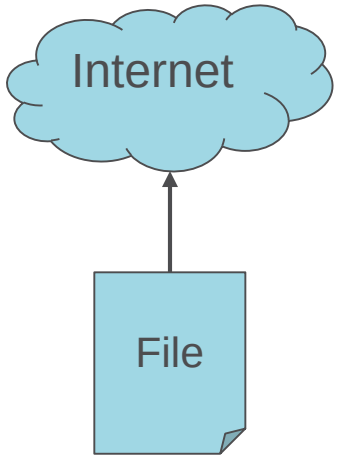
Australian Government
Geoscience Australia

netCDF-LD SKOS:
demonstrating Linked Data vocabulary use within netCDF-compliant files

Nicholas Car
Data Architect
Geoscience Australia
nicholas.car@ga.gov.au

Prepared for ISESS2017 conference (http://www.isess2017.org/)

Presented on 2017-05-11



Outline

- What we are doing and why
- Intro to main concepts
- ncskos
- ncskosdump tool
- Deployment scenarios
- Future (current) work
- Conclusions

Outline

- **What we are doing and why**
- Intro to main concepts
- ncskos
- ncskosdump tool
- Deployment scenario
- Future (current) work
- Conclusions

What we are doing and why

- **What**
  - Linking data files to reference metadata
- **Why**
  - Better metadata means better data use

- Not all metadata is equal
  - Curated vocabs better than un-curated range values
  - Curated vocabs better than static code lists
- New tools mean new metadata possibilities
  - Faster networks, better UIs, many more

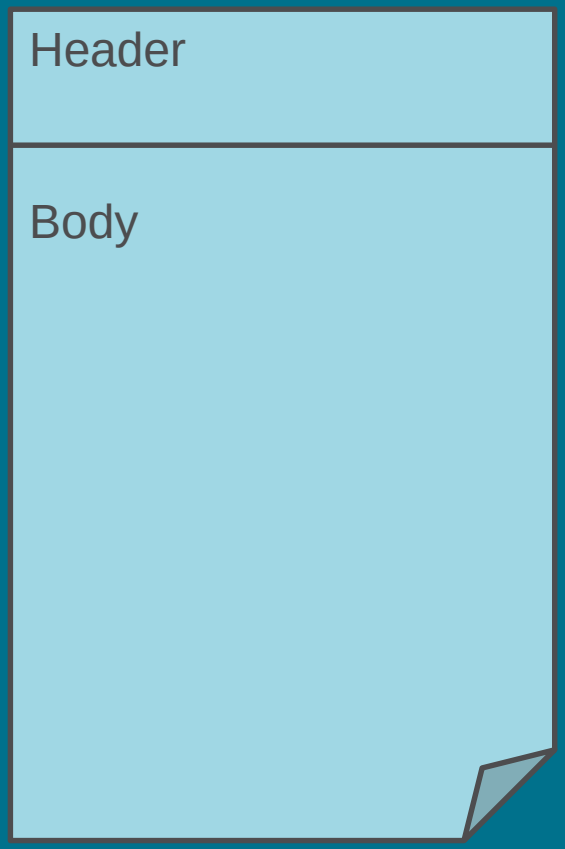
Outline

- What we are doing and why
- **Intro to main concepts**
  - netCDF
  - netCDF metadata communities
  - Linked Data
  - SKOS
- ncskos
- ncskosdump tool
- Deployment scenarios
- Future (current) work
- Conclusions

