Adoption of RDA DFT Data Model & the DTR to the Description and Structuring of Atmospheric Data

January 13, 16
Final Report

Principal Investigator: Aaron Addison

Team: Cynthia Hudson-Vitale
Rudolf Husar
Holly Weller
Sarah Weller
Kari Hoijarvi

Table of Contents

Background page 3

Project Proposal 3

Adoption Goals 4

Methods 4

Usability 7

Data Type Registry Feedback 8

Adoption Feedback 9

Process Feedback 9

Appendices

- Appendix A: Facet/Value Pairs Registered into DTR 10
- Appendix B: RDF/XML Sample Record 13

Background

The discovery, access and understanding of datasets is achieved by specifying a number of the attributes associated with each dataset.

DataFed is a distributed, web-services-based computing environment for accessing, processing, and rendering environmental data in support of air quality management and science. DataFed includes (1) the Air Quality Community Catalog (AQComCat), (2) a portable Web Coverage Service (WCS) data server software package and (3) a web-based, service-oriented workflow program for the processing and visualization of multidimensional data.

The AQComCat is a catalog of 50 heterogeneous datasets found in the DataFed.net system. The AQComCat was implemented by the Center of Air Pollution Impact and Trend Analysis (CAPITA) at Washington University in St. Louis during the period of 2009-12. In addition to finding data on atmospheric composition and emission, human population and other fields, AQComCat also serves to seamlessly access numerical and image datasets through the federated data system, DataFed.

Within DataFed, each observation-derived dataset has an array of attributes grouped into 'facets'. Both the facets (~10) and choice of attributes within each facet (about 8 attribute choices per facet) are common to all data holdings in the data system and serve as the input to the AQComCat. A user of the catalog can browse and filter data by faceted search mechanism. Once found the desired dataset, the catalog shows all the data attributes expressed in a common language.

However, broader interdisciplinary use of the catalog for finding and understanding data is hampered by significant limitations

- The naming of the attributes is in disciplinary jargon
- The data attribute descriptions are terse or incomplete
- The facets and attributes combinations are not shared with other data systems

Project Proposal

To address these limitations, the grant team sought to implement a mechanism to make the implicit knowledge of the data more explicit. The solution involved adopting the output of the Research Data Alliance's data type registry working group and the data foundations and terminology working group. Data 'typing' is the characterization of data structure, contexts, assumptions and other information needed to describe and understand the data.

The ‘types’ need to be:

- Defined and understood by data producers and consumers
- Types should have multiple levels/granularity –single observation to data sets
- Each type is to have a PID
- Permanently associated with the data they describe
- Standardized (OGC, ISO), unique (PID), and discoverable

Data typing should aid the discovery, understanding, sharing and reuse of data across domains. It should also:

- Automate the processing of large data collections
- Be machine readable
- Allow the creation of complex data types (derived data types)

Adoption Goals

Though much effort has been placed on developing metadata at the individual dataset level and in facilitating the accessibility among air quality researchers to the data within DataFed, standard terminology and models that facilitate data reuse and discoverability outside of the atmospheric science discipline have not been fully-developed.

The goals of this project were to:
1) To apply the DFT core and basic terms and DTR to the existing AQComCat dataset descriptors and facets by comparing the terms to existing air quality terminologies for any overlap, identifying any missing descriptors, and finding exact matches.

2) To test the DFT core and basic terms by adding a new data source to the AQComCat. This will evaluate the ability of the terms to integrate with new sources.

3) To assess the comprehension of DFT and DTR terms:

- in the AQComCat among researchers **within** the air quality discipline who are users of DataFed and community developers of the AQComCat.
- in the AQComCat among researchers **outside** of the air quality domain by conducting pre- and post-usability and information comprehension studies.
- through ongoing, periodic monitoring and evaluation of downloads, citations, and other measures of use.

Methods

The first task undertaken by the grant team involved adopting the data foundation and terminology (DFT) data model to the AQComCat data. This primary step established a common language for all grant team members to use when discussing various components of the project. To become DFT compliant, the team began by using the existing data model of the DataFed infrastructure (see Figure 1).

