*Formal Ontology in Information Systems*
*R. Ferrario and W. Kuhn (Eds.)*


*doi:10.3233/978-1-61499-660-6-67*

# Knowing Whether Spatio-Temporal Analysis Procedures Are Applicable to Datasets

Simon SCHEIDER [a,c,1], Martin TOMKO [b]
[a] *Department of Human Geography, University Utrecht, The Netherlands*
[b] *Department of Infrastructure Engineering, University of Melbourne, Australia*
[c] *Institute of Cartography and Geoinformation, ETH Zuerich, Switzerland*

**Abstract.** How can data analysts identify spatio-temporal datasets that are suitable for their task? Answering this question is not only dependent on the aim of the analysis and the semantic contents of the data, but also on knowing whether the required data combinations and transformations, spatio-temporal analysis methods, charts and map visualizations are *meaningfully applicable* to the data. Operators need to assess whether they can meaningfully apply analytical operations to data to derive the information required. Answering this question in a general and computationally executable way is a crucial step on our way towards supporting data analysts and their research practice in e-Science. We propose an ontology design pattern for spatio-temporal information that enables to reason about the applicability of a number of fundamental classes of analyses in relation to given data, i.e., whether data sets can be compared, transformed, combined, and whether summary statistics can be applied to them. We demonstrate this ontology implemented in OWL through a set of corresponding SPARQL queries applied to meta-data of datasets from the AURIN portal.

**Keywords.** spatio-temporal analysis, ontology design pattern, e-science, analysis support, applicability

## 1. Motivation and Related Work

The analytical opportunities opened by the wide availability of large data collections on the Web vastly exceeds analysts' manual capacities to search, identify and evaluate these datasets, assess their fitness for exploring a specific research question and extract relevant parts for analysis. Progress in *data driven e-Science* [1,2] depends to a large degree on supporting to researchers with tools enabling the identification of appropriate data for analysis. Manual assessment of a dataset's suitability is increasingly difficult, and the manual identification of appropriate analysis tools even more so.

Research in e-Science has primarily focused on *content-based* data integration. Tags, controlled vocabularies and content ontologies have been used to describe a domain of interest in a way that is general enough to integrate models [3] or data sets [4,2,5] from

[1]Corresponding Author: Department of Human Geography and Spatial Planning, University Utrecht, The Netherlands; E-mail: simonscheider@web.de

different sources. These approaches assume that once models and data are integrated, analysis can be seamlessly executed – a promise enthusiastically welcomed by big data scientists aspiring to expose vast data to automated machine learning with the aim to synthesize knowledge about all and any domain of human inquiry [6]. However, this view seems to underestimate the role of domain knowledge as a *prerequisite* to meaningful analysis, not purely as its outcome.

Consider the computation of the ratio of unemployed members of a population across a country. This task requires knowledge about the equivalence of temporal and spatial references of underlying population counts, as well as about the method of data collection. We support the argument that *"the data train needs semantic rails"* [7] precisely because knowledge is a fundamental enabler to meaningful application of analytical tools, statistical visualizations and thematic mapping to spatio-temporal data. We thus see the importance of knowledge representation [8,9,10] for analytical data workflow support. While such knowledge is subject of introductory curricula in data analysis [11], it is thus far still largely inaccessible to machine interpretation.

Though operation sequencing in workflows has been investigated before, criteria for deciding whether analysis procedures are meaningfully applicable to datasets are still lacking [10]. Type-safe programming languages [12], integrity constraints in relational databases [13], and summarizability in Online Analytical Processing (OLAP) [14,5] provide useful computational approaches. Multidimensional data quality indicators have been identified as useful means to assess a dataset's fitness for use [15,16]. However, these approaches hardly provide an automated method for decisions about applicability. Once such a method is available, OLAP models published on the Web as linked data [17], together with provenance [18] and statistics vocabularies [19,20] could be used to link data across analytical infrastructures.