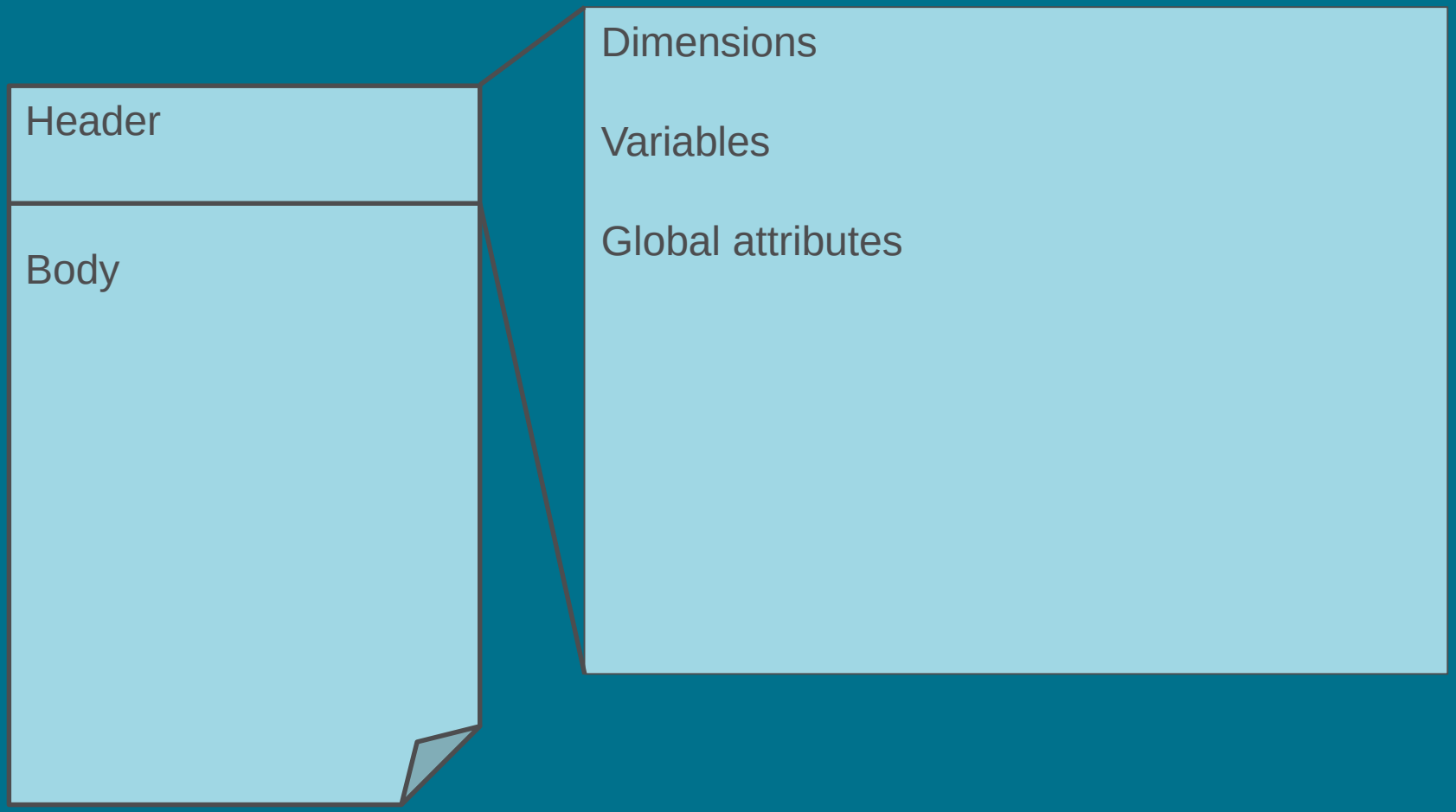
Intro to main concepts

- **netCDF – Network Common Data Form**
  - Containers that include both data – usually array-oriented scientific data – and metadata
  - Intended to be ‘self-describing’
  - Very widely used
  - Example: https://www.unidata.ucar.edu/software/netcdf/examples/sresa1b_ncar_ccsm3-example.cdl