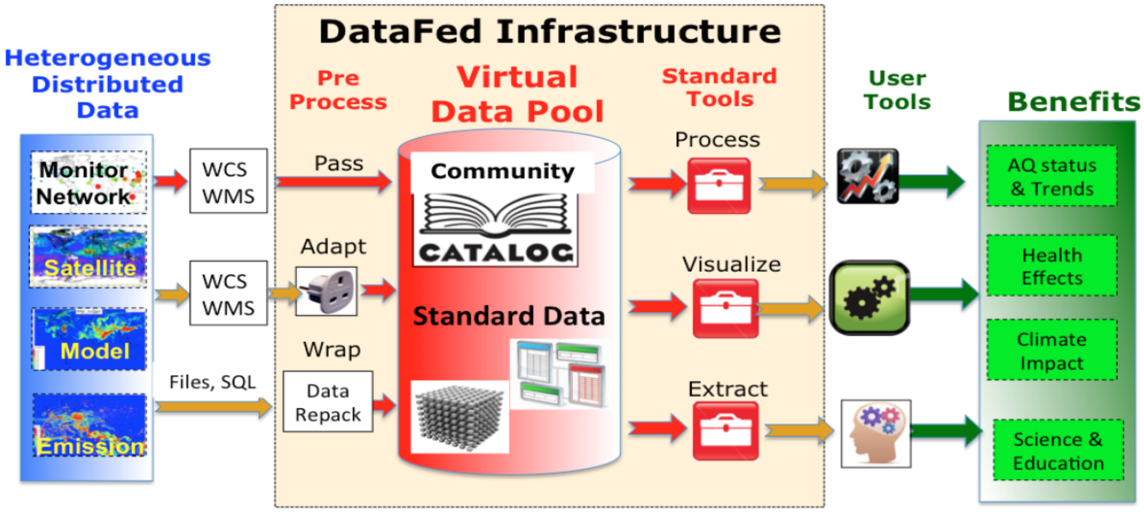


Figure 1: DataFed Data Model

As mentioned previously, the DataFed system is not a data producer, but rather a data distributor. The model created in Figure 1 reflects the data transformations and metadata creation that takes place in the DataFed system[1]. This model was then compared and transformed to be in line with the DFT guidelines (see Figure 2).

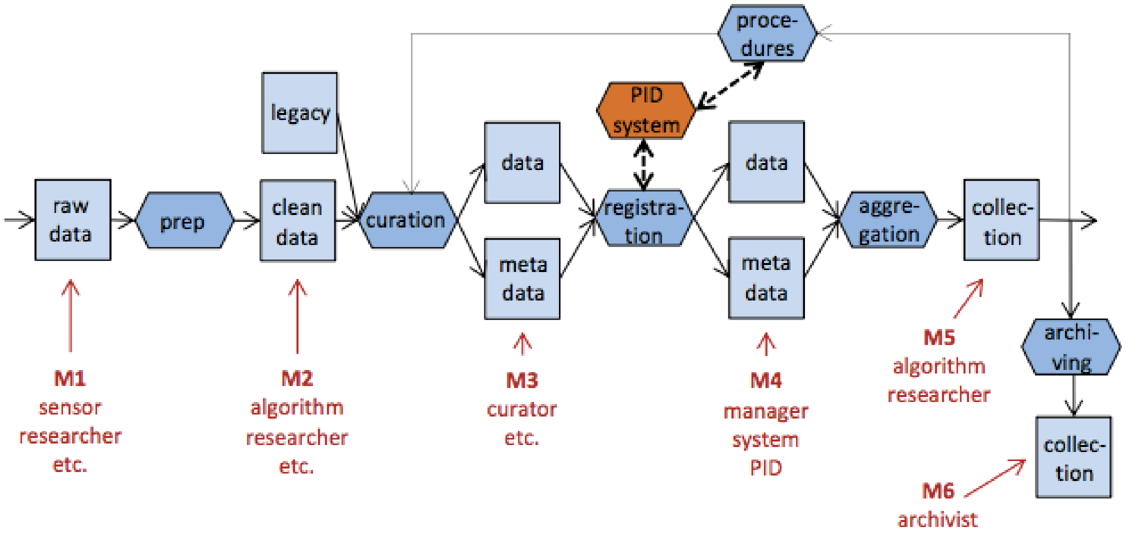


Figure 2: RDA DFT Data Model

[1] http://datafedwiki.wustl.edu/index.php/DataFed

The transformation was undertaken by comparing the existing terms of DataFed model to the DFT. If an exact term was found, it was used in the DFT compliant data model. If an exact term was not found, a similar term was used based upon the descriptions provided by the DFT. A final DFT compliant data model was developed (see Figure 3).

