Making Atmospheric Composition Observational Data FAIR

Gao Chen, Morgan Silverman, Kathleen Dejwakh, Crystal Gummo, Michael Shook, Megan Buzanowicz, Sean Leavor

NASA Langley Research Center

Franziska Stürzl

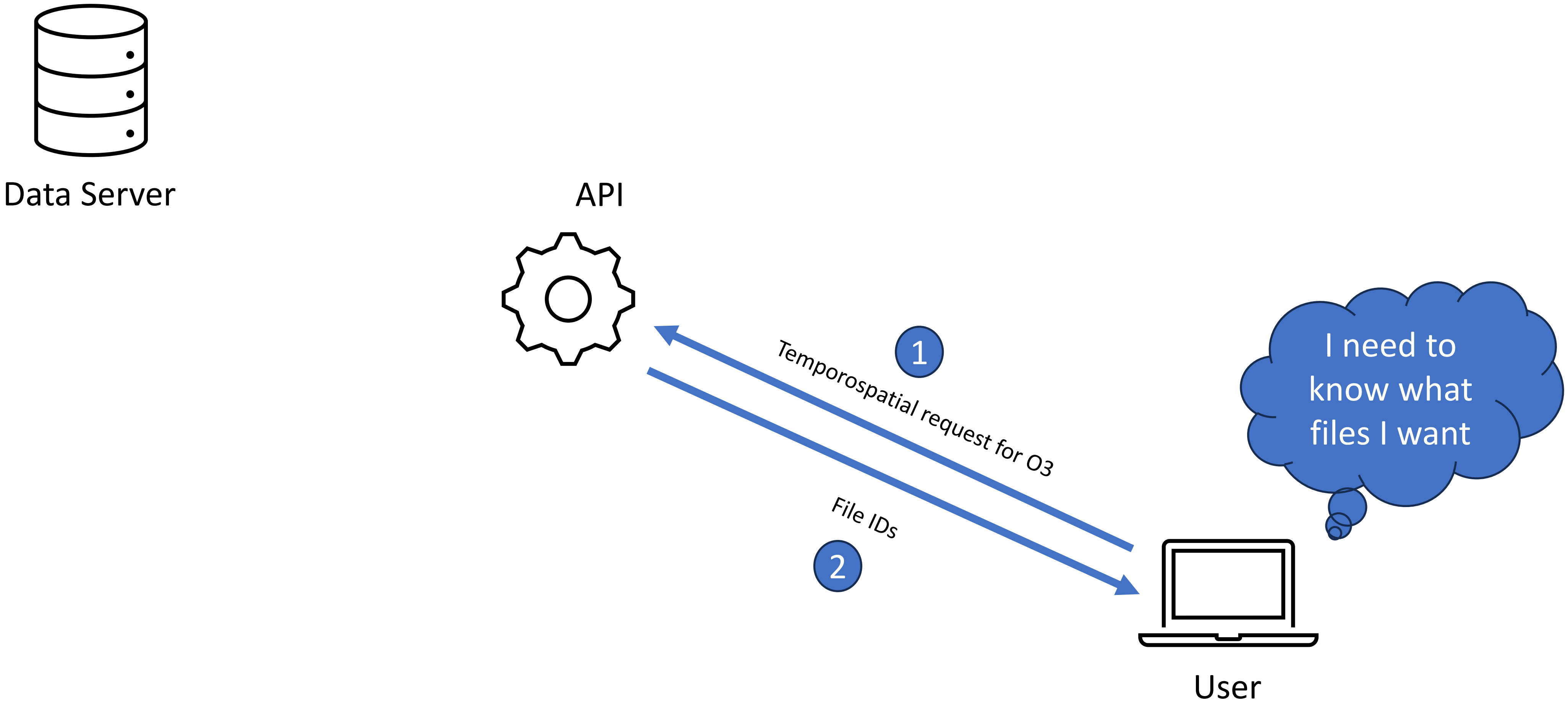
MeteoSwiss (Federal Office for Meteorology and Climatology, Switzerland)