We believe that a method to assess the applicability of an analytical procedure must be based on a semantic description of the input data on a conceptual level, thus reaching beyond data formats and data types [21,10,9], in a manner similar to semantic Web geo-processing [22]. In this article, we propose an ontology design pattern (Sect. 3) that assists analysts in assessing on a very fundamental level whether data items can be compared, statistically summarized, transformed and combined in space and time (Sect. 2). We propose corresponding queries (Sect. 4) and test the pattern on data from the data analysis portal of the Australian Urban Research Infrastructure Network (AURIN) (Sect. 5).

## 2. Spatio-temporal Analysis Tasks and Competency Questions

To design and test an ontology design pattern, it is critical to know its purpose. One approach is to design a set of explicit *competency questions* about a matter of interest [23]. This is a set of questions that are systematically answered to help analysts identifying appropriate data for their tasks. The design pattern can be tested against arbitrary datasets to evaluate whether it provides appropriate and correct answers. We now introduce four competency questions required by our ontology pattern.

### 2.1. Are Data Items Comparable with Each Other?

Analysts frequently need to visualize the spatial distribution of a phenomenon. The thematic map shown in Figure 1[2] might tell them that the fraction of volunteers in the population of Victoria, AU decreases with the proximity to the city of Melbourne but that there might be spatial outliers to this spatial decay trend.

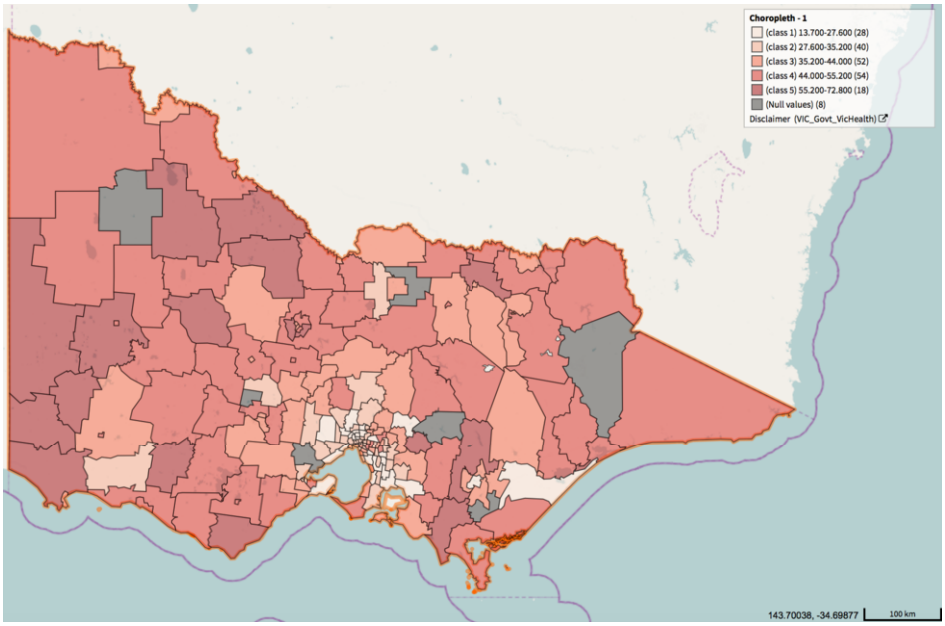



**Figure 1.** Choropleth maps show the spatial variation in the distribution of a measure (here: percentage of population taking part in volunteering work) and allow to compare the measure across irregular regions.

This thematic mapping silently makes certain assumptions which should not be taken for granted by analysts. It is implicitly assumed that not only the phenomenon measured, but also its *encoding stays invariant across space*, *while the context of measurement changes*. It is further assumed that the methodology of the census population counting methodology was homogeneously applied across all the regions shown in Fig. 1 and that the values reported in the dataset consistently denote percentages, not ranks or absolute numbers for the measures, i.e., that the data reporting is homogeneous for all the records in the dataset. Similarly, in global weather datasets we expect that temperatures are reported using the same measurement units and not change from °F to °C depending on whether they report US or European temperatures.

Comparability is a fundamental requirement for computing any *statistic* over a (univariate) measure where data items can be compared along one or several dimensions, and it basically requires *the sharing of reference systems among measured values* (compare Sect. 4.1).

[2]Source: `http://docs.aurin.org.au/portal-help/visualising-your-data/map-visualisations/choropleth/`.