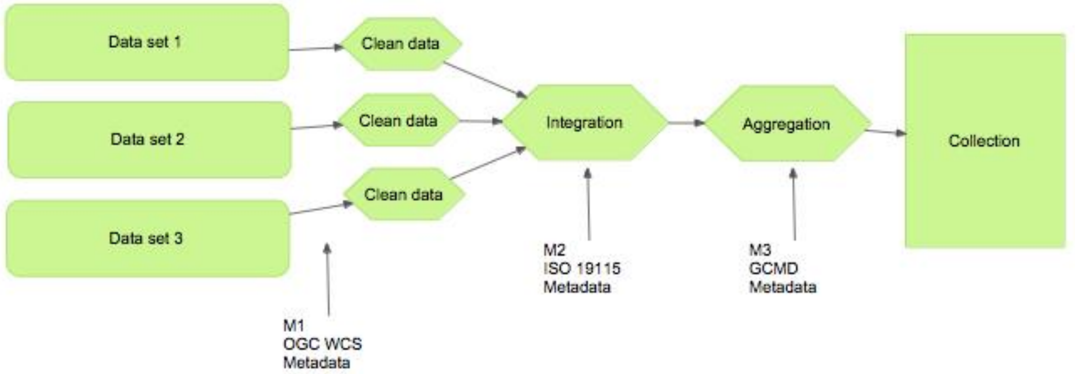


Figure 3: DataFed/RDA DFT Data Model

After the data model was compliant with RDA DFT recommendations, the project team began the process of typing the data using the outcomes produced by the RDA Data Type Registry (DTR) working group. To initiate this, the appropriate air quality facets and values were inventoried in the AQComCat. In total, seven facets and 67 values were found to be requiring some form of data typing (see Appendix A for entire list).

These data types were then registered into the Data Type Registry (DTR) developed by the Research Data Alliance Data Type Registry working group. It was decided to register each of the types as a primitives, rather than derived for the purposes of this pilot. This allowed us to focus on the definitions and working within the DTR as opposed to registering each primitive and then forming relationships to create the derived types.

Once the types were registered, the metadata staff and the data management project manager took the OGC compliant, ISO 19115 metadata records that described the datasets and transformed them into RDF/XML. This allowed the XML records to be machine readable and semantically linked. To retain the ISO 19115 and OGC domain metadata information, an OWL ontology for expressing ISO 19115 developed by Drexel University[2] was used. The links to the DTR were incorporated into the new RDF/XML records (see Appendix B for sample RDF/XML records).

Next, the AQComCat was modified to incorporate the Air Quality DTR records. These were accessed via the DTR API via the ‘get’ command and fed into the revised AQComCat using a script and the unique ID that was created for each datatype in the DTR.

[2] http://www.w3.org/2001/sw/wiki/OWL

Other enhancements of the revised AQComCat include the addition of the spatial vertical data type, which allows the user to understand in what spatial domain the dataset was collected (column, layer, none). Also, the revised AQComCat allows the user to sort the datasets according to various parameters or data types. This feature was added using Javascript.

Finally, to increase the discovery of the AQComCat an OCLC record was created and uploaded to WorldCat, by publishing and production staff in the WUSTL Libraries. OCLC WorldCat is the world's largest catalog of library holdings, content, and services. Following rules and guidelines for cataloging electronic resources, information is entered into MARC fields in OCLC's Connexion software. Some fields, for example, the 245 title statement, are transcribed from the piece. Other fields, like the 5xx (note and summary), can be taken from outside sources. Since the responsible parties were not listed on the AQComCat, this was included in a note field. This and the summary were compiled from information found on the Datafed wiki, youtube, and other sources. After the record is complete and reviewed, it is uploaded to WorldCat and then downloaded to the local catalog. OCLC later added linked data by appending Schema.org descriptive mark-up to the WorldCat record.

Usability

A small usability study (n=5) was conducted on both the original and revised AQ ComCat to compare the accessibility of the data type definitions and the speed at which they may be found. The brief usability was conducted with five individuals who had no prior knowledge of air quality data or their use. The number of clicks to find a data type definition and the length of time a tester took to access the definitions was recorded. The usability study also used retrospective probing[3] techniques to gather qualitative information from the users perspective of both the original and revised AQComCat.

As Table 1 below illustrates, usability results showed users required fewer clicks to access the definitions of the data types and also found them quicker as compared to the original AQComCat.

