The ACORN-SAT Linked Climate Dataset

Laurent Lefort [a,*], Armin Haller [a], Kerry Taylor [a] and Andrew Woolf [b]
[a] *CSIRO ICT Centre*
GPO Box 664, Canberra, ACT 2601, Australia
E-mail: {firstname.lastname}@csiro.au
[b] *Australian Bureau of Meteorology*
GPO Box 2334, Canberra ACT 2601, Australia
E-mail: a.woolf@bom.gov.au

Abstract. The Australian Bureau of Meteorology has recently published a homogenised daily temperature dataset, ACORN-SAT, for the monitoring of climate variability and change in Australia. The dataset employs the latest analysis techniques and takes advantage of newly digitised observational data to provide a daily temperature record over the last 100 years. In this article we present how ACORN-SAT can be published as linked data with the help of the Semantic Sensor Network ontology and the RDF Data Cube vocabulary. We describe how the proposed service can make such datasets more accessible and linkable with other resources and how to handle issues which are specific to such long term climate data time series. The resulting Linked Sensor Data Cube is accessible online via a pilot government linked data service built on the Linked Data API at `lab.environment.data.gov.au`.

Keywords: Climate Linked Data, Sensor Network Ontology, meteorological observations, historical climate change

1. Introduction

The Australian Climate Observations Reference Network - Surface Air Temperature (ACORN-SAT) dataset [3,14], a flagship data product of the Australian Bureau of Meteorology (BoM), has been developed for monitoring climate variability and change in Australia. The dataset provides a daily temperature record over the last 100 years. Its primary objective is to underpin better understanding of long-term climate change. To produce this dataset, climate data experts [14,15] have used all the available information about weather station relocations, changes in technology and changes in observational procedures to characterise breakpoints in time series and to compute adjustments for each station. This dataset has been released by the BoM for reuse as open data to fulfil the Australian government commitment for Open Government. In this article we describe how we transformed the originally released tabular data into a Linked Sensor Data Cube [11] based on the W3C Semantic Sensor Network ontology [4] and the W3C RDF Data Cube vocabulary [5]. It is now accessible online as a pilot government linked data service including a Linked Data API[1] at `lab.environment.data.gov.au`. This article describes the motivations and benefits of publishing the ACORN-SAT data as linked data in the context of national and international initiatives (e.g. `surfacetemperatures.org`) for the provision of long term climate data time series.

The remainder of this article is structured as follows. In Section 2, we outline the ACORN-SAT Linked Sensor Data Cube creation process. In particular, we describe what ontologies we used and extended, how we created and published the data, and how we interlinked it to other resources. In Section 3, we describe how the resulting product is already used and what opportunities exist to link the data with other available resources. We also discuss how a linked data solution can help the climate science community to publish more com-

[*] Corresponding author. E-mail: laurent.lefort@csiro.au.

