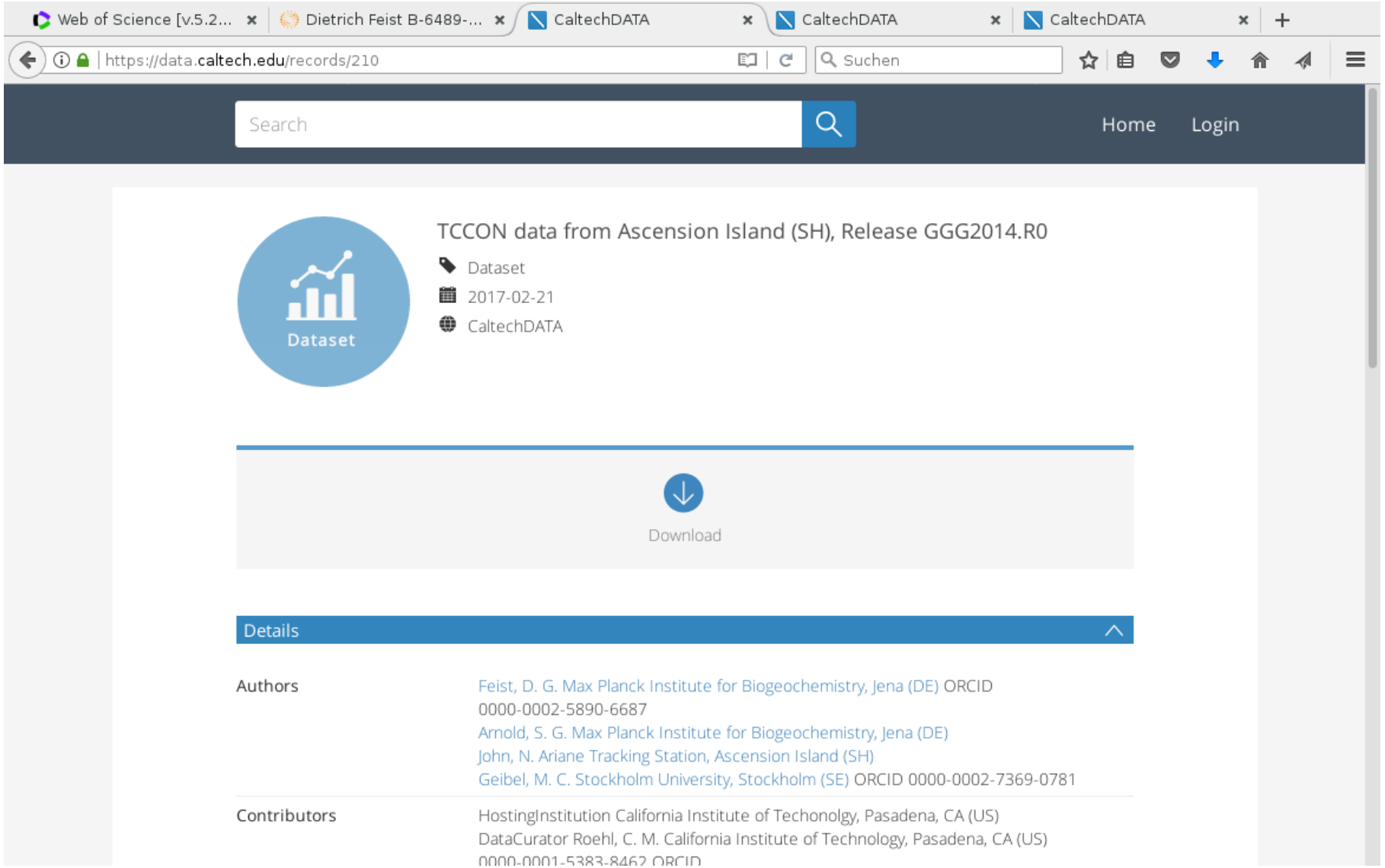



```
    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCitedBy">10.5194/acp-16-1653-2016</relatedIdentifier>
    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCitedBy">10.5194/amt-9-3491-2016</relatedIdentifier>
    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCitedBy">10.5194/amt-9-2381-2016</relatedIdentifier>
    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCitedBy">10.1002/2015JD024157</relatedIdentifier>
  ▼<relatedIdentifier relatedIdentifierType="URL" relationType="IsDocumentedBy">
     https://tccon-wiki.caltech.edu/Network_Policy/Data_Use_Policy/Data_Description
    </relatedIdentifier>
    <relatedIdentifier relatedIdentifierType="URL" relationType="IsDocumentedBy">https://tccon-wiki.caltech.edu/Sites</relatedIdentifier>
    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsPartOf">10.14291/TCCON.GGG2014</relatedIdentifier>
    <relatedIdentifier relatedIdentifierType="URL" relationType="IsPartOf">http://tccondata.org</relatedIdentifier>
  </relatedIdentifiers>
▼<formats>
    <format>application/x-netcdf</format>
  </formats>
  <version>GGG2014.R0</version>
▼<rightsList>
    <rights rightsURI="https://data.caltech.edu/tindfiles/serve/415eab89-86d4-47f4-a978-5d7b7766c8d3/">TCCON Data Use Policy</rights>
  </rightsList>
▼<descriptions>
  ▼<description descriptionType="Abstract">
     The Total Carbon Column Observing Network (TCCON) is a network of ground-based Fourier Transform Spectrometers that record direct solar absorption
     spectra of the atmosphere in the near-infrared. From these spectra, accurate and precise column-averaged abundances of atmospheric constituents
     including CO2, CH4, N2O, HF, CO, H2O, and HDO, are retrieved. This data set contains observations from the TCCON station on Ascension Island.
    </description>
  </descriptions>
▼<fundingReferences>
  ▼<fundingReference>
     <funderName>Max Planck Society</funderName>
     <funderIdentifier funderIdentifierType="GRID">grid.4372.2</funderIdentifier>
    </fundingReference>
  ▼<fundingReference>
     <funderName>Max Planck Institute for Biogeochemistry</funderName>
     <funderIdentifier funderIdentifierType="GRID">grid.419500.9</funderIdentifier>
    </fundingReference>
  </fundingReferences>
▼<geoLocations>
  ▼<geoLocation>
     <geoLocationPlace>Ariane Tracking Station (AC)</geoLocationPlace>
   ▼<geoLocationPoint>
      <pointLongitude>-14.3325</pointLongitude>
      <pointLatitude>-7.9165</pointLatitude>
    </geoLocationPoint>
   </geoLocation>
  </geoLocations>
</resource>
```