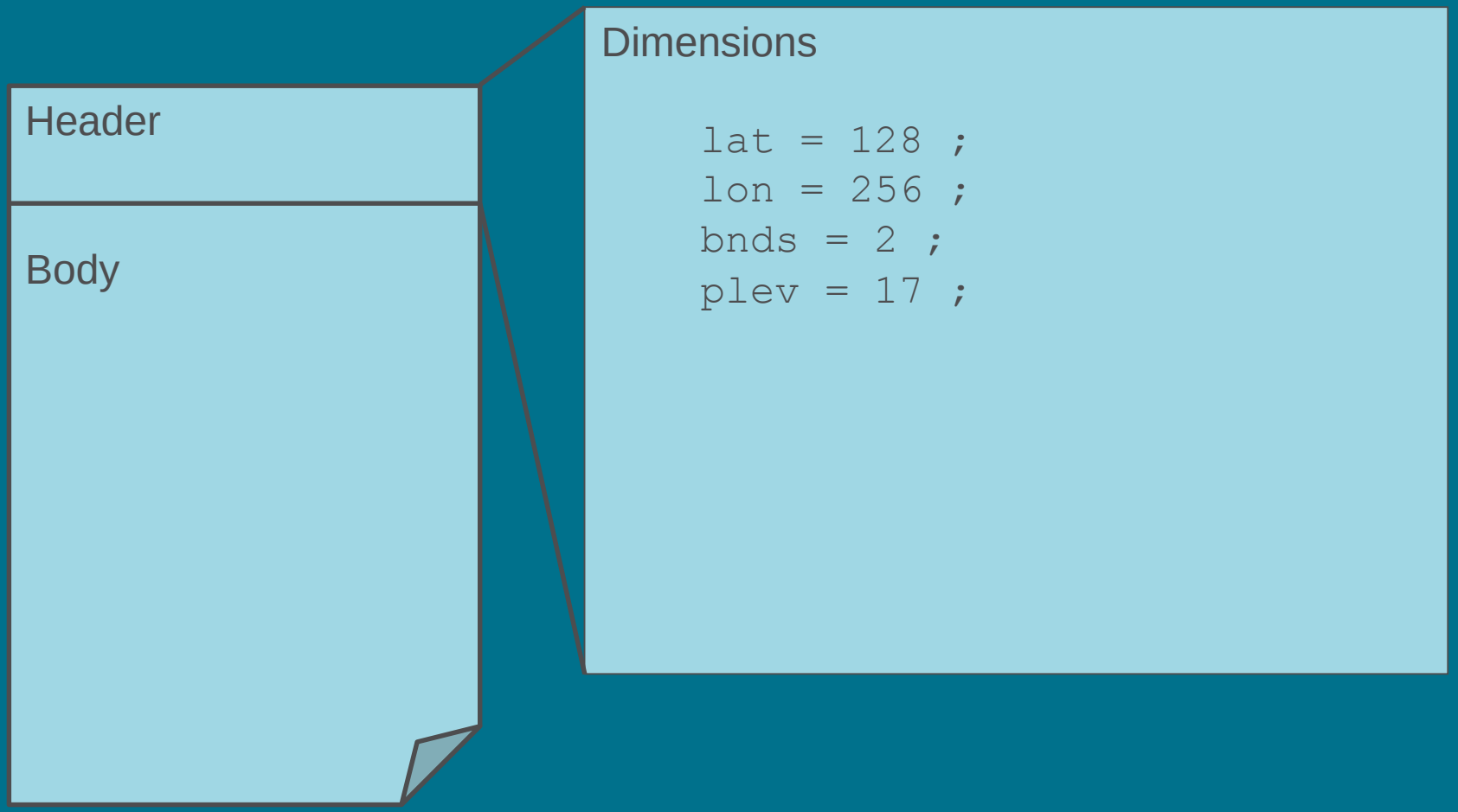
Intro to main concepts: netCDF



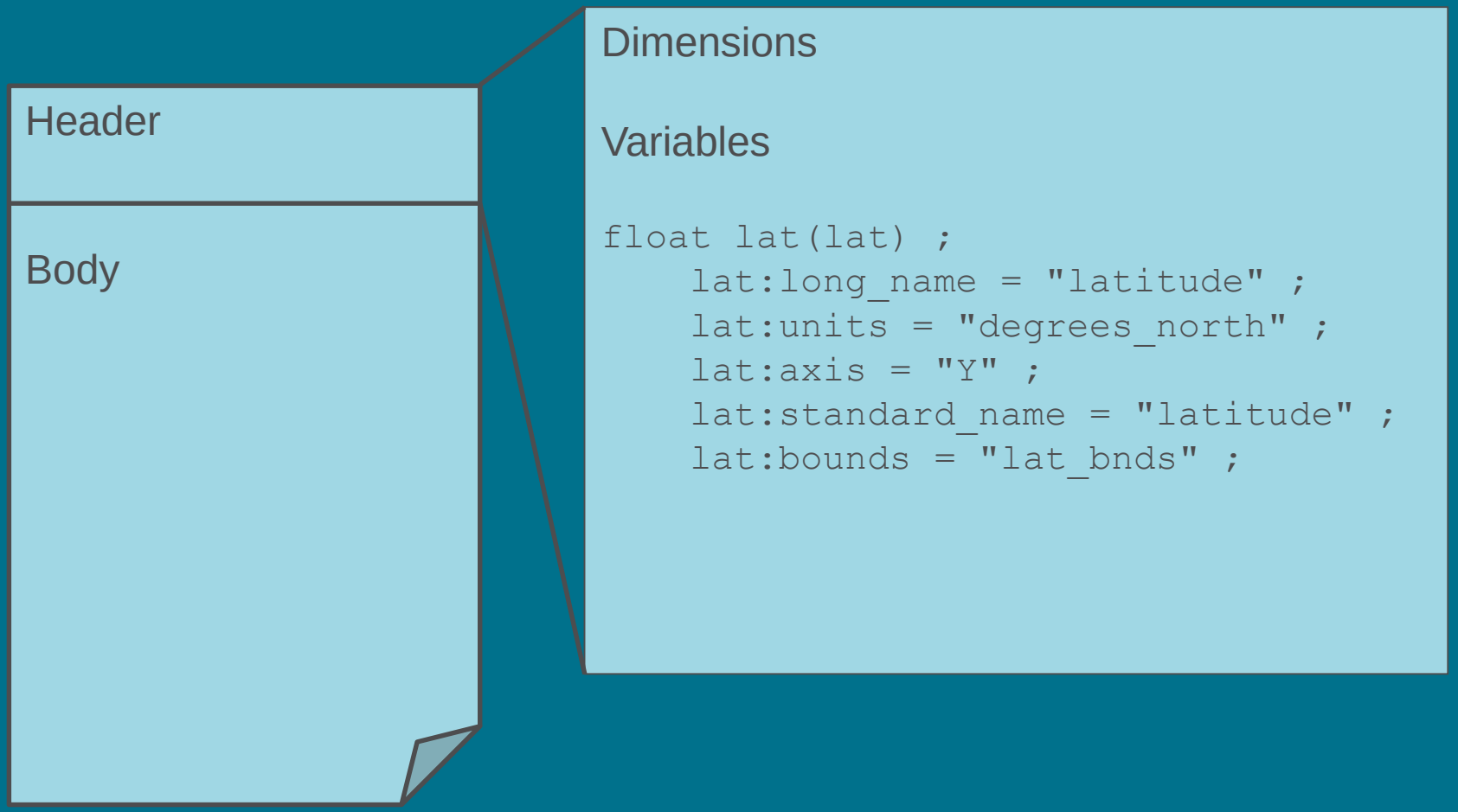
Intro to main concepts: netCDF



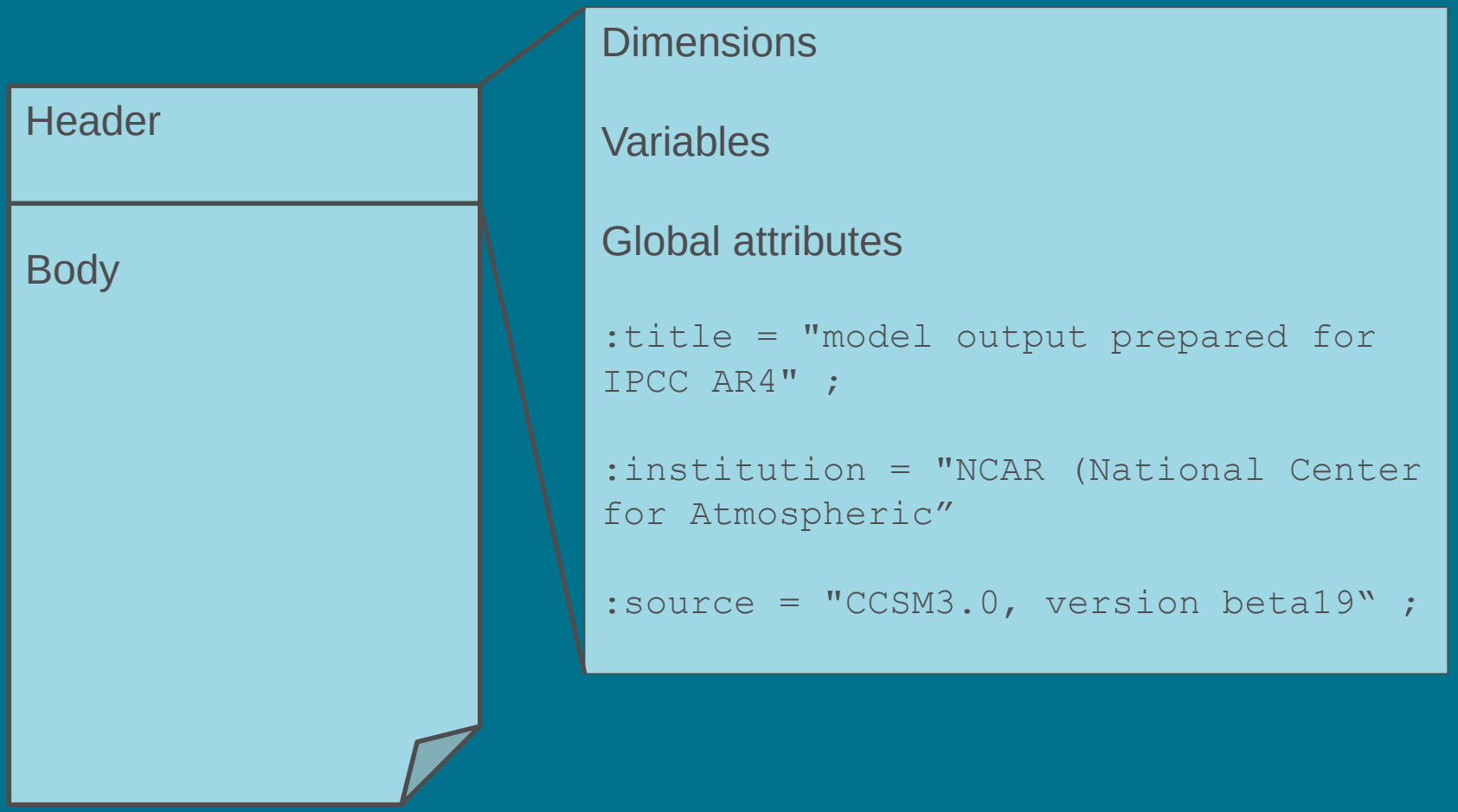
Intro to main concepts: netCDF



Intro to main concepts: netCDF



Intro to main concepts: netCDF

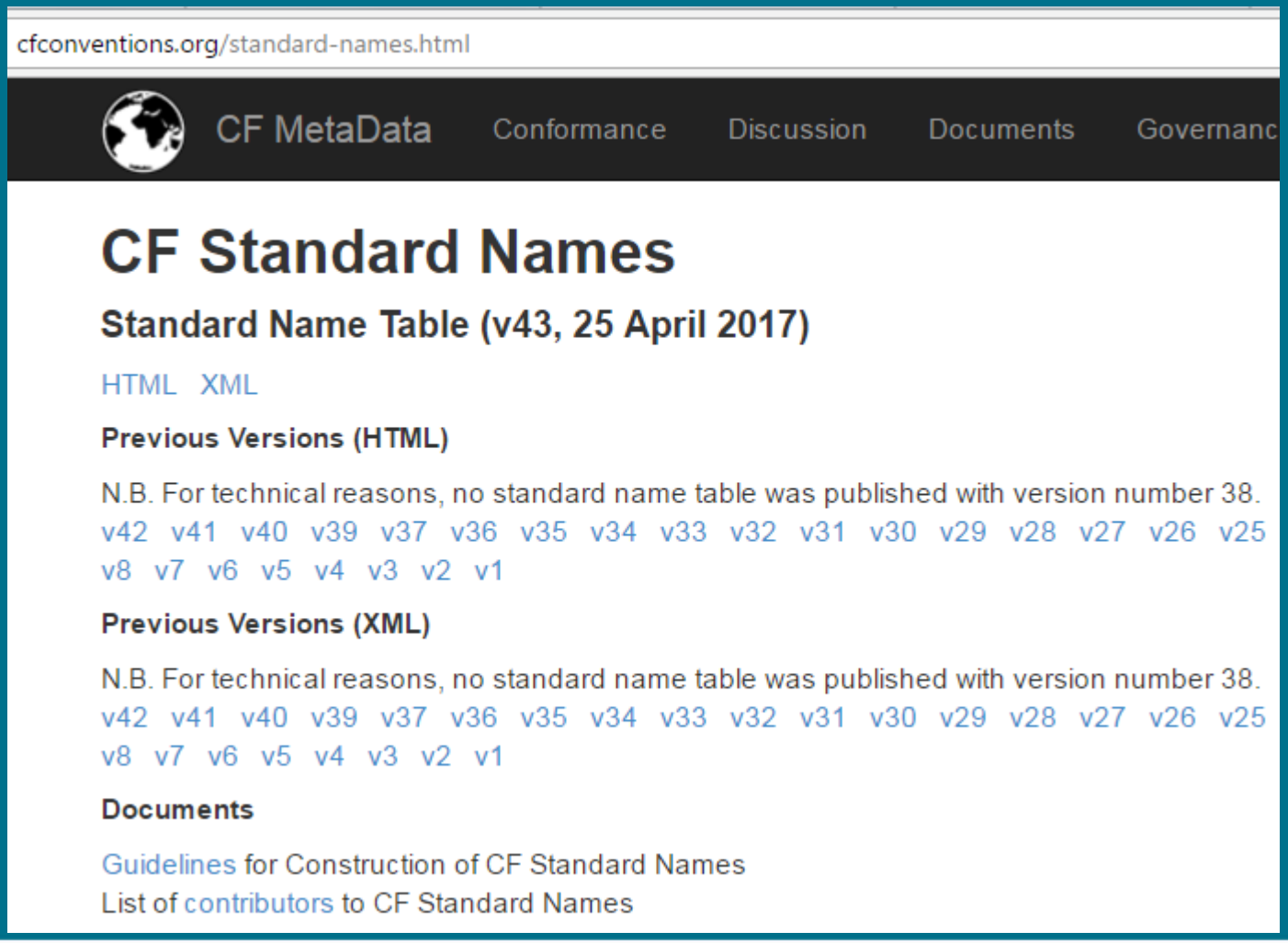


Intro to main concepts

- **netCDF metadata communities**
  - Climate and Forecast (CF) conventions… for the description of Earth sciences data, intended to promote the processing and sharing of data files
  - provide a definitive description of what the data values found in each netCDF variable represent
  - Use code lists published at http://cfconventions.org

Intro to main concepts

- **netCDF metadata communities**



Intro to ma

- netCDF m

cfconventions.org/Data/cf-standard-names/42/build/cf-standard-name-table.html

units which are physically equivalent (not necessarily identical) to the canonical units, possibly modifie
Section 1.3 of the CF conventions states: "The values of the units attributes are character strings that ar
"Units")." For example, a variable with the standard name of "air_temperature" may have a units attribu
refer to the Udunits documentation. Refer to the CF conventions for full details of the units attribute.