[1] http://code.google.com/p/linked-data-api/

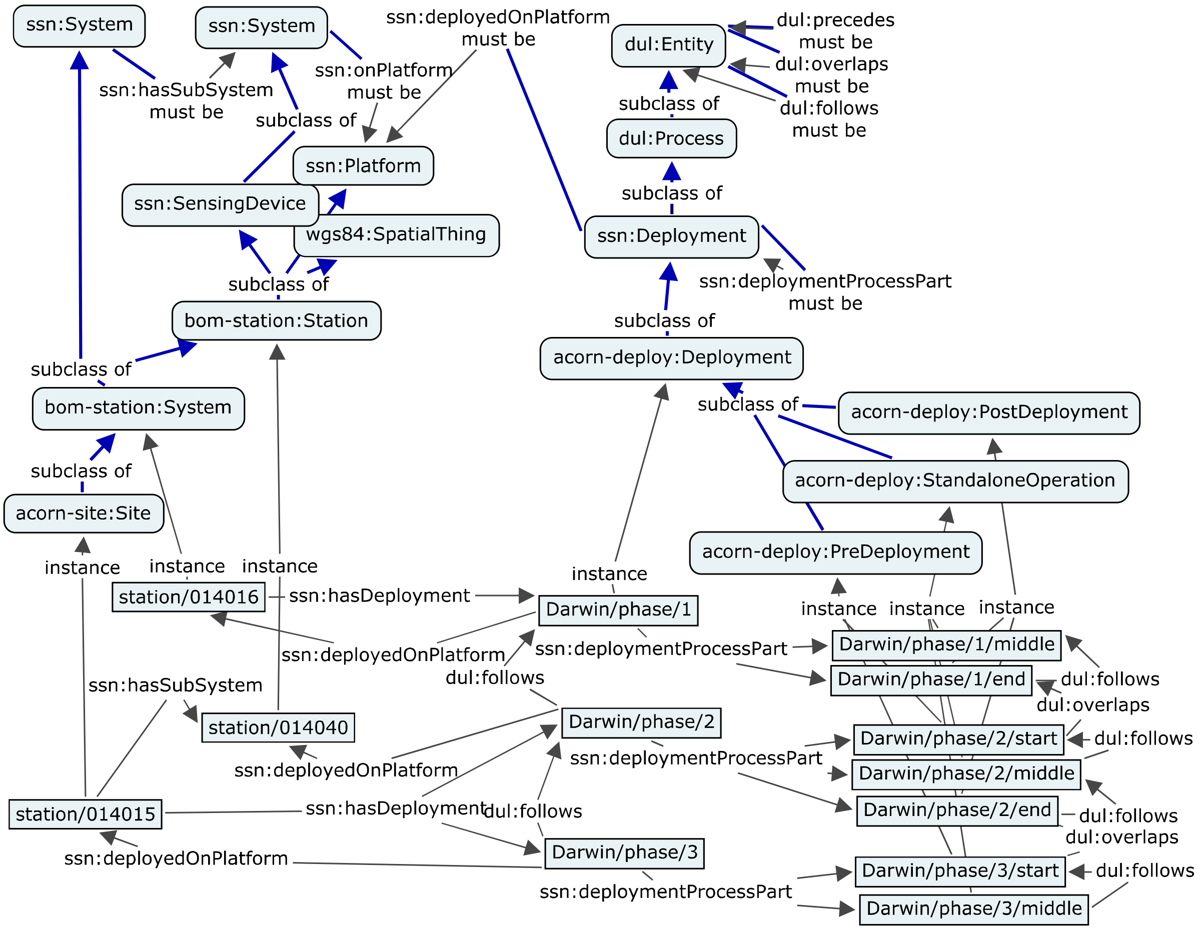

Fig. 1. ACORN-SAT system, sub-systems, deployment phases and sub-phases for Darwin [11]

prehensive climate data resources before we conclude in Section 4.

2. The ACORN-SAT Linked Sensor Data Cube Creation Process

The process of creating linked data from the ACORN-SAT dataset involved four major steps: (1) identifying and defining ontologies to represent the concepts and relations in ACORN-SAT, (2) creating the RDF data triples from the tab-delimited data and defining a URI scheme, (3) publishing the RDF triples in a linked Data fashion using ELDA[2] and (4) establishing links to other linked datasets.

2.1. *The ACORN-SAT Linked Sensor Data Cube structure*

The ACORN-SAT linked dataset is derived from three resources. The ACORN-SAT dataset originally released by the Bureau of Meteorology is available as a set of tab-delimited data files (source 1) which contain the homogenised minimum and maximum temperature and the raw rainfall data recorded daily at each selected site extending from 1910 to the present. The BoM has published the associated site metadata via a station catalogue document [2] (source 2) with a map and a photo for each site, as well as the name, number, geographical coordinates, locality and some text about the site and its history. The BoM also maintains a Weather Station Director[3] which contains metadata

[2] http://code.google.com/p/elda/

[3] http://www.bom.gov.au/climate/data/stations/

for more than 20,000 bureau stations (source 3) like the associated *rainfall district* and *rainfall state* as defined by the Bureau.