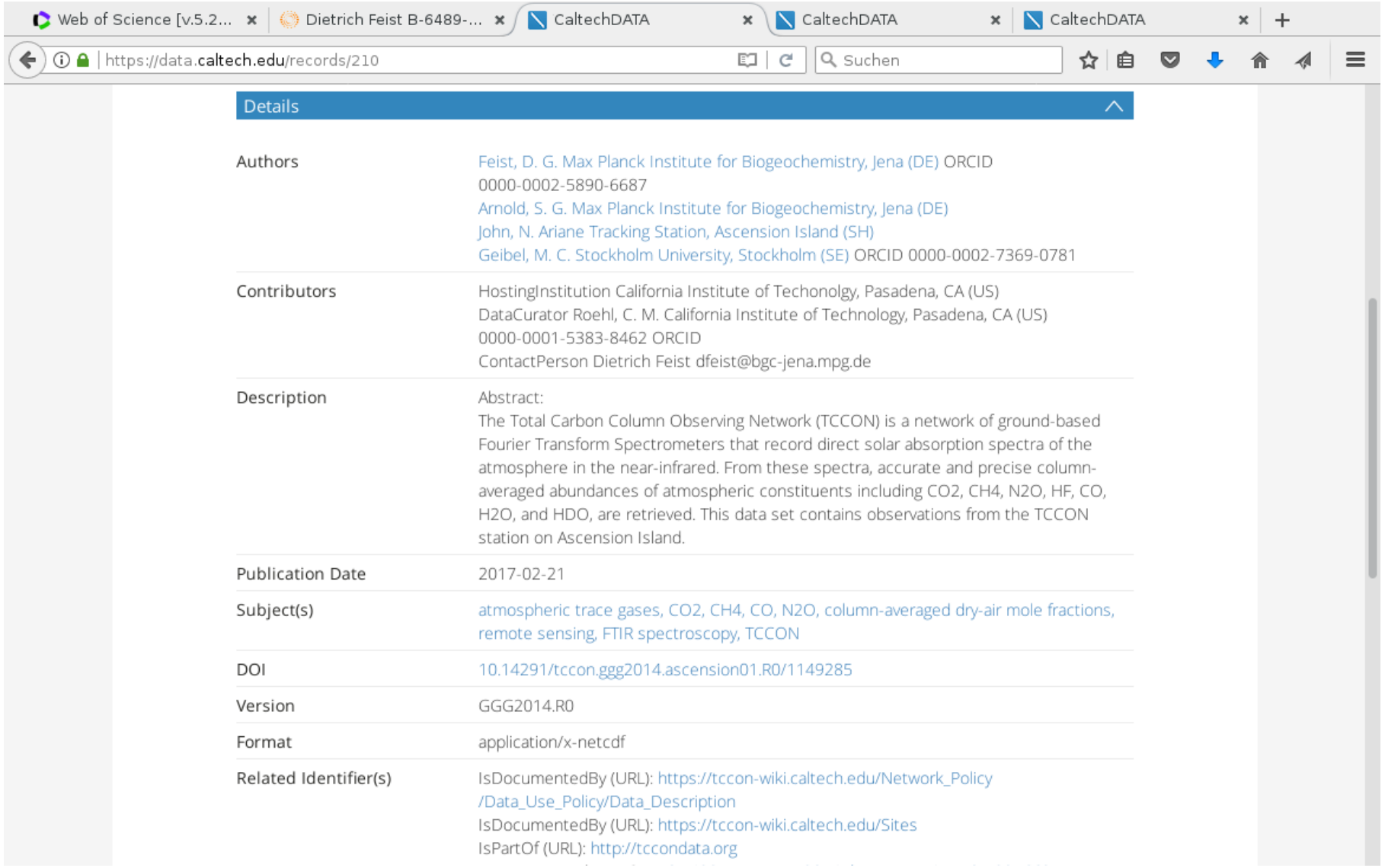
Metadata view @datacite.org



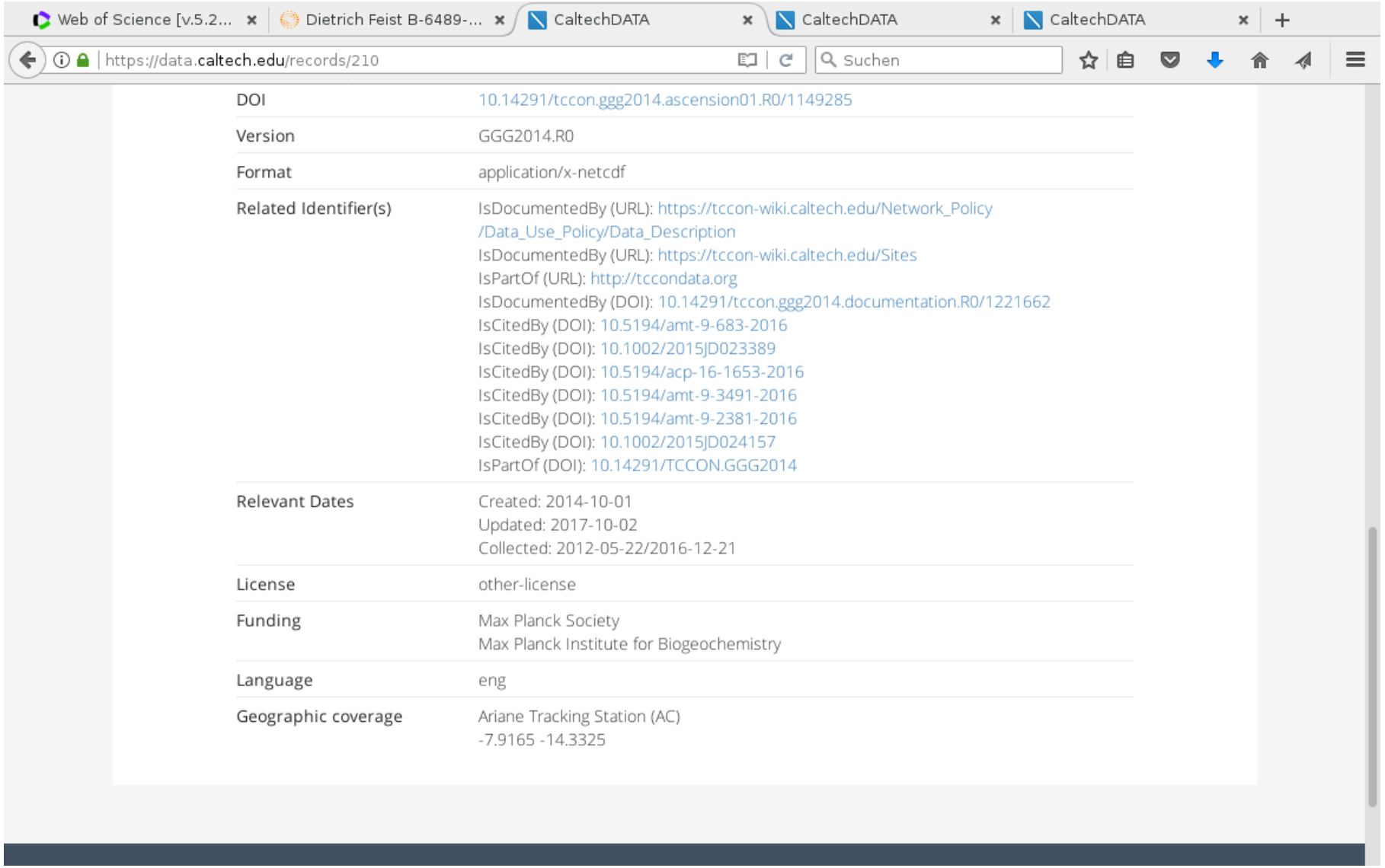
Metadata view @CaltechDATA



Metadata view @CaltechDATA



Metadata view @CaltechDATA



Getting your data cited and its usage tracked

- In theory, only two steps are needed:
  - Get your data users to cite your data set references with the DOI.