FAIR Data Principles

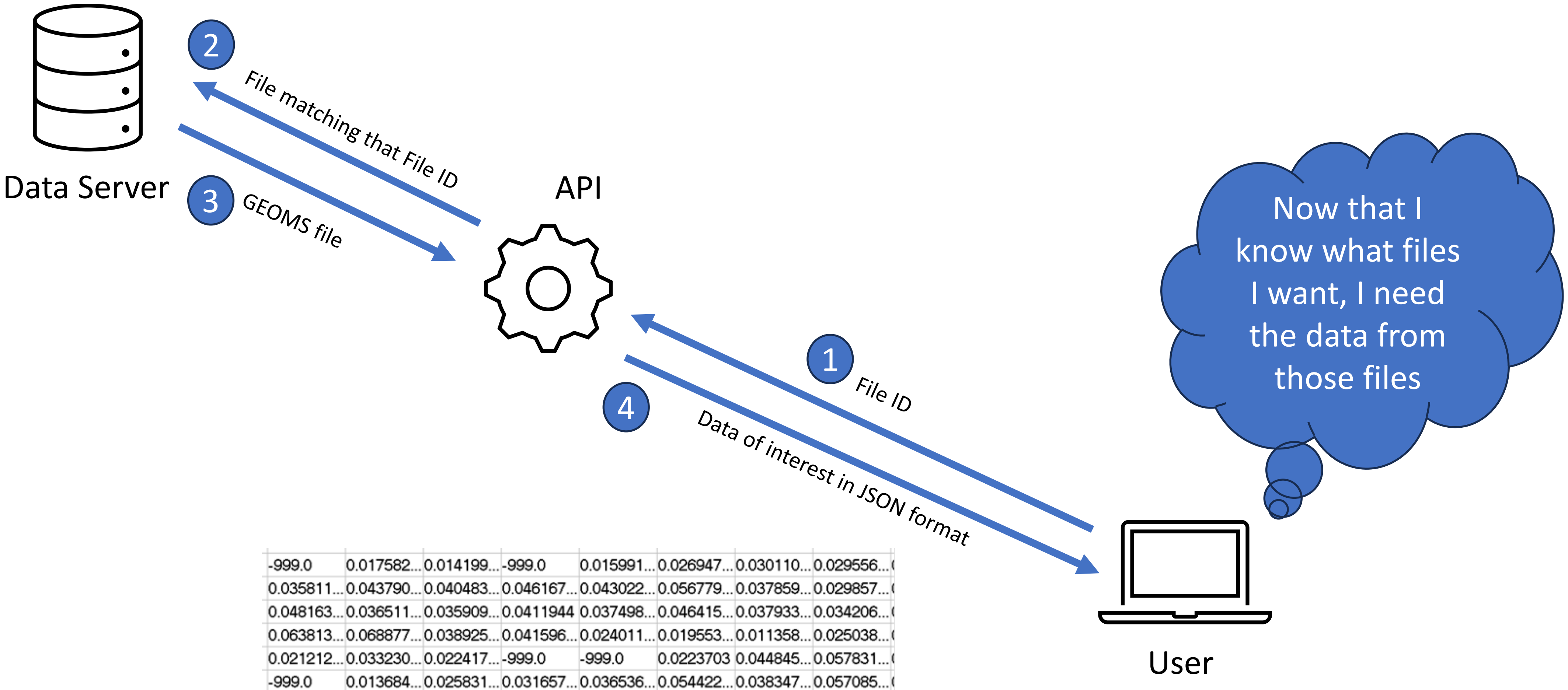
https://www.go-fair.org/fair-principles/

- First published in *Scientific Data* in 2016
  - ✓ rapidly endorsed by research & data science communities
- Intended to substantially lower the barriers for data sharing & use and to improve machine-actionability, specifically by enhancing:
  - ✓ Findability: Human and machine findable
  - ✓ Accessibility: well-defined protocol for data and metadata
  - ✓ Interoperability: data fusion/integration and data use in various applications
  - ✓ Reusability (usability): well-described to enable data properly used
- Making data FAIR is still work in progress
- Essential to implement Open-Source Science and Open data policy

FAIR Data Sharing M2M Example: find Tropospheric Ozone Lidar Network (TOLNet) Data (https://tolnet.larc.nasa.gov)