The modelling of the ACORN-SAT dataset posed two main challenges for the design of the ontologies, (1) the modelling of sensing assets changes over time and (2) the modelling of a dimensional array of observation values.

The first challenge arises from the selection of the stations by the Bureau of Meteorology for the ACORN-SAT locations. The locations (112 in total) have been sourced from a set of single or composite stations selected according to the availability and quality of the data [14,15]. The published documentation [3] explains the numbering system used by the Bureau of Meteorology and the methods used to manage the changes of stations at each site ([15], section 2.4 and 3.4). During each transition period, one of the sites, generally the old one, is kept as a comparison site for a minimum period of five years of parallel observations. These modifications of the network structure are related to factors such as the urbanisation of the original site, in particular, the construction of new buildings affecting the observations, and the systematic transfer of bureau-staffed sites from city centres to airports. For example, the Darwin ACORN-SAT site is sourced from three successive deployment phases (1910-1942, 1941-2007 and 2001-now) [11]. The first phase corresponds to the observations done at the Darwin Post Office (PO) from 1910 to 1942. The BoM code for this station is 014016. The second and third phases correspond to two separate sites at the Darwin Airport (AP), located one kilometre away from each other. One common BoM code (014015) is used for the observations made at these two sites. During and after the overlap period, the decommissioned site is renamed and referred to as the Darwin Airport Comparison station (014040).

ACORN-SAT utilises improved analysis techniques which exploit pair-wise comparison between one station and up to 10 comparable stations used as references. The algorithm applies differential adjustments to different parts of the daily temperature frequency distribution to better estimate the deltas with reference stations before and after an inhomogeneity. The ACORN-SAT method uses all the available background information or metadata about station moves, changes in technology and in observational procedures to identify and locate the breakpoints and to evaluate and validate the adjustments.

The Semantic Sensor Network ontology can help to unambiguously describe how the time series data for each sensing site have been assembled for all stations. To do so, we have extended it with the `acorn-deploy:Deployment` class and its three sub-classes (see Fig. 1) to explicitly model the start, middle and end phases of a deployment. We reuse the `dul:follows` and `dul:overlaps` properties from the DOLCE Ultra Lite (DUL) upper ontology [7] to specify the temporal relationships between phases and sub-phases. The “logical" codes used by the Bureau of Meteorology for the publication of observation data and station metadata are also captured at different levels (time series, phases and sub-phases).

For the second challenge, the modelling of a dimensional array of observation values, we have identified the RDF Data Cube vocabulary [5], a vocabulary for the publication of statistical data in RDF, published by the W3C Government Linked Data working group, to be suitable.

The ACORN-SAT data cube follows the general design principles defined by the Statistical Data and Metadata Exchange (SDMX) initiative [13]. It has two primary dimensions, a spatial one for the ACORN-SAT site and a temporal one with three levels (year, month, day) for the date of the observation. Each observation in the original ACORN-SAT dataset contains three daily measures referring to a 24 hour interval, encoded as `interval:CalendarInterval` instances: the minimum and maximum temperature, and the rainfall amount, plus some extra boolean attributes to signal missing values. The data cube itself is divided into slices using the site id first, and then the year and month of the observation. All the slices are compound observations and are enriched with extra statistical attributes. For the temperature measures, we pre-compute the minimum, maximum, mean, and standard deviation indicators over the relevant time period and add them to each slice. For the rainfall measure, we have added the maximum, and the sum. The count of missing measurements is also provided. Finally, the start and end dates of the period are encoded as `interval:CalendarInstant` instances.

Fig. 2 shows how we have integrated the RDF Data Cube vocabulary (QB) and the Semantic Sensor Network ontology (SSN):

- `acorn-sat:Observation` defined as `qb:Observation` and `ssn:Observation`,
- `acorn-sat:TimeSeries` defined as `qb:Slice` and `ssn:Observation`.