### 2.2. Are descriptive statistics applicable to data sets?

Descriptive statistics provide a summary representation of a dataset whose measured values are comparable. The way how a descriptive statistics summarizes data, however, additionally requires a certain *scale level* [24] (for a criticism of scale levels from the spatio-temporal analyst perspective, see [25]). Thus, this requirement is more restrictive than comparability discussed prior. For example, if we summarize a dataset based on a measure of central tendency, such as the mean, then the values in this dataset need not only be comparable, they also need to be on an interval scale level. The median, in contrast, is applicable to ordinal scale levels as well.

Analogously, summary visualizations are only meaningfully used with values on a particular scale level. A scatterplot should be used to visualize the distribution of a continuous measure as a function of another continuous one, in contrast to the histogram which should only be used to depict the frequency of binned continuous measurement values. This is again in contrast to the bar chart with which it is often confused, and which allows for a discrete, arbitrarily ordered x-axis applicable to nominal values. Yet, any numeric data could be input to the dependent y-axis of a bar chart–although not meaningfully–making it difficult to filter away numbers expressing nominal values (e.g., IDs of regions are often integers).

### 2.3. Are data items transformable into each other?

Analysts often need to relate values that are not directly comparable because they are encoded in different ways. For example, a common problem in Geographic Information Systems (GIS) is the transformation of geo-coordinates to and from different spatial reference systems. It is necessary whenever *identical* locations captured in two maps are represented with different numeric coordinates [26]. Similarly, temperature values or heights with different datums and units of measure need to be transformed into each other to become comparable [25]. This problem is that of reference system transformations and it holds for all kinds of measurable or observable phenomena [27] such as when the method of census population counting changed [28].

The transformability requirement is *less restrictive* than comparability, since comparable datasets are (trivially) transformable, but not vice versa. If data are transformable, then it becomes possible to re-establish comparability within or across datasets. In demographic data, a socio-demographic measure such as unemployment may be represented as a regional rank, and in other cases as percentage, simply transformable into ranks by ordering.

### 2.4. When Can Data Items Be Combined?

Analysts often need to merge (possibly incomparable or even non-transformable) values from different sources to generate new data items. In the simplest case, they form *multidimensional data spaces*, where attributes from multiple tables representing the same entities become the dimensions. They may join datasets with health indicators for administrative regions from a dataset about type 2 diabetes and from a dataset about hypertension rates, for correlation analysis (see Fig. 2[3]), [5]. A different kind of combination of

[3]Soure: `http://docs.aurin.org.au/tutorials-and-use-cases/creating-a-thematic-map/`

datasets occurs when *deriving new variables*. One example are *part-whole ratios*. These are ratios of two values, one representing a subpopulation or proper part of the other, e.g., the percentage of population that engages in volunteering work (compare Fig. 1).

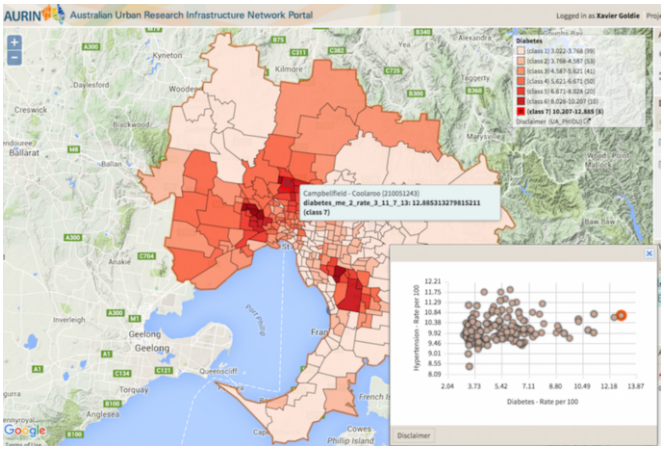



**Figure 2.** An example for combined health records about hypertension and type 2 diabetes.

In all these cases, analysts implicitly assume that, not the kinds of measures themselves, but the *context* in which combinable measures were taken are identical. In distinction to comparability or transformability, combinability allows different kinds of measures but requires identical measurement contexts. For example, in building a health record, we need to make sure that measures refer not only to the same region (identical identifier) but that these regions have not changed boundaries through time [29].