  - Get index providers (e.g. Web of Science) to count these references.
- In practice, this turns out to be more complicated.

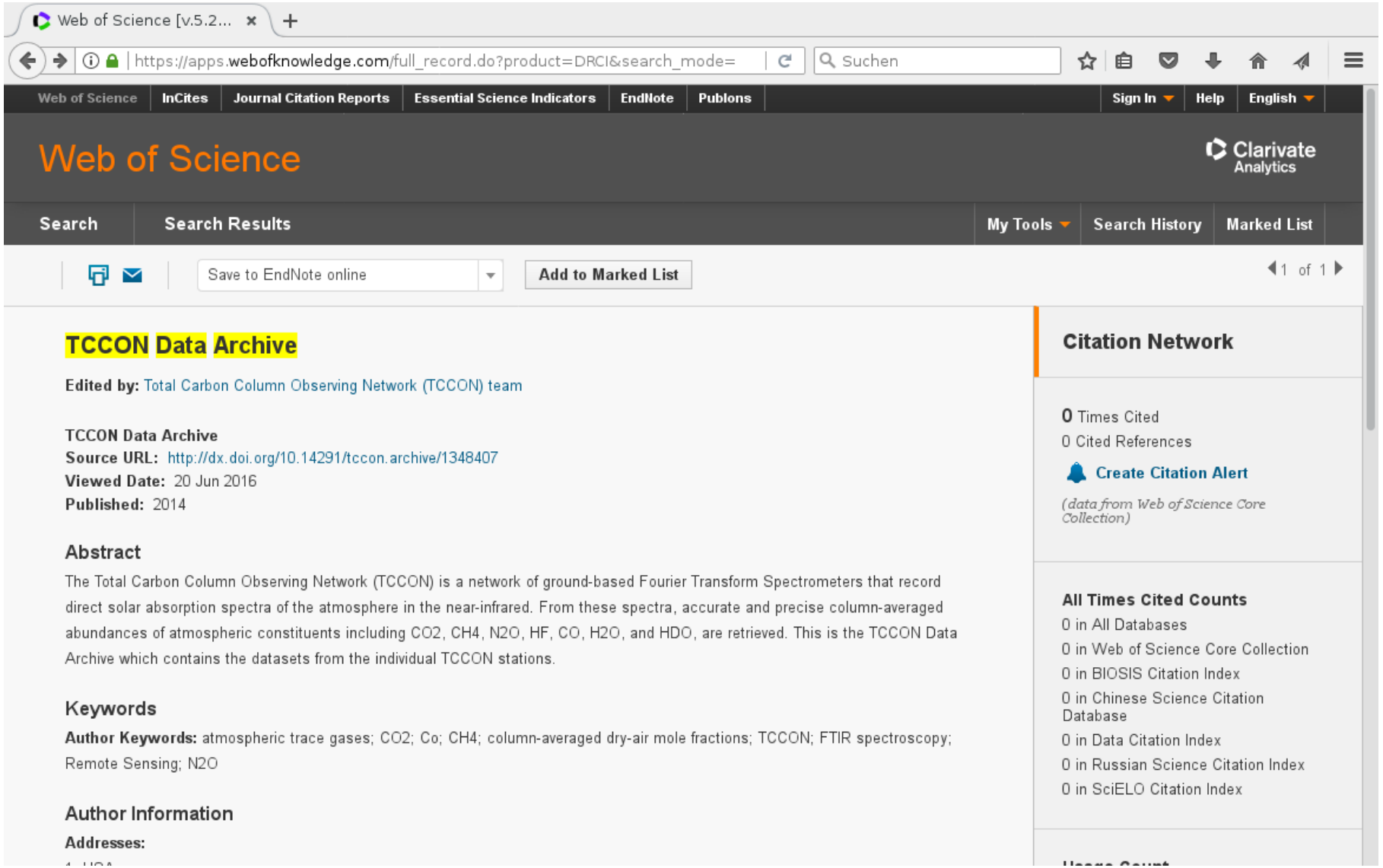
Data Citation Index (DCI) in the Web of Science (WoS)

- Dedicated database for indexing scientific datasets in WoS
- Not part of WoS Core Collection: requires separate subscription!