	Revised AQComCat	Original AQComCat
Average # of clicks to find 'grid' data type definition	1	4 (2 users gave up)
Average duration to find 'grid' data type definition	1:09	4:03
Average # of clicks to find 'NAAPS' data type definition	2	3
Average duration to find 'NAAPS' data type definition	1:42	2:05

Table 1: Usability results

[3] Retrospective probing is a usability technique that requires waiting until the session is complete and then asking questions about the participant's thoughts and actions.

Feedback after the usability from participants provided additional information about the revised and original AQComCat. All participants felt both sites could improve by adding the following features/information:

- Description of site and its purpose
- Site search
- Complete missing data/information

Data Type Registry Feedback

Creating new data types was a very straightforward process using the data type registry created by the RDA working group. The automated process of assigning a persistent identifier was helpful and further facilitated the machine readability of the data types.

When accessing the created data types via the API, the JSON did not parse as expected and our programmer received a .NET deserializing error shown in Table 2 below. Ultimately the programmer treated it as a string and was able to access the JSON DTR documents.

There was an error deserializing the object of type uFind.Provenance. DateTime content '2015-07-21T14:54:53.865Z' does not start with '\/Date(' and end with ')\/' as required for JSON.

Table 2: JSON parsing error

The individuals who participated in the usability also had a number of comments about the DTR interface. Comments fell into two categories; visual display and organization of information.

Visual Display of DTR

- Users felt the gray on gray display of data type records made reading the descriptions difficult; additionally, some variables of the used gray font, which also made user squint to read
- Users felt the DTR required lots of scrolling
- The DTR often took quite a while to load the records and users felt the experience could be improved with faster downloads

Organization of information on DTR

- Remove repeated ‘title’ on each record
- All users commented that they would have liked to see how terms were related in a table rather than the ‘experimental’ variable at the bottom of the record – and moved higher on the record display
- Users felt the identifier at the top of the page could be move lower to accommodate the relationships among datatypes to be displayed without scrolling

The team is mindful that the “real” usability testing will be in the months and years to come. The long tail of future users engaging AQComCat and the broader functions of the DataFed portal will make the ultimate determination on the usability of the DFT data model. The team believes that the timeframe of the adoption grants is a limitation to conducting full usability testing.

Adoption Feedback

Overall, the WUSTL team members found the process of adopting a common terminology and of data typing to be extremely beneficial. The foundational terms were helpful in assisting team members in establishing a common language to describe the different data points within the model or workflow. As many of our team members were working remotely, this helped in facilitating communication. The data typing was also helpful in making the implicit assumptions of the data more explicit. The RDA working groups were available to discuss their outcomes at length and provide feedback on adoption and implementation throughout the process.

The WUSTL team has a few recommendations for the working groups to improve the adoption process in the future. First, the team felt it would helpful to have best practices available on how to create a data type in the registry. Currently, the registry exists and there are examples populated, but they are incredibly complex. Additionally, any assistance in conceptually thinking how to break down derived types into primitives would be helpful. For example, the creation of a framework or a set of questions to help the adopter include the necessary components of what comprises a type would be beneficial.

Examples or use cases of how the information placed into the DTR could be made machine-readable would be useful. If the DTR working group could develop a common ontology, vocabulary, XMLS or RDFS it would help in standardizing the machine readability of these data objects and further facilitate linked data operations.

Process Feedback

The timeframe for the adoption in 8 months required a significant amount of communication and coordination among the grant team members, the working groups, and stakeholders. Coming to the RDA DTR and DFT groups as an implementation candidate and adopting their outcomes to an established community resource meant the resource had to be technically modified, which cannot be underestimated. Designing real-world solutions for extant data repositories such as AQComCat must include the socialization and politics of any implementation, which often takes additional time, extending well beyond the conclusion of the grant. To put it another way, the RDA solution must bring clear value to the community (not simply a different process) for it to achieve ultimate goals of changing behavior within the targeted community.

It may be useful to the larger goals of the RDA\US to convene adopters periodically to better understand (and guide) future working groups on the viability of outcomes. These teams may also be able to provide a core group for the implementation or regional “Tiger Team” concept that has been discussed within RDA.