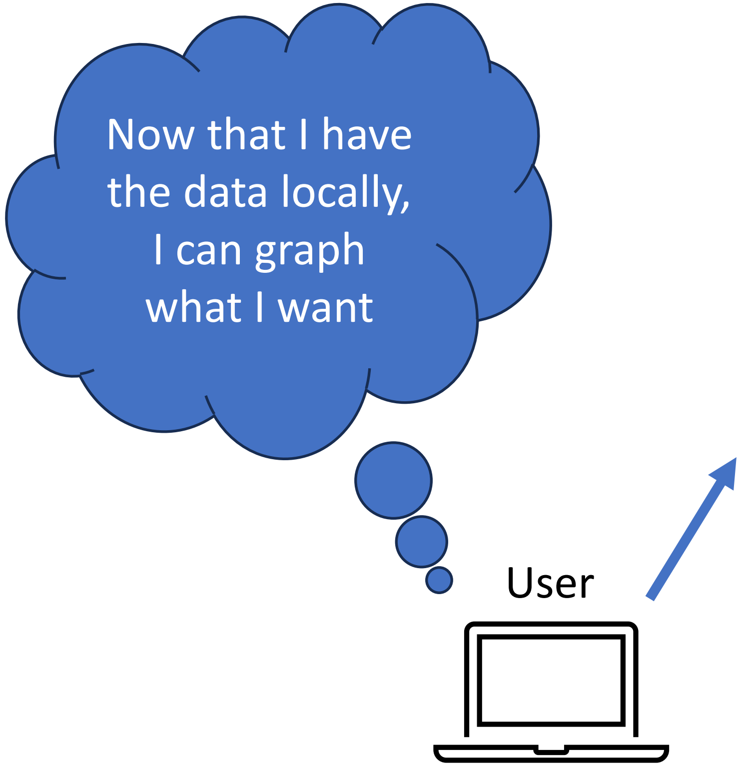


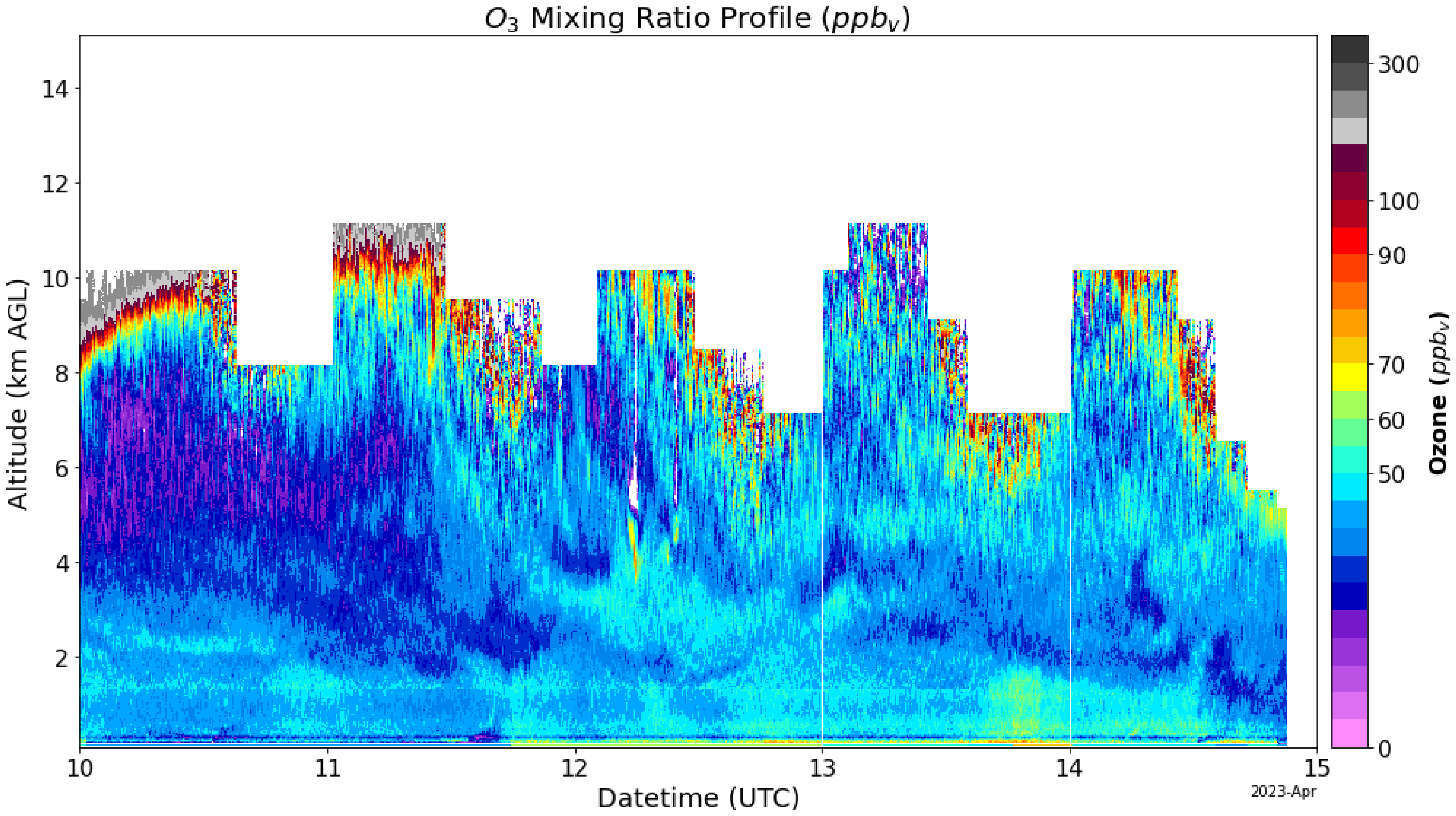
FAIR Data Sharing M2M Example: interoperable access of Tropospheric Ozone Lidar Network (TOLNet) Data (https://tolnet.larc.nasa.gov)