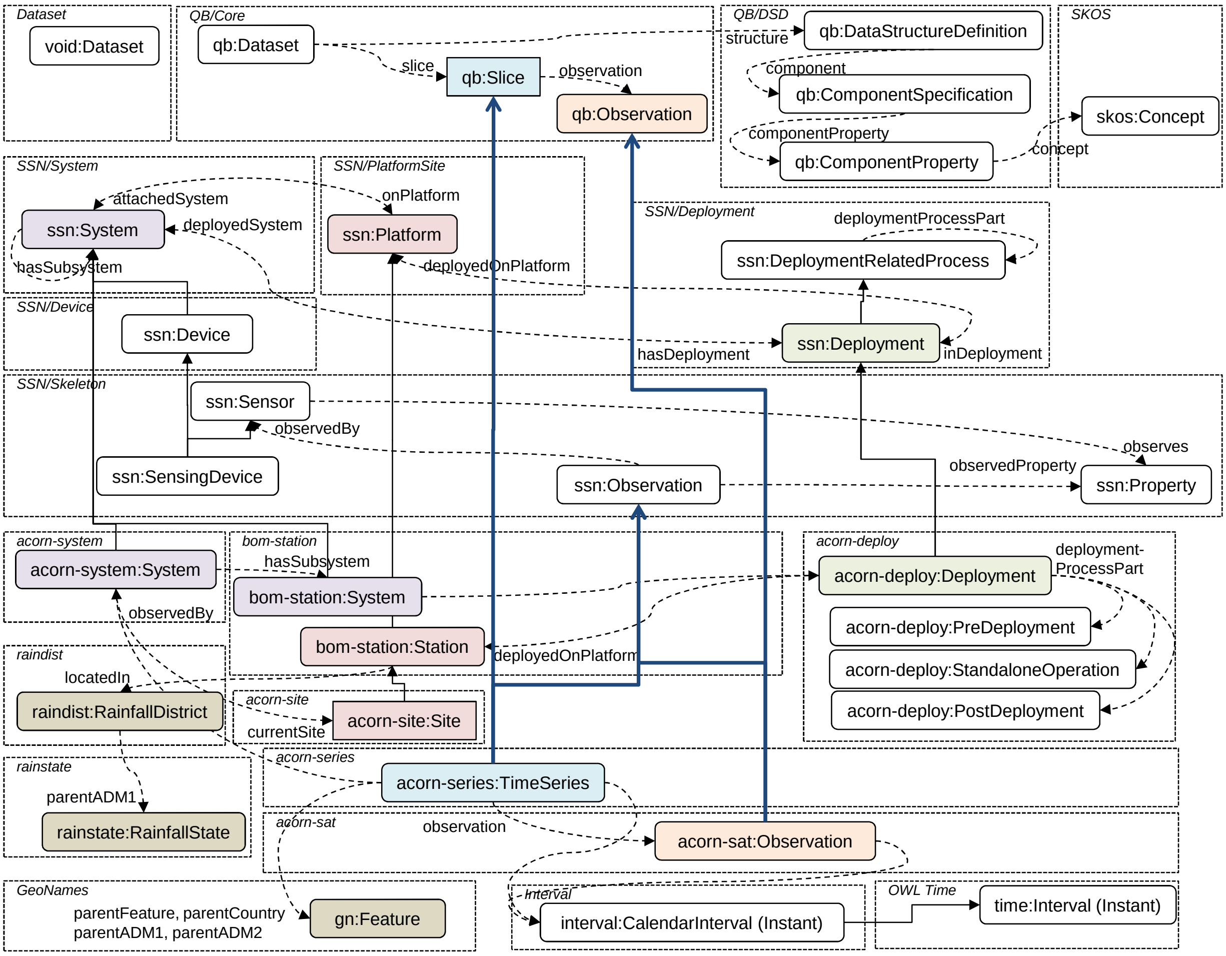

Fig. 2. The ACORN-SAT Linked Climate dataset key concepts and relationships, split by conceptual modules: plain lines are used for sub-class-of relationships and dashed lines for object properties linking classes

The boxes in Fig. 2 delineate single ontologies except for SSN [4] and QB. The colours indicate which classes from which ontology modules are coupled together via inheritance relationships. The prefixes used in Fig. 2 are listed in Table 1 (existing vocabularies and ontologies) and in Table 2 (new ones). Table 2 also gives a more complete list of the reused classes.