Acknowledgements

The WUSTL team would like to thank the leaders of the DFT working group, specifically Larry Lannom, Raphael Ritz, Giridhar Manepalli and Gary Berg-Cross for their willingness to share the process and decisions leading up to the release of the DFT Data Model & DTR RDA outcome. The team would also like to thank everyone at RPI for their leadership on the “early adopter” grants, including Dr. Fran Berman and Dr. Kathy Fontaine for challenging each of us to build on the incredible work of the Research Data Alliance.

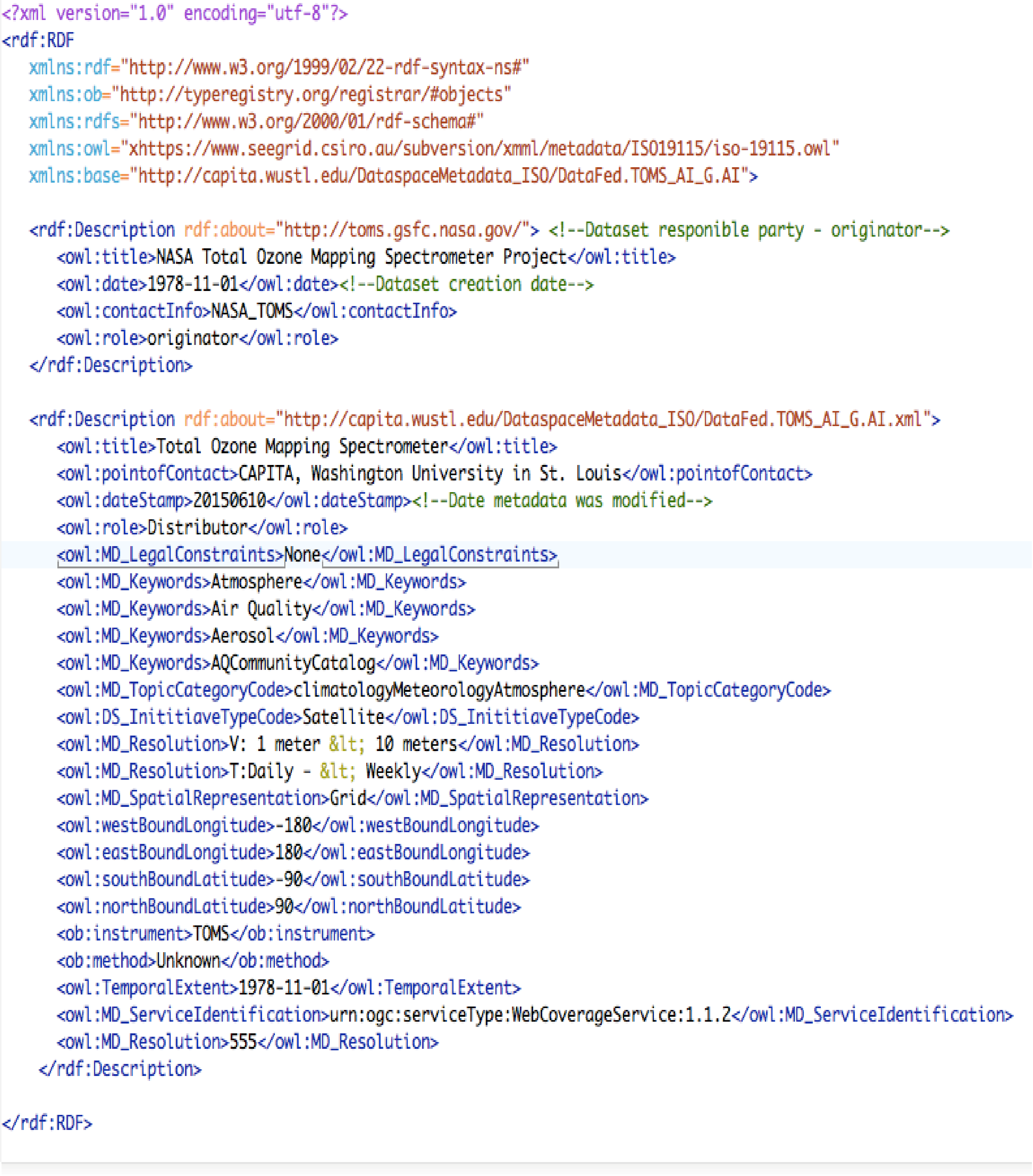
Appendix A: Facet/Value Pairs Registered into DTR