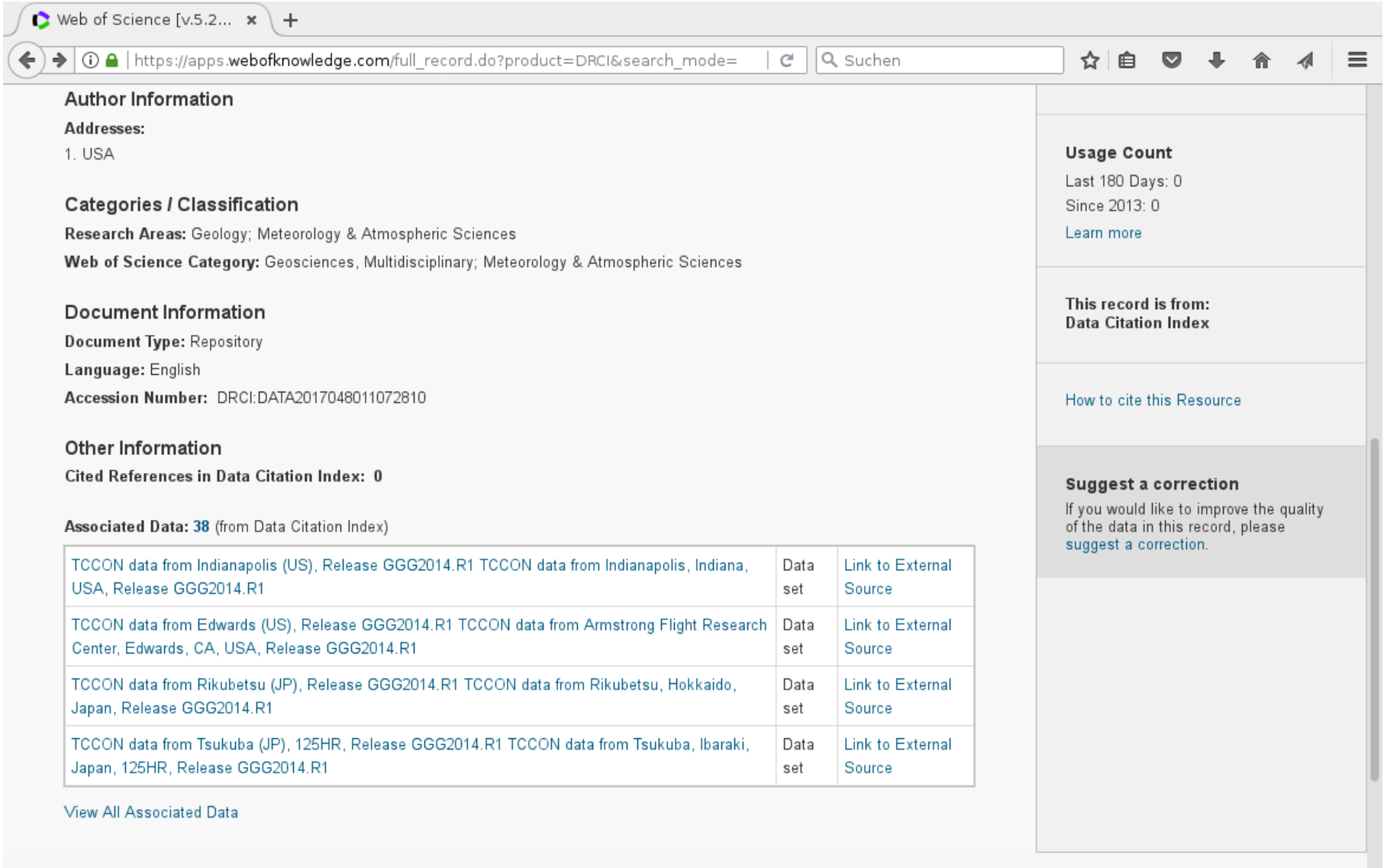
Getting your DOIs and metadata registered in the DCI

- Contact WoS/DCI representative directly.
- Prepare your metadata (largest part of the work!)
- DCI creates a parent repository in the DCI.
- Send your meta data for all datasets.
- Set up a way to update your entries.
- Wait for WoS to start counting ...