-999.0	0.017582...	0.014199...	-999.0	0.015991...	0.026947...	0.030110...	0.029556...
0.035811...	0.043790...	0.040483...	0.046167...	0.043022...	0.056779...	0.037859...	0.029857...
0.048163...	0.036511...	0.035909...	0.0411944	0.037498...	0.046415...	0.037933...	0.034206...
0.063813...	0.068877...	0.038925...	0.041596...	0.024011...	0.019553...	0.011358...	0.025038...
0.021212...	0.033230...	0.022417...	-999.0	-999.0	0.0223703	0.044845...	0.057831...
-999.0	0.013684...	0.025831...	0.031657...	0.036536...	0.054422...	0.038347...	0.057085...
0.028461...	0.068662...	0.031239...	-999.0	-999.0	-999.0	0.012128...	0.0491939
0.036685...	0.060575...	0.046171...	0.018497...	-999.0	-999.0	-999.0	-999.0
0.041369...	0.104973...	0.050575...	0.031792...	0.019604...	0.016898...	-999.0	0.023881...
0.012317...	0.018789...	0.016852...	-999.0	-999.0	0.034882...	-999.0	0.033992...
0.042673...	0.0814322	0.035534...	0.018709...	0.024948...	0.023957...	-999.0	0.013147...

FAIR Data Sharing M2M Example: reuse Tropospheric Ozone Lidar Network (TOLNet) Data (https://tolnet.larc.nasa.gov)





API

API: Application Programming Interface

- ✓A way to enable computer systems or software to communicate with each other and share data and functionality
- ✓Like a restaurant server between guests and chef
- ✓Platform agnostic
- ✓Reduce user efforts

The TOLNet API

What did the TOLNet API do:

- ✓Handled data request from user's Python code
- ✓Communicated with server to retrieve relevant data files submitted by TOLNet science team
- ✓Extracted data of interest and output data in JSON files (a text-based self-describing format) – minimized user efforts needed to understand the native data format

TOLNet’s FAIR Data Practice

- Science Team:
  - Provided sufficient metadata to support data discovery and data use/reuse by using:
    - A community standard: GEOMS (Generic Earth Observation Metadata Standard) data format (https://evdc.esa.int/documentation/geoms/)
      - well maintained and updated to ensure relevance to data use needs
      - Metadata standards developed by researchers and data scientists with controlled vocabulary clearly describing the measurement and the data
      - widely accepted by the ozone lidar community
  - Provided data use cases to data management team for tool development
- Data Management Team:
    - Findability and accessibility: ensured user interface and API fully leverage metadata and closely collaborated with science team to design and develop tailored tools
    - Interoperability and Reusability : developed data checker to ensure data and metadata integrity

It takes a village to make data FAIR

- Data providers
  - Provide rich metadata through community data format standards
  - Use standardized metadata when available and register relevant persistent identifiers, e.g., ORCID
  - Contribute to development of domain-relevant metadata standards
- Data publishers (e.g., NASA Distributed Active Archive Centers or “DAACs”)
  - Ingest and curate data
  - Assign persistent identifiers, e.g., data DOIs
  - Provide secure stewardship and maintain/improve distribution tools for data and metadata, including M2M tools
- Data Scientists
  - Maintain/update the existing metadata standards and support their application
  - Develop domain-relevant metadata standards in collaboration with data providers, e.g., measurement specific metadata