We have noted [11] that the declarations of the observed properties in the SSN ontologies as classes are not directly compatible with their declarations in the RDF Data Cube as properties. Further coordination with the editors of the RDF Data Cube Vocabulary specification [5] is needed to evaluate if it is acceptable to bundle together the SSN classes defining the observed properties and the SKOS concepts which can be attached to the Data Cube properties via the `qb:concept` property.

2.2. *Creating the ACORN-SAT linked data and defining the URI scheme*

We mapped the tabular time series data of the original ACORN-SAT to RDF using D2RQ and custom-built XSLT and Python scripts. These scripts produce RDF data based on the ACORN-SAT ontologies listed in Table 2. We have largely followed the URI guidance issued for the publication of public sector data in the UK [6] (see Table 3, Table 4): in particular, we use data.gov.au as the root domain for URI sets that are promoted for re-use within the Australian Government and “domain" prefixes like “environment" to split the governance of these URI sets into sectors matching the competencies of agencies owning shareable data.

The scheme also supports *Concept Identifiers* with a URI starting with `def`, based on a word capturing the essence of the real-world “thing" that the set

Prefix	Name	URI
void	Vocabulary of Interlinked Datasets	`rdfs.org/ns/void`
qb	RDF Data Cube Vocabulary	`purl.org/linked-data/cube#`
skos	Simple Knowledge Organization System	`www.w3.org/2004/02/skos/core`
ssn	Semantic Sensor Network ontology	`purl.oclc.org/NET/ssnx/ssn`
gn	GeoNames ontology (version 3.1)	`www.geonames.org/ontology#`
wgs	Basic Geo (WGS84 lat/long) vocab.	`www.w3.org/2003/01/geo/wgs84_pos`
time	Time Ontology in OWL	`www.w3.org/2006/time`
interval	Interval URI Sets	`reference.data.gov.uk/def/intervals`

Table 1: Reused vocabularies

Ontology	URI	Reused Classes
Observation	root/def/acorn/sat	`ssn:Observation`,`qb:Observation`
Time-Series	root/def/acorn/time-series	`ssn:Observation`,`qb:Slice`
System	root/def/acorn/system	`ssn:System`
Station	root/def/stations/station	`wgs:SpatialThing`,`ssn:SensingDevice`,`ssn:Platform`
Site	root/def/acorn/site	`wgs:SpatialThing`,`ssn:SensingDevice`,`ssn:Platform`
Deployment	root/def/acorn/deployment	`ssn:Deployment`
Rainfall District	root/def/stations/raindist	`gn:Feature`
Rainfall State	root/def/stations/rainstate	`gn:Feature`

Table 2: New ontologies. The URI `root` is `lab.environment.data.gov.au`.

Resource Type	URI Pattern	Description
Identifier URI (Slice)	root/data/acorn/climate/slice/station/{id}	Aggregate observation data and metadata for a given station location
Identifier URI (subSlice)	root/data/acorn/climate/slice/station/{id}/year/{year}	Observation slice for a given location and year
Identifier URI (subSlice)	root/data/acorn/climate/slice/station/{id}/year/{year}-/month/{month}	Observation slice for a given location, year and month
Identifier URI (Observation)	root/data/acorn/climate/series/{id}/date/{date}	One Observation
List URI	root/data/acorn/climate/slice/station	A list of all observation slices

Table 3: URI patterns. The URI `root` is `lab.environment.data.gov.au`.