cfconventions.org/sta

CF

CF St

Standard

HTML XML

Previous Ve

N.B. For tech
v42 v41 v
v8 v7 v6

Previous Ve

N.B. For tech
v42 v41 v
v8 v7 v6

Documents

Guidelines f
List of contri

Search

Search Standard Names | Show All Standard Names

AND OR (separate search terms with spaces)

Also search help text

View by Category

Atmospheric Chemistry	Atmosphere Dynamics	Carbon Cycle	Cloud	Hydrology
Ocean Dynamics	Radiation	Sea Ice	Surface	

Standard Name
acoustic_signal_roundtrip_travel_time_in_sea_water
aerodynamic_particle_diameter
aerodynamic_resistance
age_of_sea_ice
age_of_stratospheric_air
age_of_surface_snow

Intro to ma

- netCDF m

cfconventions.org/Data/cf-standard-names/42/build/cf-standard-name-table.html

units which are physically equivalent (not necessarily identical) to the canonical units, possibly modifie
Section 1.3 of the CF conventions states: "The values of the units attributes are character strings that ar
"Units")." For example, a variable with the standard name of "air_temperature" may have a units attribu
refer to the Udunits documentation. Refer to the CF conventions for full details of the units attribute.

cfconventions.org/sta

CF

CF St

Search

Search Standard Names | Show All Standard Names

AND OR (separate search terms with spaces)

air_potential_temperature

Potential temperature is the temperature a parcel of air or sea water would have if moved adiabatically to sea level pressure.

	K	theta	13

Previous Ve

N.B. For tech
v42 v41 v
v8 v7 v6

Documents

Guidelines f
List of contri

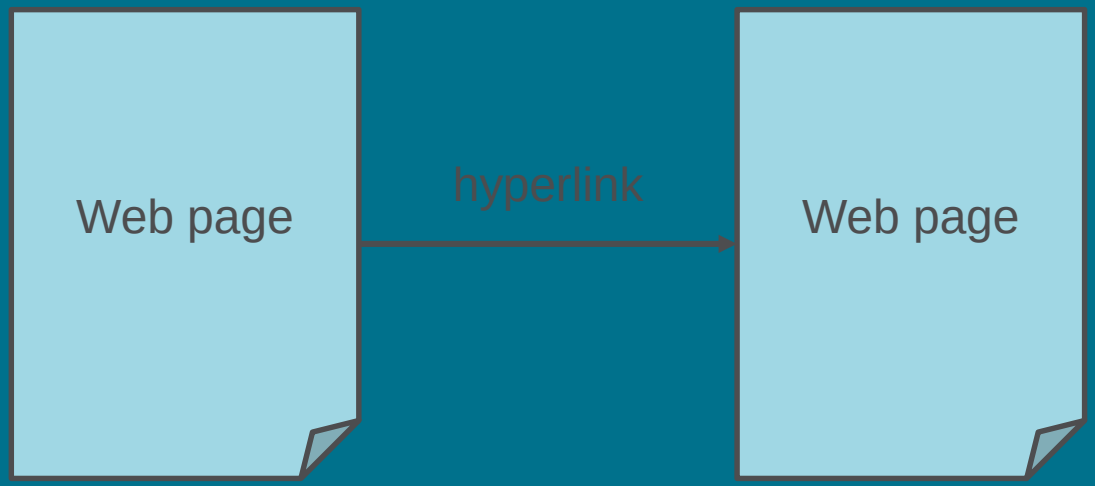
Standard Name
acoustic_signal_roundtrip_travel_time_in_sea_water
aerodynamic_particle_diameter
aerodynamic_resistance
age_of_sea_ice
age_of_stratospheric_air
age_of_surface_snow

GEOSCIENCE AUSTRALI

Intro to main concepts

- **Linked Data**
  - "Linked Data provides a publishing paradigm in which not only documents [web pages] but also data can be first class citizens of the Web"

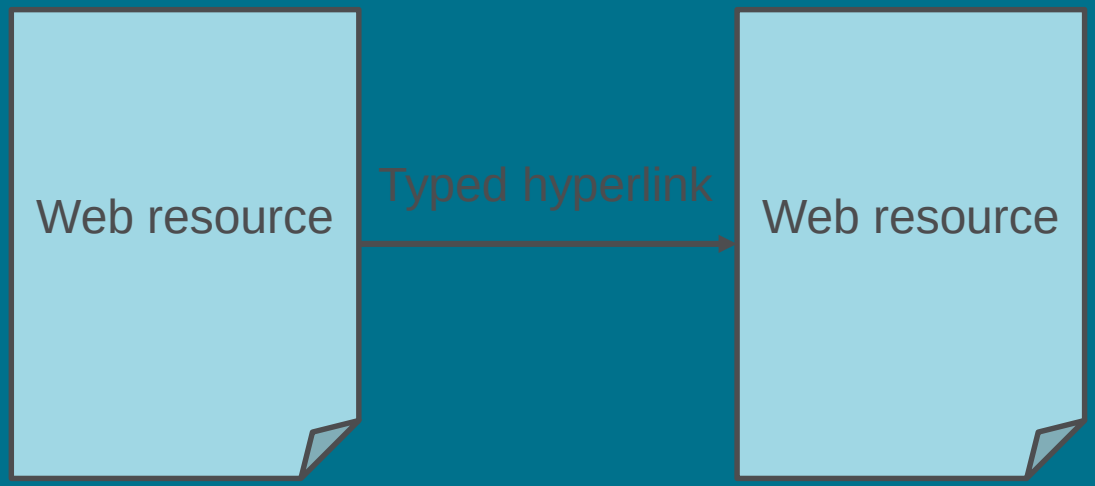
    – Heath & Bizer, 2011
  - Normal web pages:



Intro to main concepts

- **Linked Data**
  - "Linked Data provides a publishing paradigm in which not only documents [web pages] but also data can be first class citizens of the Web"