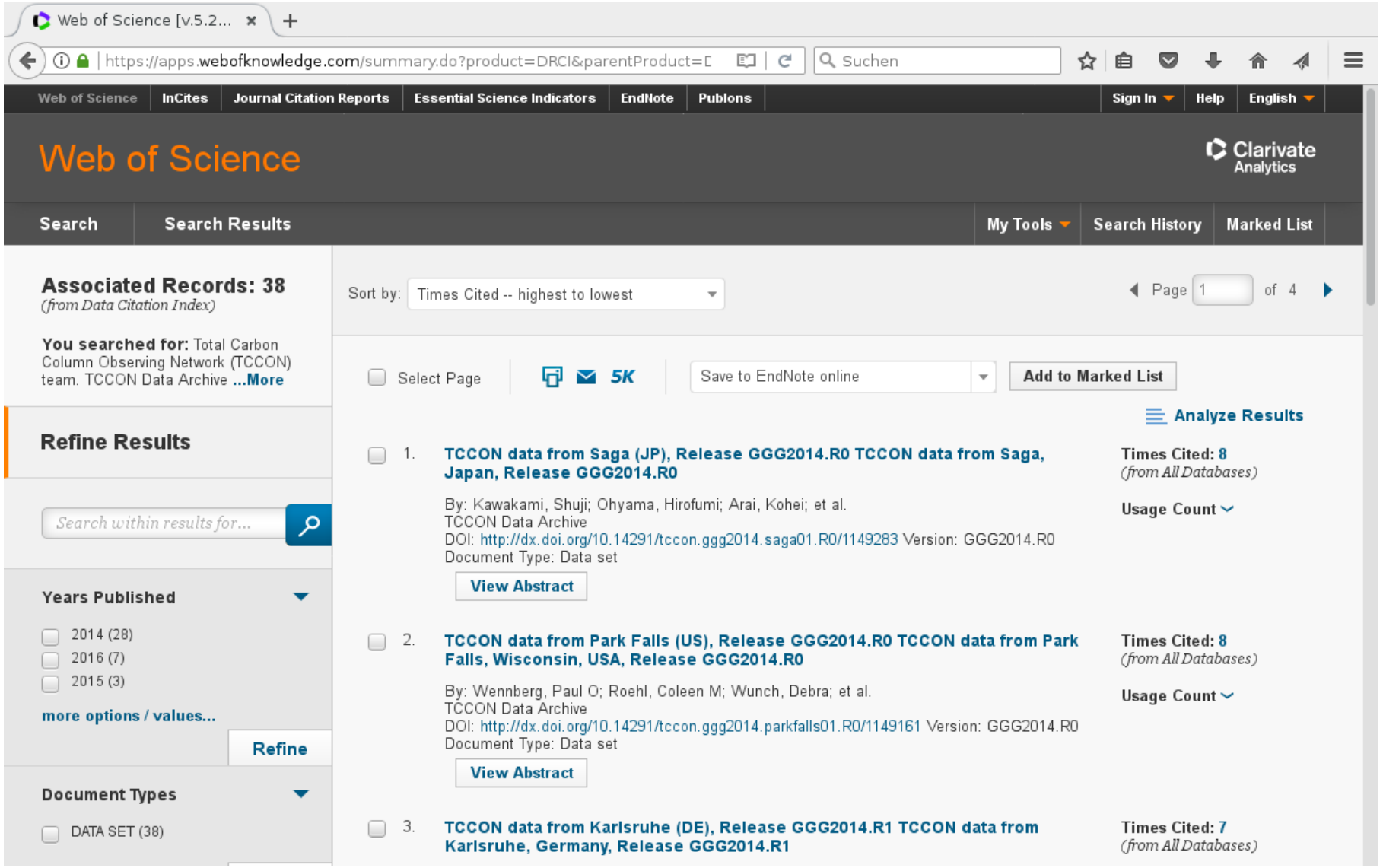
Metadata view @DCI



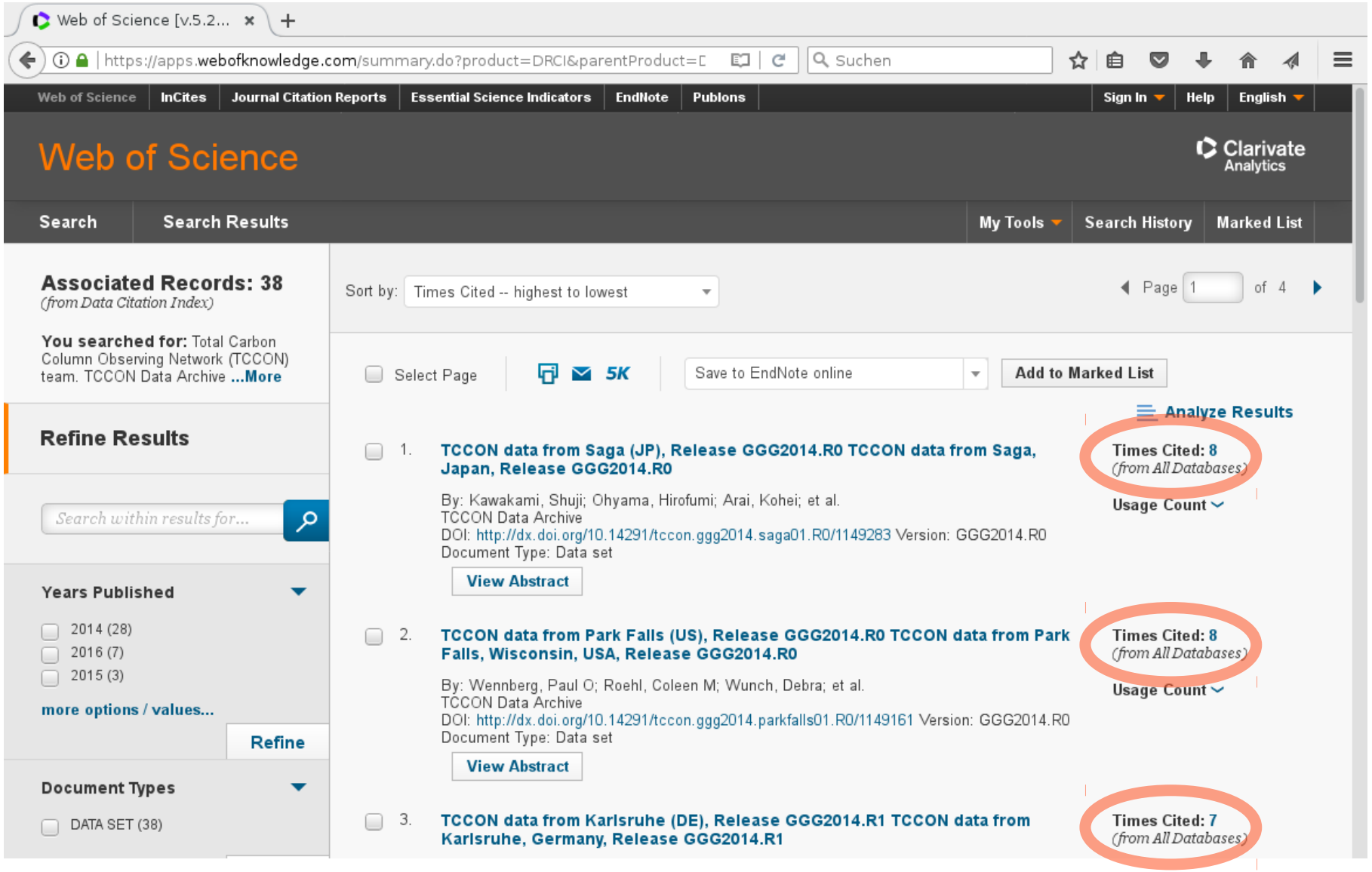
Metadata view @DCI



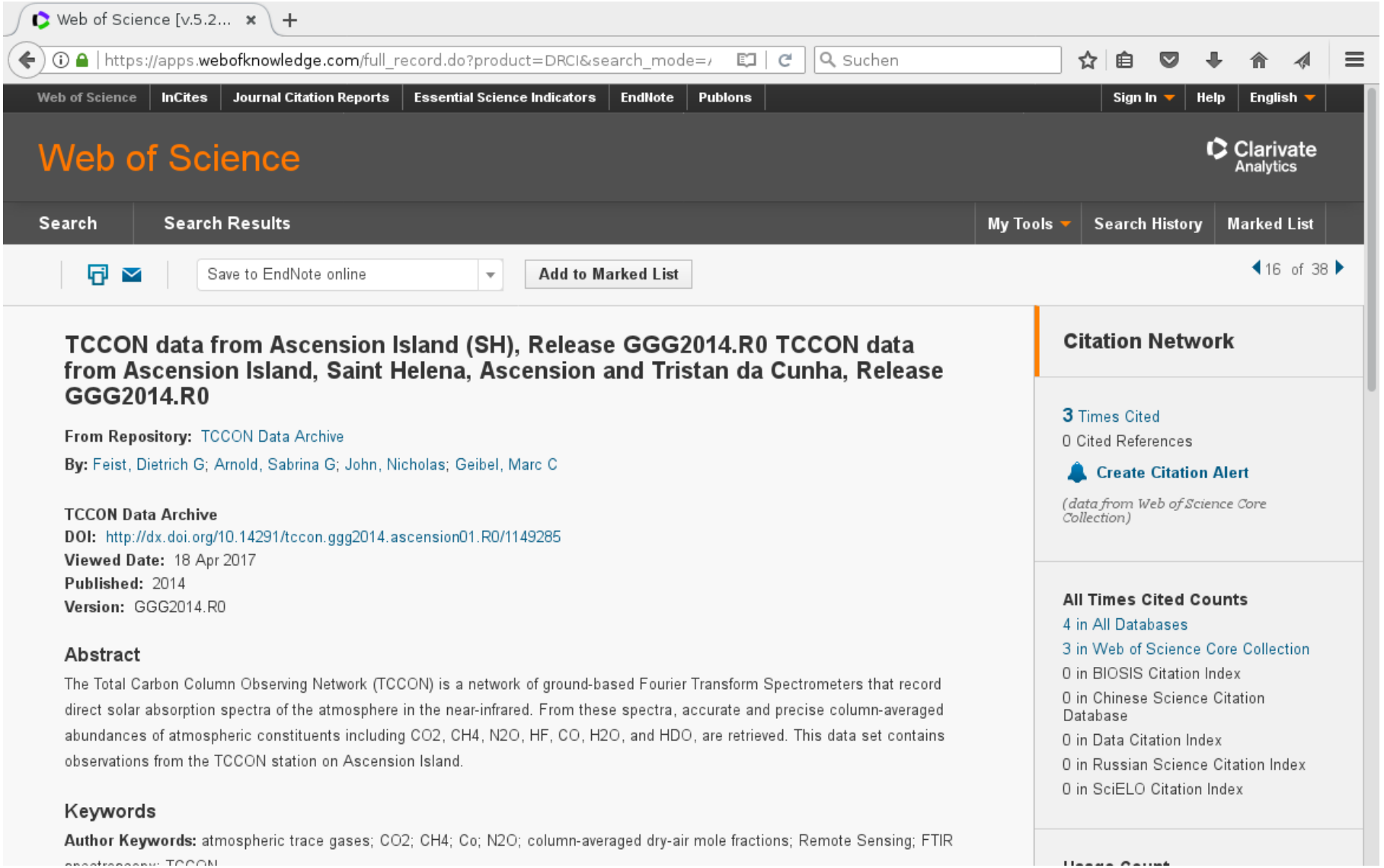
Metadata view @DCI



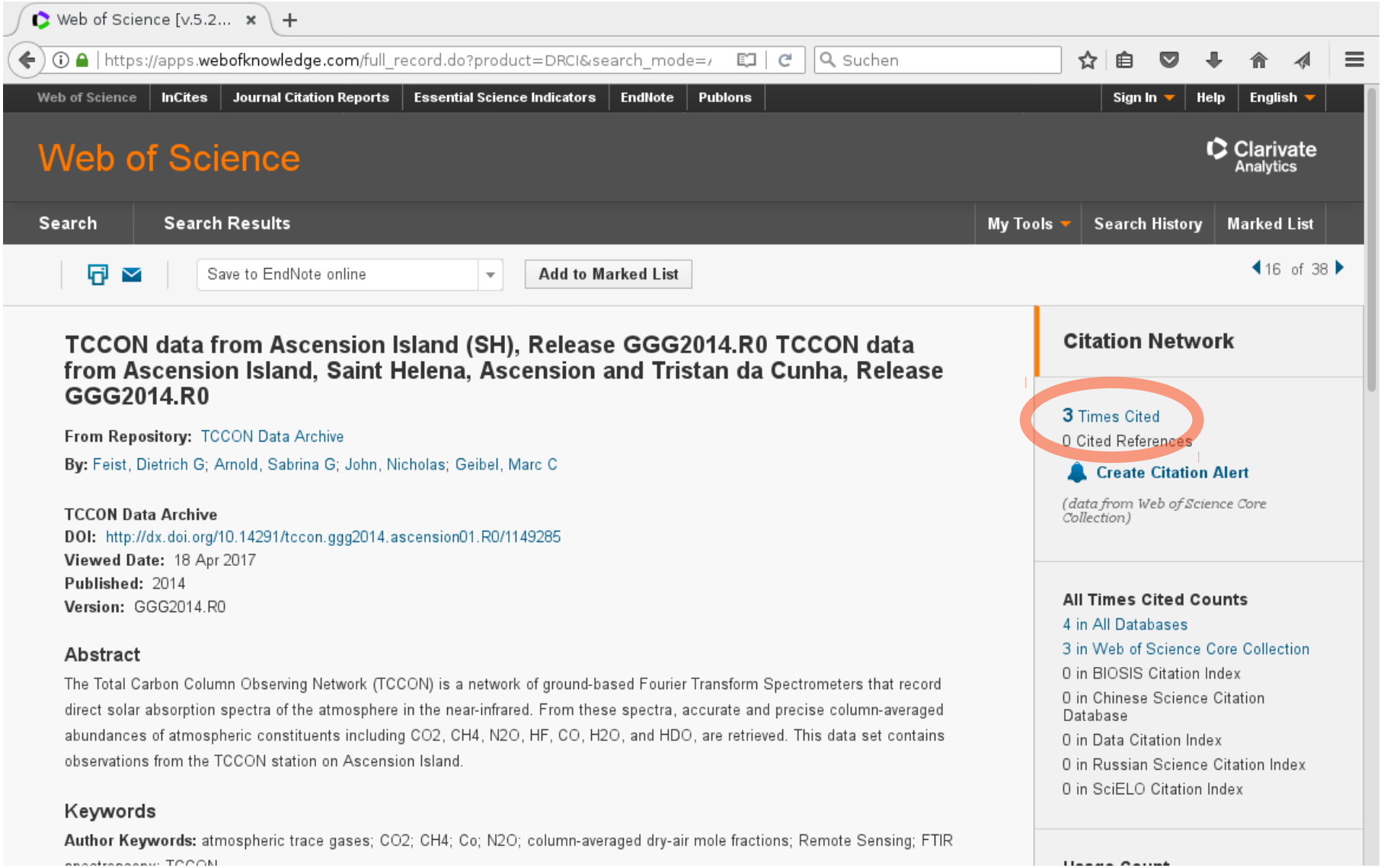