Class	URI Pattern	Description
Station	root/id/station/{station}	Weather station listed in the BoM Directory
Rainfall District	root/id/raindist/{raindist}	Rainfall Districts defined by BoM
Rainfall State	root/id/rainstate/{rainstate}	Rainfall State defined by BoM

Table 4: Linkable IDs. The URI `root` is `lab.environment.data.gov.au`.

names (e.g. `lab.environment.data.gov.au/def/station/Station`) and *Individual Identifiers* with a URI starting with `id`, based on a code used to identify an individual instance of a concept, if possible based on existing ID schemes: for example, `lab.environment.data.gov.au/id/station/023090` reuses the code defined by BoM for a `Station` located near Adelaide.

This URI scheme supports access to the published data with third party tools based on the Linked Data API.
Table 3 describes the URI patterns for API calls for individual data items (Identifier URIs) based on URLs finishing with identifiers (i.e. {id},{year},{date}) and URI patterns for API calls for lists of items (List URIs) based on URLs finishing with a keyword like station, year or month. API calls using these nested keywords/identifiers patterns are easier to learn and memorise, especially when their design exactly mirrors the structure of the underlying data.

2.3. Publishing the ACORN-SAT Dataset

The publication of this dataset represents a milestone in e-government in Australia – it is the first linked data published by the Australian Government's open data sharing initiative known as `data.gov.au`. It is also the first attempt to publish a 100 year climate time series as an RDF Data Cube. The ACORN-SAT Linked Sensor Data service [11] uses the ELDA open source implementation of the Linked Data API. Due to the size of ACORN-SAT (∼61 million triples) we have put particular focus on the usability (performance) of the exposed APIs and defined custom viewers for the different API endpoints in order to avoid expensive SPARQL CONSTRUCT queries. Our production environment serving `lab.environment.data.gov.au` uses a Virtuoso triple store and runs on an Amazon cloud.

The ACORN-SAT dataset has also been uploaded on the CKAN archive: its key characteristics are listed in Table 5 and Table 6.

URL	`lab.environment.data.gov.au/`
SPARQL	`lab.environment.data.gov.au/sparql`
DataHub	`datahub.io/dataset/acorn-sat`
VoID	`lab.environment.data.gov.au/data/acorn.rdf`
Licensing	`www.bom.gov.au/other/copyright.shtml?ref=ftr`

Table 5: Technical details

Category	Resources
All	61164662
Observation	4172560
TimeSeries	149968
Deployment	203
PreDeployment	72
StandaloneOperation	203
PostDeployment	72
Station	200
System	112
Site	112
RainfallDistrict	114
RainfallState	4
Links to GeoNames	576

Table 6: Key Statistics

2.4. Interlinking - enrichment of existing resources

Like LinkedSensorData [12] and AEMET [1], we have linked the BoM stations to their associated GeoNames features with the help of the GeoNames API. The BoM stations are also published with a different URI scheme (Table 4) in anticipation of the release of larger chunks of the Weather Station Directory for climate datasets which may use different Weather Stations. External datasets can link to ACORN-SAT, either temporally or spatially.

We provide temporal slices for each year and month of observations for a given station and consequently, we have 1300 (100 year slices plus 12 x 100 month slices) temporal slices for each location that could be linked from other temporal datasets. Table 3 shows the URI pattern of these slices and of their associated observations. We have not published any spatial slice, because in ACORN-SAT the ratio between the number of sites and the number of rainfall districts is close to one. So, we only have one level in the spatial dimension. However, we have also linked all the base observations made on the same day to the corresponding UK interval[4] object. Thus, we can run a SPARQL query like the one below to get all the available data for the particular day defined via its reference.

```
SELECT ?x
  WHERE {
     ?x rdf:type acorn-sat:Observation .
     ?x acorn-sat:dailyPeriod
       <http://reference.data.gov.uk/
       id/gregorian-interval/
       1935-07-27T09:00:00/PT1D }
```

[4] http://reference.data.gov.uk/def/intervals

Further, the inclusion of the links to the rainfall district and rainfall state also supports the comparison of observations from ACORN-SAT with data from other datasets by their geographical location defined by the `gn:locatedIn` and `gn:parentADM1` relations.

