Linked Data breakout session

Attendees
Jonathan Yu (chair), Mark Hedley (co-chair, remote dial in), Denis Nadeau, Aaron Sweeney, Jim Biard, Dave Blodgett, Larry Oolman, Stefano Nativi.

1. Introductions

2. Present background and context (Jonathan and Mark)
slides
https://docs.google.com/presentation/d/1S8_WOpsIL7Sw27sa4ylGoDoMcAhPUppsPM-xyQdcj0s/edit#slide=id.g1334a8bd12_4_1123

Problem and challenges
- Semantics in the data and across data formats
- Range of conventions and practices e.g. ACDD, CF, oceans, cross-org, project and team
- Keeping up to date, validation, discovery and use, referencing external content

Linked Data (LD) background
- Foundational principles
- Traction and growth of LD datasets since 2007
- Lots of content available and being published - science domains, features, etc.

Not alone
- Some precedence in JSON and CSV for Linked Data profiles, approaches and tooling
- Lessons to learn from XML

We have building blocks available to test LD approaches for netCDF
Benefits
- Enhance discoverability
- Conformance checking
- Potentially transforming between formats and formalisms (groups vs flat)
- Pull in external content and providing additional info
- Ultimately have better discoverability, easier to use, and wider impact

Current approaches and thought exercises
- People are already injecting web links (links to external content) into netCDF but fragmented approach - would be good to just have one approach
- netCDF-LD strawman approach in a paper in 2014 - basically puts a boilerplate and attributes as decorators alongside CF
    - Ability to harvest semantics from the netCDF metadata, aggregate and enable discovery
- Binary Array LD - BALD proposal

    - able to represent Containers and Arrays as subjects we can reference
    - Method to add prefixes to attributes so we can build web links from netCDF metadata and render machine readable Linked Data content
    - Provide ability to do compliance checking of any convention
        - E.g. represent conventions (CF/ACDD/etc) as machine readable form, do compliance checking of these convention

3. Discussion and questions

Is there value in exploring Linked Data approach for netCDF(-CF)?
There was consensus from participants on the potential for Linked Data

- SN: Value is that we don't have to embed full semantics into netCDF file
    - Note: This is a change - as referencing external content
    - Personally, feel that proposal is reasonable
    - CF - control semantics?
- DB: Current problem is that identifiers and vocabularies/standard names are opaque - no way to find more info directly, so LD approach could address this problem

JB: Is this CF? Or is this CF compatible?

- Scope of whether relevant for CF only or wider (netCDF)?
- DB - CF should say how to embed LD web links, e.g. how standard names, units referenced
- JB: precedence in ACDD, which is consciously not in CF
    - AS: CF is 'use' metadata, ACDD is 'discovery' metadata. Linked data approach is both use and discovery.

DB: What is the minimal step to get us to netCDF-LD?

- SN: As the principle is in being able to include external references - should be useful to enhance discovery and usability
- (need to demonstrate this…)
- MH: suggests small steps - suggest netCDF + CF community explore how we capture URIs - current proposal is to use namespace prefixes and using a double '_' to represent prefixes in attribute names, e.g. cf__standard_name <equivalent> standard_name
- LO: Worry about having to handle additional stuff in software for parsing namespaces

Discussion about the ability to blend conventions and accommodate multiple conventions

- Able to validate them

JB: Side note - netCDF 3.6.3 - anything except a slash in a name

DB: Use case - being able to reference other kinds of metadata more broader than variable level - features (discrete geometries, stations, palatforms, instruments)

- Ability to link out to these features

Draft use cases:

- Discovery
- Use
    - Machine readable content
    - Pull in external content - like labels, additional content
    - Understand context of datasets with general scientific concepts, sensors, features
- Encoding
    - help data providers to reference external sources in netCDF
    - e.g. reference features (geoms, stations, platform, instrument, sensor)
- Compliance checking
    - help data providers check conventions bound - e.g. practice of 1 or more conventions

Challenges:

- Wary of introducing XML-ism into netCDF
    - Perhaps have default namespaces for each convention
    - People like netCDF because there's no namespace
    - Namespace prefixes not as elegant
- Governance of prefix namespace
    - netCDF LD - falls under unidata?
    - governance of other namespaces with community e.g. CF
- Persistence of URIs
    - injecting fragility
    - Fragility already exists - references to convention documentation Doi?

Principles

- Doesn't break classic CF - Backwards compatible
- Prefer elegance of classic CF
- Forward looking approach

What would we need to make this work?

- principles (see prev slide)
- project use cases
- endorsement - CF/ACDD/CMIP (conventions level) or netCDF (at an API level)

What would an activity look like?

- Github
- test BALD software on github
- Provide use cases
- Monthly telecons

How do we organise it? Next steps and timeframes

- Agreed to pilot this activity in the next 6 months, propose monthly telecon in this period
- github

Summary presented in plenary.

Plenary discussion

- There is potential - probably still early stages
- Standards process is that experiment with something - come back to community to get approved
    - Suggestion for developing on a branch on netCDF APIs on Github and coming back to propose in CF
    - Suggestion for engineering report to netCDF SWG for endorsement

Actions:

- Need feedback CF community - is it acceptable to reference external resources?
- Demonstrate value of leveraging external sources e.g. vocabs, compliance checking
    - E.g. CRS - spatialreference.org links embedded in netCDF?
- Carry out some tests and experiments