Metadata

- Data about data or description of data in various aspects, e.g., smart phone photo location, time/date, camera settings
- Metadata for FAIR
  - Most frequently used noun in https://www.go-fair.org/fair-principles/
  - "Typically, the FAIRification process begins when a community of practice considers its domain-relevant metadata requirements…" (https://www.go-fair.org/how-to-go-fair/)
  - FAIRness critically depends on "rich metadata", which are accurate, standardized, and compatible with data system
    - Basic metadata: what, why, who, where, how (e.g., instrument), and provenance
    - Domain-relevant metadata examples: assumption, sampling, detection, processing, and limitation

Metadata challenges

Variable Descriptions

- Measurements often not representative or incomplete, e.g., blind men and an elephant
- Traditional approaches often inadequate to address the complexity (application of ontological approach encountered ambiguities due to its inherent limitations)
- Data system limitations on potential changes

“Rich” Metadata

- Sufficient but not overkill for data discovery and distribution tools
- Application and user dependent requirements

Domain-relevant metadata standards

- Expertise required in both data science and measurement to determine the information needed
- Controlled vocabulary for accurate description

I-ADOPT Recommendations for variable representation

https://elib.dlr.de/148699/1/I_ADOPT_recommendations.pdf

I-ADOPT Recommendations:

- ✓ Variables in both a human- and a machine-actionable format
- ✓ Minimum reliance on external free-text documentation
- ✓ "use of FAIR terminologies (e.g., controlled vocabularies, ontological relationships)"
- ✓ Follow decomposition approach to translate human-readable to machine actionable
- ✓ Preferably reuse existing I-ADOPT terminologies
- ✓ Mapping/translations from other schema

- Implementation of I-ADOPT recommendations can be found at: https://vocabulary.actris.nilu.no/skosmos/actris_vocab/en/
- WIGOS (WMO Integrated Global Observing System) and CF standard names are under development following I-ADOPT recommendation
- Similar efforts include NASA ESDIS UMM-Var

Ontology and Literature Terminology Balance

Literature terminology – commonly used in literature publications

- Popular community terms, e.g., NOy (reactive nitrogen species), SSA (single scattering albedo – ratio of scattering over extinction)
- Not systematically derived, sometimes deviate from English definition, e.g., “aerosol”.
- Not always interpretable using an ontological approach

Ontology

- Widely used in data science for machine readability
- Difficulty with complicated measurement variables, ambiguity, subjectivity, and consistency issues in implementation
- Words alone cannot uniquely identify some VOC variables

Atmospheric Composition Variable Standard Names

ESDS-RFC-043: Atmospheric Composition Variable Standard Name Convention

Synergistically developed by instrument scientists, modelers, and data scientists to enhance discoverability, interoperability, and usability

Successful implementation in NASA Atmospheric Composition field campaigns

Translatable to other variable naming schema, ACTRIS

Approach generally consistent with I-ADOPT

- Decomposition based on physical /chemical properties and measurement characteristics
- Controlled literature-friendly vocabulary with description
- Consistent and predictable structure

Actively maintained, supported, and updated

Checker under development

A NASA ESCO Convention

Atmospheric Composition Variable Standard Names Examples

- **General Structure:**

MeasurementCategory_CoreName_AcquisitionMethod_DescriptiveAttributes

- **Trace Gas**

Gas_CO2_InSitu_S_DMF: *an in-situ measurement of CO2 gas reported in molar fraction with respect to dry air*

Gas_NO2_SlantColumn_S_CNDAMB: *a remote sensing measurement of slant column NO2 gas reported column number density with respect to ambient air*

- **Aerosol**

AerMP_NumSizeDist_InSitu_RHd_Aerodynamic_Coarse_STP: *an in-situ measurement of aerosol particle number size distribution reported at reduced relative humidity derived from an aerodynamic sizing technique for coarse-mode aerosols at standard temperature and pressure*

AerOpt_Absorption_InSitu_red_RHd_Bulk_AMB: *an in-situ measurement of absorption measured at a red wavelength under reduced humidity conditions with a bulk aerosol particle size range reported in ambient conditions*

Simple Takeaways

- Work with existing data formats to make data more FAIR
  - ✓Provide useful and accurate information to required metadata entries
  - ✓Raise questions if unsure where to put essential information
- Use standardized metadata when available
  - ✓Use NASA ESCO Atmospheric Composition Variable Standard Names whenever possible
  - ✓Use WMO WIGOS unit standards
- Provide feedback to data scientists for improving metadata standards
  - ✓Report issues encountered in reporting metadata, e.g., variable standard names, measurement relevant information
  - ✓Provide research examples (use cases) showing how data are used and the targeted user groups
  - ✓Try to understand from the data science perspective and provide constructive suggestions