Facet:Name	Facet:Value	FacetValue	Description
Domain	Demographic	Domain:Demographic	Demographic spatio-temporal data
Domain	Emissions	Domain:Emissions	Pollutant Emissions data in space and time
Domain	Fire	Domain:Fire	Fire occuring in space and time
Domain	Gas	Domain:Gas	Gaseous concentration in air
Domain	Generic	Domain:Generic	Generic domain
Domain	GIS	Domain:GIS	Georefernceced spatial features
Domain	Chemistry	Domain:Chemistry	Model-derived concentrations
Domain	Land	Domain:Land	Data on land
Domain	Meteorology	Domain:Meteorology	Data on meteorology
Domain	GIS	Domain:GIS	Geospatial Stndard-based data
Domain	Test	Domain:Test	Experimental data domain
Domain	URL	Domain:URL	Uri type
TimeResolution	Day	TimeResolution:Day	Data available with resolution of a day
TimeResolution	Hour	TimeResolution:Hour	Data available with resolution of an hour
TimeResolution	Minute	TimeResolution:Minute	Data available with resolution of a minute
TimeResolution	Month	TimeResolution:Month	Data available with resolution of a month
TimeResolution	NA	TimeResolution:NA	Time resoultion not applicable
TimeResolution	Second	TimeResolution:Second	Data available with resolution of a second
TimeResolution	Year	TimeResolution:Year	Data available with resolution of a year
DataType	Grid	DataType:Grid	Regularly spaced spatial data in two or three dimensions, X-Y-Z
DataType	Image	DataType:Image	Geo-referenced image representing spatial data
DataType	Point	DataType:Point	Observations at specific geographic locations
DataType	Trajectory	DataType:Trajectory	A line made of sequential points. Usually represents transport path of moving objects.
Platform	Catalog	Platform:Catalog	Catalog
Platform	Count	Platform:Count	Count
Platform	Emissions	Platform:Emissions	Emissions
Platform	General	Platform:General	Miscellaneous content
Platform	GIS	Platform:GIS	Geographical Information Systems
Platform	Model	Platform:Model	Numerical Model
Platform	NA	Platform:NA	NA
Platform	Network	Platform:Network	Monitoring Network
Platform	Satellite	Platform:Satellite	Satellite
Platform	Unknown	Platform:Unknown	Unknown
Method	Catalog	Method:Catalog	Catalog data entry including unique identifier

Method	Cont	Method:Cont	Data collected continuously over the measurement period
Method	Continuous	Method:Continuous	Data collected continuously over the measurement period
Method	FilterSmp	Method:FilterSmp	Sample collected over specific period. Usually contains materials for further analyasis
Method	Human_Obs	Method:Human_Obs	Qualitative observation or judgement by human
Method	Model	Method:Model	Data derived from a physico-chemical, statistical or other model
Method	NA	Method:NA	Method is not relevant to the data
Method	Network	Method:Network	Time-synchronized set of observations at multiple locations
Method	None	Method:None	Method is not applied
Method	Point	Method:Point	Sample collected irregularly in space and/or time
Method	RemSensAir	Method:RemSensAir	Remote sensor based on airborne platform
Method	RemSensSat	Method:RemSensSat	Observation based on Earth orbiting satellite data
Method	RemSensSurf	Method:RemSensSurf	Remote sensor based on surfece
Method	Unknown	Method:Unknown	Method is unknown
Instrument	Aerosol Sampler	Instrument:Aerosol Sampler	The instrument is an aerosol sampler
Instrument	Cimel	Instrument:Cimel	CIMEL Electronique 318A spectral radiometer that measures Sun and sky radiances at a number of fixed wavelengths within the visible and near-infrared spectrum. The Cimel Sun photometer is used in the global-scale Aeronet network that measures aerosol properties
Instrument	CMAQ	Instrument:CMAQ	CMAQ consists of a suite of programs for conducting air quality model simulations. It combines current knowledge in atmospheric science and air quality modeling with multi-processor computing techniques in an open-source framework to deliver fast, technically sound estimates of ozone, particulates, toxics, and acid deposition.