    – Heath & Bizer, 2011
  - Linked Data:



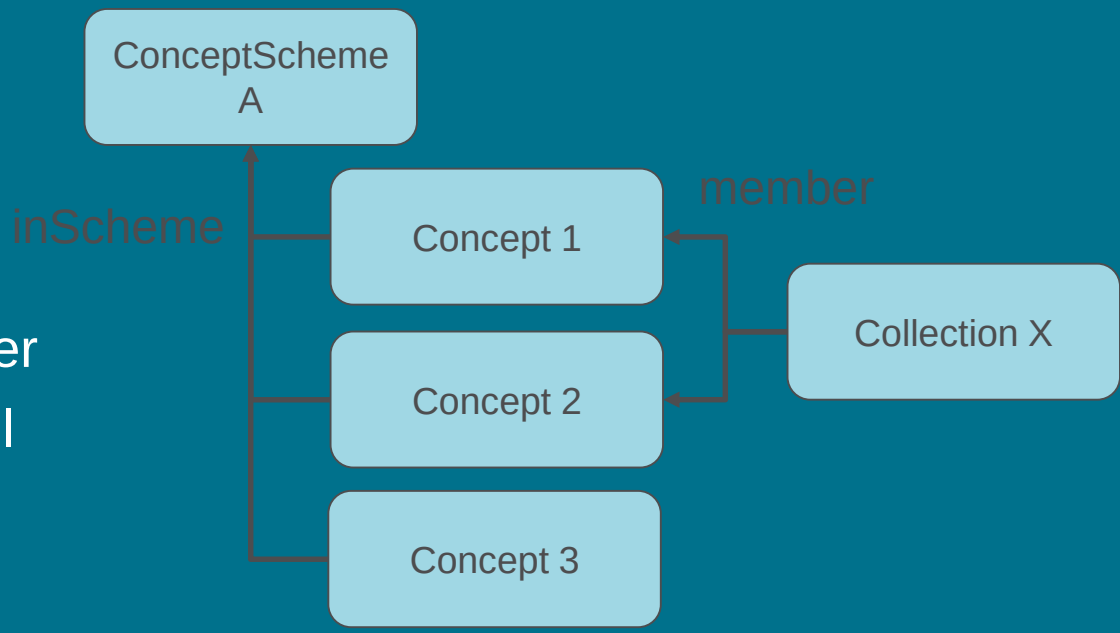
Intro to main concepts

- **Linked Data**
  - "Linked Data provides a publishing paradigm in which not only documents [web pages] but also data can be first class citizens of the Web"

    – Heath & Bizer, 2011
  - Linked Data:

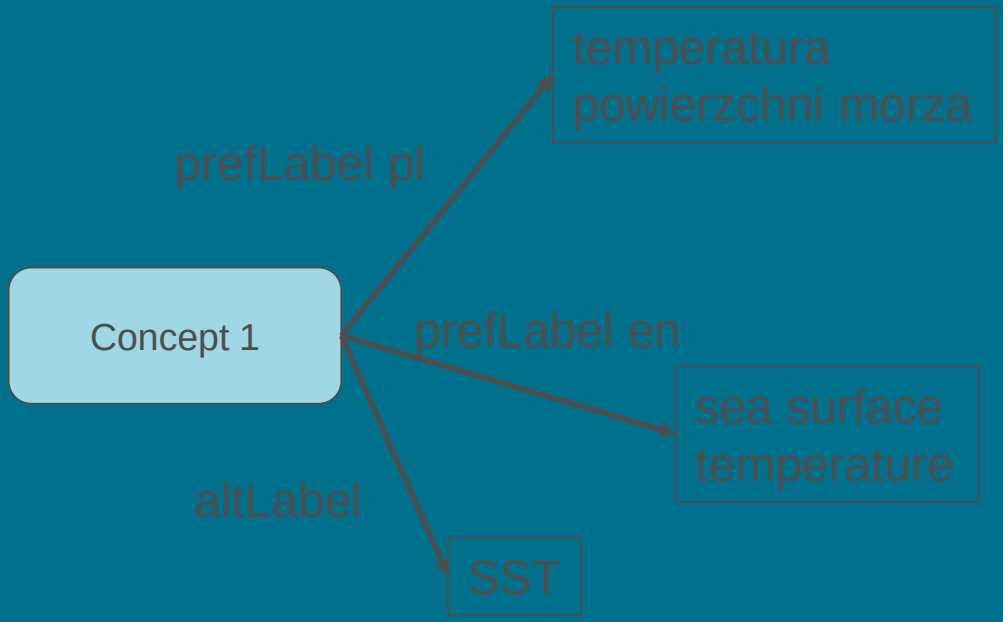
Intro to main concepts

- **SKOS – Simple Knowledge Organization System**
  - A vocabulary data model
  - Main classes:
    - Concept
    - ConceptScheme
    - Collection
  - Main relations:
    - hasTopConcept
    - inScheme, member
    - prefLabel, altLabel