Low-Hanging Fruit

- Lack of standardized unit system limits data interoperability and usability
  - Non-standard units often used, some may not be human-readable
  - A wide variety of units used for dimensionless quantities: none, N/A, fraction, ratio, unitless…
- Use standard measurement units, e.g., WMO WIGOS Metadata Representation: Measurement Unit (http://codes.wmo.int/wmdr/unit)
  - Adopted SI units and units widely used in literature
  - URL-friendly
  - Properly defined
  - Conversion to SI units provided
  - SI equivalent unit provided, e.g., “ppmv, Unit of volume fraction, equivalent to SI derived unit: 1E-6”

WIGOS Metadata Development Example:

Sample treatment - (http://codes.wmo.int/wmdr/SampleTreatment)

- Domain relevant metadata standard
- Used literature friendly, controlled vocabulary for in-situ sample processing before analysis/detection
- Developed by Ad-hoc working group of well-established aerosol experts, user, and data scientists (https://github.com/wmo-im/wmds/wiki/wg-acv):
  - To identify the sample treatment methods commonly deployed in field measurements
  - To ensure the controlled vocabulary is adequate
- Accurate description, applicable to multiple types of measurements
- Selected examples: 5 out of 16

Name	Notation	Description	Types	Status
Chemical conversion	conversion	Sample is chemically converted before analysis	Concept	stable
Drying	drying	Sample is dried before analysis	Concept	stable
inapplicable	inapplicable	Inapplicable or no sample treatment	Concept	stable
other	other	Inapplicable or no sample treatment	Concept	stable
Thermal decomposition	decomposition	Sample is thermally decomposed before analysis	Concept	stable

Needs and Directions

Making data FAIR requires synergistic collaboration between researchers, data centers, and data scientists

- Improve/develop detailed guidelines, standards, conventions, and best practices, especially those specific to measurements/instruments to support data use (reuse)
- Improve/develop tools to support both targeted communities and communities at large
- Develop data file checker to ensure data and metadata integrity
- Modify/Develop FAIRness checkers capable of handling community data standards
- Standards need to be maintained, updated, and supported by dedicated staff

Franziska

data integration team at MeteoSwiss (Federal Office for Meteorology and Climatology, Switzerland)

- **My Education:**
  - ✓ B.Sc. (Germany) in Civil Engineering
  - ✓ B.Sc. (Germany) in Mathematics with additional courses in linguistics
- **What I do:**
  - ✓ Working with WMO ET-Metadata to improve the current WIGOS metadata standards and to develop new standards as needed
  - ✓ Reviewing measurement concepts against internal and external requirements, including WMO guidelines and drafting recommendations
  - ✓ Supporting users related to data warehouse service from data providers to end users
  - ✓ Developing scripts to manage the content of a WMO metadata repository
- **What I like:**
  - ✓ Enjoyable working atmosphere
  - ✓ Diverse and challenging projects, including investigating and trouble-shooting issues
  - ✓ Communications with a broad range of people

Megan

Atmospheric Science Data Center (ASDC) at NASA Langley Research Center

- **My Education:**
  - ✓ M.S. in Atmospheric Science
  - ✓ B.S. in Meteorology with minors in Environmental Geography, Mathematics
- **What I do:** Lead in the data ingest, archive, and distribution of sub-orbital campaigns for long-term preservation and distribution
  - ✓ Creating metadata records from the incoming data and ancillary information
  - ✓ Work with science teams to ensure data discoverability and interoperability
- **What I like:**
  - ✓ I get to ensure these useful and valuable datasets are properly archived and distributed to the public, especially for those legacy datasets which haven't been available to the public in the past.
  - ✓ Get to work at NASA

Michael

Chemistry and Dynamics Branch at NASA Langley Research Center

- **My Education:**
  - ✓ B.S. in Atmospheric Sciences with a minor in Mathematics
- **What I do:**
  - ✓ Developing data acquisition and control software tools for Aerosol measurement system data processing
  - ✓ Processing and QA/QC field observation data
  - ✓ Improving the current data format and metadata standards and developing best practices
  - ✓ Supporting science teams to ensure data product integrity, including data file format and metadata checks, generating value-added data products, and trouble-shooting various issues during the process

What I like:

- ✓ Good people and work environment
- ✓ Challenging and interesting projects – every day is different