Instrument	FRM	Instrument:FRM	Federal Reference Methods are EPA-developed methods for accurately and reliably measuring these six criteria pollutants in ambient air. These methods are used by states and other monitoring organizations to assess implementation actions needed to attain National Ambient Air Quality Standards. FRMs ensure that air quality data collected at different sites are accurate and can be used for purposes of inter-comparison.
Instrument	Gas Analyser	Instrument:Gas Analyser	Measures the concentration of gaseous species in air
Instrument	IMPROVE Sampler	Instrument:IMPROVE Sampler	The IMPROVE sampler was designed to collect on filter airborne particles. It consists of four sampling modules. The filters are analyzed for chemical and optical components that affect visibility.
Instrument	NA	Instrument:NA	Not Applicable
Instrument	NAAPS	Instrument:NAAPS	The Naval Research Laboratory (NRL) in Monterey, CA, has developed a near-operational system for predicting the distribution of tropospheric aerosols. The model is a modified form of that developed by Christensen (1997). The NRL version uses global meteorological fields from the Navy Operational Global Atmospheric Prediction System (NOGAPS) (Hogan and Rosmond, 1991; Hogan and Brody 1993) analyses and forecasts on a 1 X 1 degree grid, at 6-hour intervals and 24 vertical levels reaching 100 mb
Instrument	Unknown	Instrument:Unknown	The instrument used for the measurement is unknown
Vertical	Column	Vertical:Column	Column concentration over a vertical slab
Vertical	Layer	Vertical:Layer	Data layer in a model
Vertical	NA	Vertical:NA	Not Applicable
Vertical	None	Vertical:None	None
Vertical	Profile	Vertical:Profile	Entity profine in vertical dimansion
Vertical	Surface	Vertical:Surface	Entity value at the Earth surface

Appendix B: RDF/XML record



```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:ob="http://typeregistry.org/registrar/#objects"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="xhttps://www.seegrid.csiro.au/subversion/xmml/metadata/ISO19115/iso-19115.owl"
   xmlns:base="http://capita.wustl.edu/DataspaceMetadata_ISO/DataFed.TOMS_AI_G.AI">

   <rdf:Description rdf:about="http://toms.gsfc.nasa.gov/"> <!--Dataset responible party - originator-->
      <owl:title>NASA Total Ozone Mapping Spectrometer Project</owl:title>
      <owl:date>1978-11-01</owl:date><!--Dataset creation date-->
      <owl:contactInfo>NASA_TOMS</owl:contactInfo>
      <owl:role>originator</owl:role>
   </rdf:Description>

   <rdf:Description rdf:about="http://capita.wustl.edu/DataspaceMetadata_ISO/DataFed.TOMS_AI_G.AI.xml">
      <owl:title>Total Ozone Mapping Spectrometer</owl:title>
      <owl:pointofContact>CAPITA, Washington University in St. Louis</owl:pointofContact>
      <owl:dateStamp>20150610</owl:dateStamp><!--Date metadata was modified-->
      <owl:role>Distributor</owl:role>
      <owl:MD_LegalConstraints>None</owl:MD_LegalConstraints>
      <owl:MD_Keywords>Atmosphere</owl:MD_Keywords>
      <owl:MD_Keywords>Air Quality</owl:MD_Keywords>
      <owl:MD_Keywords>Aerosol</owl:MD_Keywords>
      <owl:MD_Keywords>AQCommunityCatalog</owl:MD_Keywords>
      <owl:MD_TopicCategoryCode>climatologyMeteorologyAtmosphere</owl:MD_TopicCategoryCode>
      <owl:DS_InititiaveTypeCode>Satellite</owl:DS_InititiaveTypeCode>
      <owl:MD_Resolution>V: 1 meter &lt; 10 meters</owl:MD_Resolution>
      <owl:MD_Resolution>T:Daily - &lt; Weekly</owl:MD_Resolution>
      <owl:MD_SpatialRepresentation>Grid</owl:MD_SpatialRepresentation>
      <owl:westBoundLongitude>-180</owl:westBoundLongitude>
      <owl:eastBoundLongitude>180</owl:eastBoundLongitude>
      <owl:southBoundLatitude>-90</owl:southBoundLatitude>
      <owl:northBoundLatitude>90</owl:northBoundLatitude>
      <ob:instrument>TOMS</ob:instrument>
      <ob:method>Unknown</ob:method>
      <owl:TemporalExtent>1978-11-01</owl:TemporalExtent>
      <owl:MD_ServiceIdentification>urn:ogc:serviceType:WebCoverageService:1.1.2</owl:MD_ServiceIdentification>
      <owl:MD_Resolution>555</owl:MD_Resolution>
    </rdf:Description>

</rdf:RDF>
```