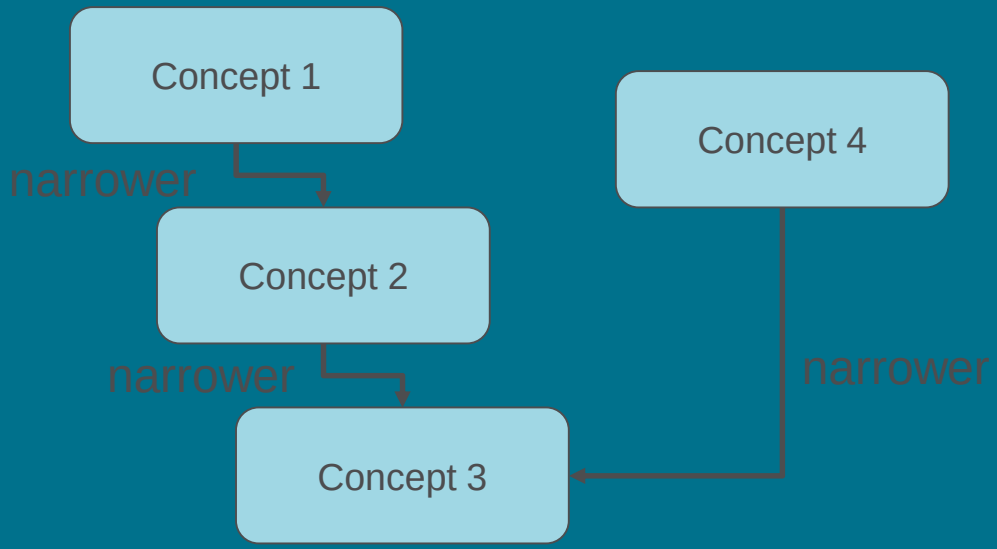
Intro to main concepts

- **SKOS – Simple Knowledge Organization System**
  - A vocabulary data model
  - Main classes:
    - Concept
    - ConceptScheme
    - Collection
  - Main relations:
    - hasTopConcept
    - inScheme, member
    - prefLabel, altLabel



Intro to main concepts

- **SKOS – Simple Knowledge Organization System**
  - A vocabulary data model
  - Main classes:
    - Concept
    - ConceptScheme
    - Collection
  - Main relations:
    - hasTopConcept
    - inScheme, member
    - prefLabel, altLabel
    - broader, narrower



Outline

- What we are doing and why
- Intro to main concepts
- **ncskos**
- ncskosdump tool
- Deployment scenarios
- Future (current) work
- Conclusions

ncskos

- We proposed a semi-sophisticated LD encoding for netCDF files in ISESS 2015
  - “Towards linked data conventions for delivery of environmental data using netCDF”
  - Used JSON-LD style “context” variables to specify namespaces
  - Proposed but did not define a netCDF structural ontology

```
var:ref= "http://vocab.nerc.ac.uk/collection/P07/current/CFSN0600/";
var:unit = "Meters";
var:unit_ref = " http://qudt.org/vocab/unit#Meter";
```

ncskos

- Now, a simpler method of including only SKOS definition links in netCDF files
- Aiming for
  - Scale deployment to make case for utility
  - Tool generation
  - Establishment of useful reference datasets

```
float tos(time, lat, lon) ;
    tos:standard_name = "sea_surface_temperature" ;
    tos:long_name = "Sea Surface Temperature" ;
    tos:units = "K" ;
    tos:original_units = "degC" ;
    …
```

ncskos

- Now, a simpler method of including only SKOS definition links in netCDF files
- Aiming for
  - Scale deployment to make case for utility
  - Tool generation
  - Establishment of useful reference datasets

```
float tos(time, lat, lon) ;
    tos:skos__concept_uri =
"http://pid.geoscience.gov.au/def/voc/netCDF-LD-eg-
ToS/sea_surface_temperature";
    tos:standard_name = "sea_surface_temperature" ;
    tos:long_name = "Sea Surface Temperature" ;
    tos:units = "K" ;
    tos:original_units = "degC" ;
    …
```

ncskos

- Now, a simpler method of including only SKOS definition links in netCDF files
- Aiming for
  - Scale deployment to make case for utility
  - Tool generation
  - Establishment of useful reference datasets

```
float tos(time, lat, lon) ;
    tos:skos__concept_uri =
"http://pid.geoscience.gov.au/def/voc/netCDF-LD-eg-
ToS/sea_surface_temperature";
```

ncskos

Concept of “sea surface temperature”, RESTful link:

http://pid.geoscience.gov.au/def/voc/netCDF-LD-eg-ToS/sea_surface_temperature

Whole vocab:

http://pid.geoscience.gov.au/def/voc/netCDF-LD-eg-ToS

Also in RDF for machines

Outline

- What we are doing and why
- Intro to main concepts
- ncskos
- **ncskosdump tool**
- Deployment scenarios
- Future (current) work
- Conclusions

ncskosdump tool

- Many well used netCDF tools:
  - ncdump
  - ncgen
  - ncview
  - Operators Toolkit

ncskosdump tool

- Many well used netCDF tools:
  - ncdump
  - ncgen
  - ncview
  - Operators Toolkit

```
# ncdump -h some_file.nc

sst:skos__concept_uri =
"http://pid.geoscience.gov.au/def/voc/netCDF-LD-
eg-ToS/sea_surface_temperature" ;
```

ncskosdump tool

- Many well used netCDF tools:
  - ncdump
  - ncgen
  - ncview
  - Operators Toolkit
- ncskosdump
  - Wraps ncdump, adds commands for SKOS info retrieval

```
# ncskosdump -h some_file.nc --lang=pl altlabels=True

sst:skos__prefLabel_pl = "temperatura powierzchni morza" ;
sst:skos__altLabels = "SST" ;
```

ncskosdump tool

- Many well used netCDF tools:
  - ncdump
  - ncgen
  - ncview
  - Operators Toolkit
- ncskosdump
  - Wraps ncdump, adds commands for SKOS info retrieval

```
# ncskosdump -h sst.nc --narrower=True

sst:skos__narrower = "http:…01, http:…02, http:…03"
```

ncskosdump tool

- Many well used netCDF tools:
  - ncdump
  - ncgen
  - ncview
  - Operators Toolkit
- ncskosdump
  - Wraps ncdump, adds commands for SKOS info retrieval
  - Session (memory) and location (file) metadata retrieval caching enabled

ncskosdump tool

- Commands:

--verbose to enable verbose output

--lang=<lang_code> where <*lang_code*> is a two-character ISO 639-1:200 code for the language in which the results are sought

--narrower to recursively create complete tree of narrower concepts, not just ones resolved directly from URIs

--altLabels=<altLabel_list> where <*altLabel_list*> is a comma-separated list of altLabels to match in order to list their associated datasets

--retries=<max_retries> where <*max_retries*> is the maximum number of retries to attempt for unresolved URIs. Default retries = 0