3. Benefits and opportunities for linked data producers and consumers

3.1. Transparency and reproducibility

One key challenge for publishers of climate data is to answer the public demand to have a more transparent and reproducible homogenisation process. By coupling the SSN ontology and the RDF Data Cube vocabulary, we are able to capture the station history and to attach it to the data at the right level of temporal and spatial granularity. We have used the semi-structured information available in the ACORN-SAT reports [2,14,3] and made the stations changes metadata shown in Fig. 2 directly accessible via a range of API endpoints. Approximately half of the adjustments done on the ACORN-SAT minimum and maximum temperature values [15] are supported by metadata records of which 80% were linked to station moves. We have captured the major moves occurring mainly in the 1940s and the 1990s when the sites of observations were transferred from town centres to airports. But there are changes in the sensor locations, technologies and observation procedures [14] mentioned in the Station Catalogue document [2] which have not been published yet. For example, the construction of buildings and the growth of vegetation close to a weather station can have a significant impact on the observations. We would also have liked to capture the known changes of observation times which affect the long-term time series of extremes measurements: in Australia, before 1964 which is the date when the BoM switched to the current 0900-0900 standard, about 30% of the weather stations used a 0000-0000 day.

Meteorological agencies like BoM are also packaging the climate observations into data products which are complex to compare, because they are based on collections of weather stations assembled into observation networks built at a national (e.g. Reference Climate Stations), regional (e.g. Regional Basic Climatological Network) and global scale (e.g. the Global Climate Observing System). Until now one of the challenges of combining national and international climate datasets has been the numbering scheme of observation locations. With unambiguous metadata about station locations and their sensing device type, national and international climate datasets will be easier to integrate.

3.2. Developing Mashups

Recognizing that the ACORN-SAT dataset with its view dimensions is not particularly suitable for a faceted browsing interface like ELDA we have developed some additional mashups. For example, we provide a web mapping service where we embed the 112 sensor locations of the ACORN-SAT dataset in a Google map widget[5] to let a user explore the yearly, monthly and daily (min, max, mean) temperature for a chosen location on the map in a Google Area Chart. We also have a simple query interface[6] where a user can provide a date range for a chosen location via a dropdown box. Based on the user input a SPARQL query is constructed and the resulting JSON document is used to plot the min and max temperature and the rainfall data on a Google Area chart. These services are described in more details in our previous publication [11].

3.3. Identification of new opportunities

The ACORN-SAT dataset is primarily a *dense* [13] data cube. Its structure can be reused for other data cubes suppplying long term time series like census or biodiversity data, which then will become easy to integrate together via links established at the slice level. Linking at the observation level is also possible e.g. to use the archived data to compare an observation done on a particular day or week to measurements done on the same day or week in the past 100 years. This type of comparison can provide valuable information on the rarity or severity of extreme weather events, but it should be done with caution because of the low number of ACORN-SAT locations. Another long term Australian dataset could be used in this context: AWAP[7], the Australian Water Availability Project which uses gridded cells with a surface of approximatly 25 km^2. The World Bank has published gridded data as linked data[8], but not yet at the level of detail

[5] http://lab.environment.data.gov.au/mashup/drilldown

[6] http://lab.environment.data.gov.au/mashup

[7] http://www.eoc.csiro.au/awap/

[8] http://data.worldbank.org/developers/climate-data-api

which would enable point-based or cell-based comparison of available datasets.

We are also interested in opportunities to link the ACORN-SAT dataset to *sparse* climate data like cyclone tracks. To our knowledge, no one has systematically linked observations done on a cyclone path to ground observations done at nearby weather stations. With the approach presented here, this could be done without extra duplication of the published observation data, by adding new slices to the data cube structure.

4. Conclusion

In this pilot project, an important public dataset, ACORN-SAT has been made available as Climate Linked Data. We believe that the explicit support for metadata attachment as offered by the linked data approach presented here is of prime importance to the publication of climate data, and may help to enrich the public debate about the scientific foundations for climate science.