Metadata view @DCI



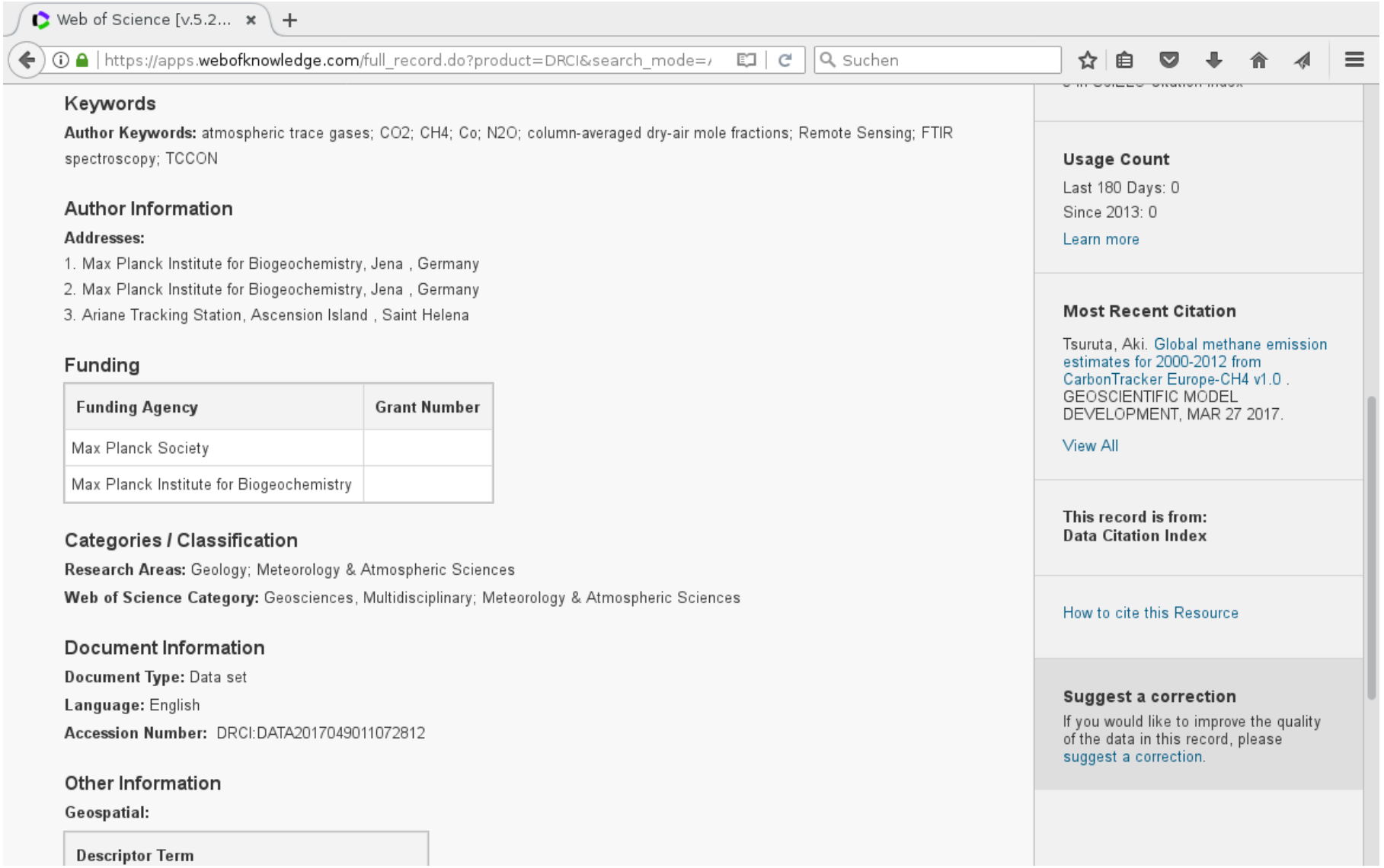
Metadata view @DCI



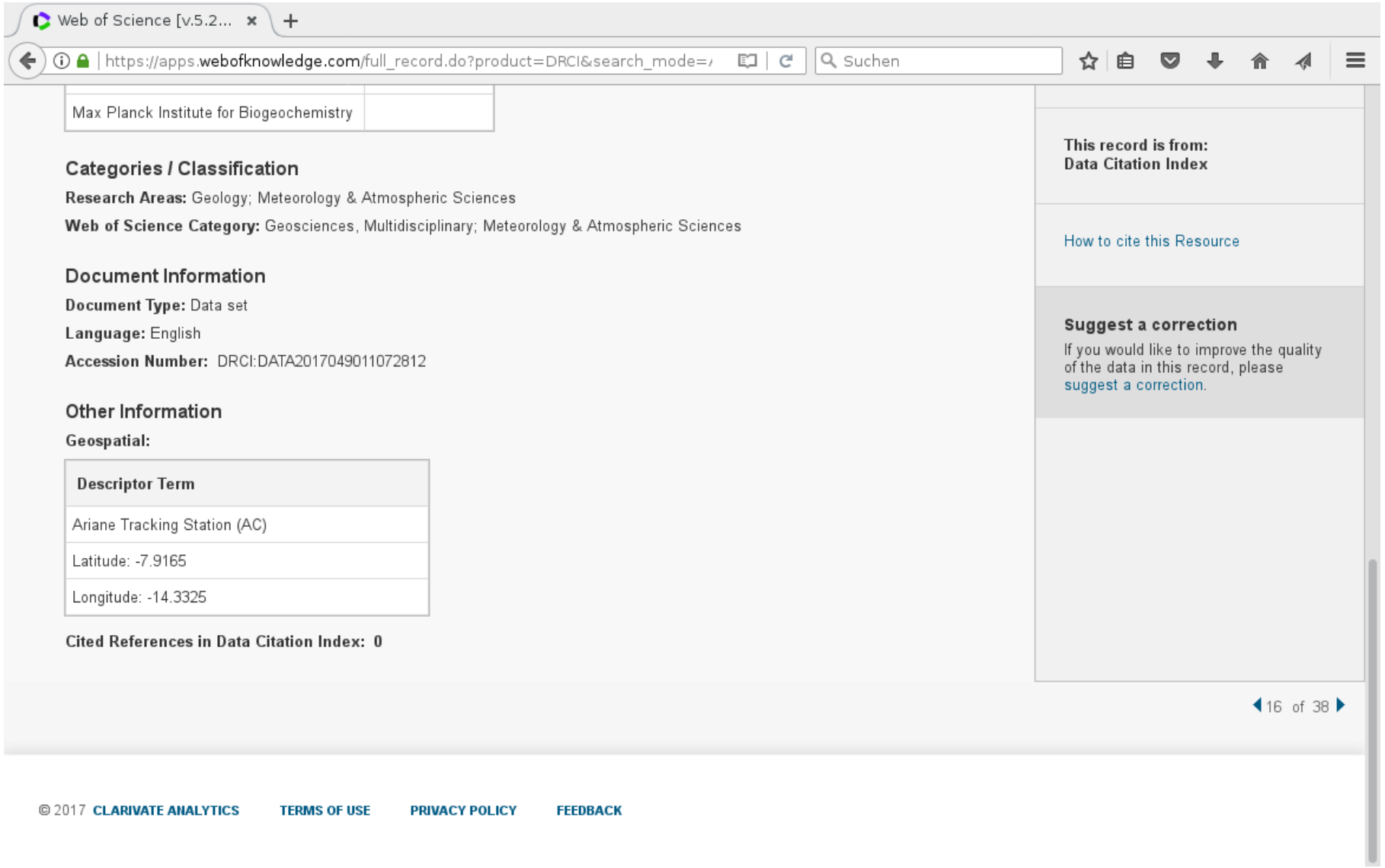
Metadata view @DCI



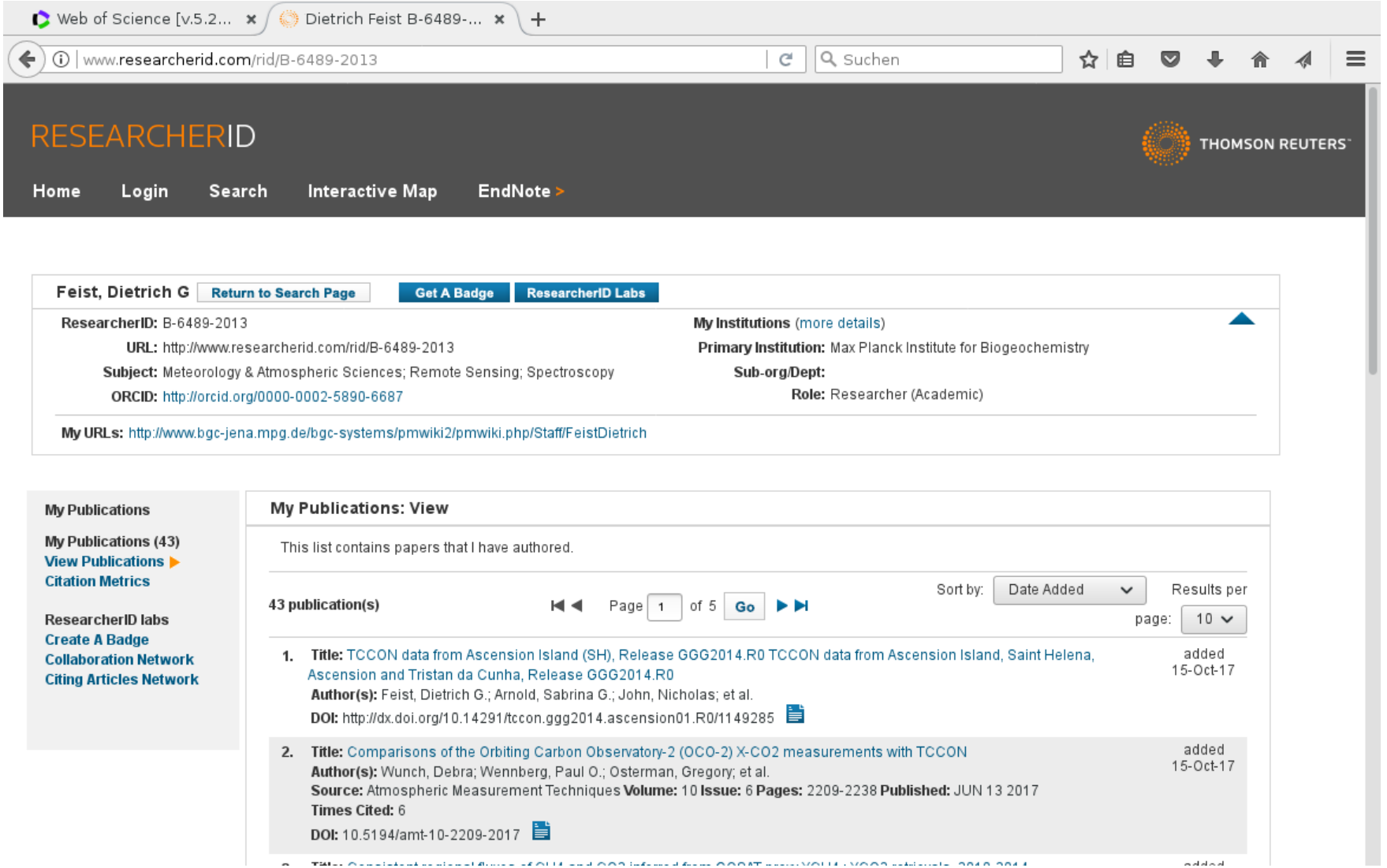
Metadata view @DCI



Metadata view @DCI



Citations to your data sets contribute to your citation metrics



Things we had to learn the hard way

- URLs don't work for citations!
- Make sure you have full control over your own DOIs and metadata!
- Think about metadata from the beginning!
- Set up a reasonable level of granularity that works for
  - yourself
  - your data users
  - index providers
- Your archive may not exist forever.
- Not everybody will follow your data licence or use your data in the way you want.