--delay=<retry_delay_seconds> where <*retry_delay_seconds*> is the number of seconds to wait before each retry. Default delay = 2s

--refresh to discard current file cache and repopulate the cache from scratch

Outline

- What we are doing and why
- Intro to main concepts
- ncskos
- ncskosdump tool
- **Deployment scenarios**
- Future (current) work
- Conclusions

Deployment scenarios

- Example in the paper & GitHub
  - Series of empty netCDF files
  - Test SKOS vocab: http://pid.geoscience.gov.au/def/voc/netCDF-LD-eg-ToS
  - Demo program to sort files, based on hierarchical relations

- Real-world scenario
  - 10,000+ netCDF files at the Aust. Nat. Comp. Center
  - Geophysics data, multiple variables per file
  - Variables being described in a Geophysics SKOS vocab
  - Automated selection using ncskosdump

Outline

- What we are doing and why
- Intro to main concepts
- ncskos
- ncskosdump tool
- Deployment scenario
- **Future (current) work**
  - PROV
  - netCDF-LD

Future (current) work

- **PROV**
- We have extended the use of a variable name prefix for the PROV ontology
  - file:prov__{var_name}
  - file:prov__wasGeneratedBy

Future (current) work

- **PROV**
- We have extended the use of a variable name prefix for the PROV ontology
  - file:prov__{var_name}
  - file:prov__wasGeneratedBy

:prov__wasGeneratedBy = "http://pid.geoscience.gov.au/survey/801"

pid.geoscience.gov.au/survey/801

Australian Government
Geoscience Australia

Applying geoscience to Australia's most important challenges

GA Home | Surveys API Home | Surveys Register

Survey

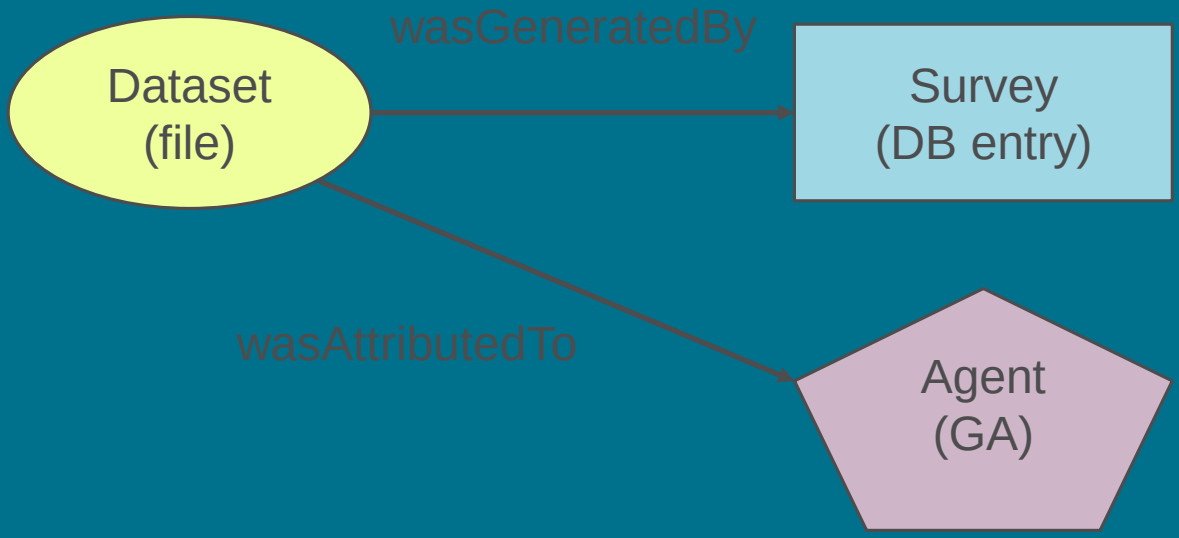
GAPD view

These properties are those listed in the GAPD Model of a Survey.

Property	Value
ID	801
Name	Basin/Quondong, Vale, SA Company, 1988 (88SA14) (74pt)
State	SA
Agents: **Contractor** **Operator** **Processor**	 Kevron Geophysics Pty Ltd MESA & Dominion Mining Aerodata Holdings Pty Ltd
survey_type	Semi-detailed
data_types	MAG
Vessel: **Type** **Name**	 Plane None
release_date	1989-01-01 00:00:00
onshore_offshore	Onshore

Future (current) work

- **PROV**
- We have extended the use of a variable name prefix for the PROV ontology
  - file:prov__{var_name}
  - file:prov__wasGeneratedBy



Future (current) work

- **netCDF-LD**
- We are working on a full netCDF LD implementation
  - WG with netCDF developers and users
- Making an encoding format for any LD data in netCDF files
  - It would be a new netCDF specification
  - Interest looks high (AGU, EGU, other confs)
- Using Spatial Data on the Web WG conventions
- Re-using JSON-LD encoding where possible

Outline

- What we are doing and why
- Intro to main concepts
- ncskos
- ncskosdump tool
- Deployment scenario
- Future (current) work
- **Conclusions**

Conclusions

- Just using SKOS is:
  - Easy
  - Useful – standardised vocab, hierarchy
  - Better than CF – open world, curation still possible
- Using limited PROV looks to be:
  - Easy
  - Useful – people really want provenance, reuse important
- Use of SKOS & PROV:
  - Paving the way to full LD

Conclusions

- Final thoughts:
  - This work brings well managed yet Open World to scientific data
  - Overcomes several major problems with Linked Data:
    - Relevance to non ontologists – domain application
    - Computational efficiency – appropriateness
    - Tooling – wrapping existing tools

Australian Government
Geoscience Australia

netCDF-LD SKOS:

demonstrating Linked Data vocabulary use within netCDF-compliant files

Nicholas Car
Data Architect
Geoscience Australia
nicholas.car@ga.gov.au

Code: http://pid.geoscience.gov.au/dataset/103620
Vocab: http://pid.geoscience.gov.au/def/voc/netCDF-LD-eg-ToS