Thanks to the coupling of the Semantic Sensor Network ontology and the RDF Data Cube vocabulary, we can now publish valuable metadata alongside the observation data. Like the UK Bathing Water project (environment.data.gov.uk/lab/), this project has also demonstrated the flexibility of the RDF Data Cube Vocabulary especially when it paired with the Linked Data API.

There are a number of use cases [8] and linking opportunities [9] which depend on the availability of complementary ontologies and vocabularies for the publication of geospatial and statistical linked data. The Provenance ontology [10] directly answers the demand for increased transparency and reproducibility.

We are now working on event detection on live data feeds from a soil moisture wireless sensor network deployed on a farm near Armidale in New South Wales, Australia. This sensor network has 112 wireless nodes and a 5 minute sampling period.

The lessons learned from this project are also shared through the newly established Australian Government Linked Data Working Group (AGLDWG) with participants from eight government agencies. The working group is developing technical guidances (URI rules, vocabularies and ontologies, deployment architecture, etc.) and best practice on the use of linked data by the Australian Government. The working group also consults in the launch of a new CKAN-based `data.gov.au` platform.

References

[1] G. Atemezing, O. Corcho, D. Garijo, J. Mora, M. Poveda Villalón, P. Rozas, D. Vila Suero, and B. Villazón Terrazas. Transforming meteorological data into linked data. *Semantic Web, (to appear).*, 2012.

[2] BOM. ACORN-SAT station catalogue - Report 5 for the Independent Peer Review of the ACORN-SAT data-set. Technical report, Bureau of Meteorology, 2011.

[3] BOM. Australian Climate Observations Reference Network - Surface Air Temperature (ACORN-SAT). http://www.bom.gov.au/climate/change/acorn-sat/, 2012.

[4] M. Compton, P. Barnaghi, L. Bermudez, R. Garcia-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. L. Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, and K. Taylor. The SSN ontology of the W3C semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17(0):25–32, 2012.

[5] R. Cyganiak, D. Reynolds, and J. Tennison. The RDF Data Cube Vocabulary. W3C Working Draft, W3C, April 2012.

[6] P. Davidson. Designing URI Sets for the UK Public Sector. Technical report, UK Chief Technology Officer Council, 2010.

[7] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, R. Oltramari, and L. Schneider. Sweetening Ontologies with DOLCE. In *Proc. of 13th Int. Conf. on Knowledge Engineering and Knowledge Management*, pages 166–181. Springer, 2002.

[8] B. Kämpgen and R. Cyganiak. Use Cases and Requirements for the Data Cube Vocabulary W3C Working Group Note Editor's Draft. Technical report, World Wide Web Consortium, February 2013.

[9] S. Kramer, A. Leahey, H. Southall, J. Vompras, and J. Wackerow. Using RDF to Describe and Link Social Science Data to Related Resources on the Web. Working paper, German Council for Social and Economic Data (RatSWD), 2012.

[10] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology W3C Candidate Recommendation. Technical report, W3C, December 2012.

[11] L. Lefort, J. Bobruk, A. Haller, K. Taylor, and A. Woolf. A Linked Sensor Data Cube for a 100 year homogenized daily temperature dataset. In *Proc. of the Semantic Sensor Networks Workshop (SSN 2012)*, volume 904. CEUR-Proceedings, 2012.

[12] H. Patni, C. Henson, and A. Sheth. Linked Sensor Data 2010. In *Proc. of the Int. Symp. on Collaborative Technologies and Systems (CTS 2010)*, pages 362–370. IEEE, 2010.

[13] SDMX. SDMX Guidelines for the Design of Data Structure Definitions. Technical Report Version 0.7, Statistical Data and Metadata Exchange initiative, September 2012.

[14] B. Trewin. A daily homogenized temperature dataset for Australia. *Int. J. Climatol.*, 2012.

[15] B. Trewin. Techniques involved in developing the Australian Climate Observations Reference Network Surface Air Temperature (ACORN-SAT) dataset. Technical report, Centre for Australian Weather and Climate Research, 2012.