## 3. The Design Pattern in a Nutshell

In this section, we explain our design pattern and discuss its classes and properties. Throughout the paper, we use Description Logic (DL) expressions wherever possible [30]. The pattern is depicted in Fig. 3. An OWL[4] version of this pattern is available online[5]. Each node in this figure stands for one of the classes in Table 1. The main classes of the pattern are all meant to be disjoint:

$$Disjoint(Data, DataSet, Reference, RefSys) \tag{1}$$

Each black arrow stands for a property being a relation between instances of some class. The "from" class is called the domain, and the "to" class is called the range. Properties and their domains and ranges are listed in Table 2. Dotted arrows indicate that the first class is a subclass of the second class.

[4]Web Ontology Language `https://www.w3.org/TR/owl-features/`

[5]`http://geographicknowledge.de/vocab/analysisdata.rdf`. We use the prefix *ada* for this vocab. Note that it is richer than the classes shown in this paper.

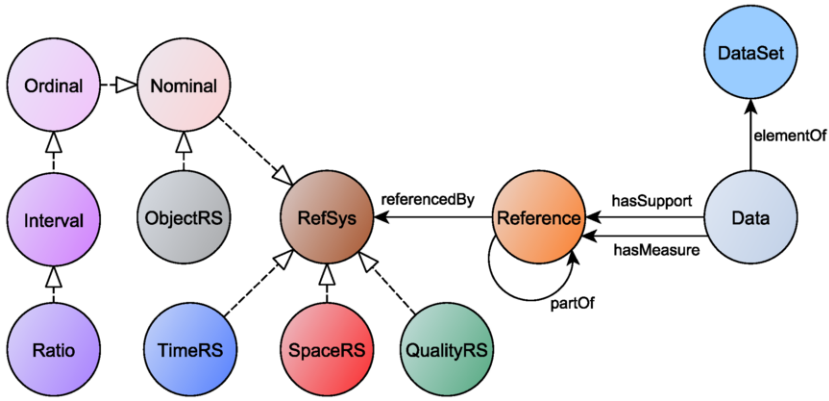



**Figure 3.** The pattern in a nutshell. White dotted arrows denote subclass relations, and black arrows denote binary relations.

**Table 1.** Pattern vocabulary: Classes

| Class | Explanation | Class | Explanation |
|---|---|---|---|
| *Data* | The class of data items | *SpaceRS* | Spatial reference systems |
| *DataSet* | The class of datasets | *TimeRS* | Temporal reference systems |
| *Reference* | The class of references | *Ratio* | Ratio scaled reference systems |
| *RefSys* | The class of reference systems | *Interval* | Interval scaled ref. sys. |
| *ObjectRS* | Ref. sys. for objects | *Ordinal* | Ordinal scaled ref. sys. |
| *QualityRS* | Ref. sys. for quality values | *Nominal* | Nominal scaled ref. sys. |

**Table 2.** Pattern Vocabulary: Properties

| Property | Domain | Range | Explanation |
|---|---|---|---|
| *hasMeasure* | *Data* | *Reference* | Measured reference of a data item |
| *hasSupport* | *Data* | *Reference* | Supporting reference of a data item |
| *partOf* | *Reference* | *Reference* | Reference is part of another reference |
| *elementOf* | *Data* | *DataSet* | Data item is element of a data set |
| *referencedBy* | *Reference* | *RefSys* | Reference system of a reference |

### *3.1. Reference Systems, Their Types and Scale Levels*

References are symbols for observable phenomena in the world [31], e.g., coordinate pairs of a statistical region's boundary refer to locations on the Earth surface, and its attribute number might refer to the population. These symbols need to belong to a reference system, a conventionally established way of interpreting symbols in observations [31]. Spatio-temporal reference systems include systems for geographic space (SpaceRS), time (TimeRS), measurable or observable qualities (QualityRS) as well as perceivable objects (ObjectRS). Note our distinction of objects (discrete entities such as buildings, cars, hurricanes) from qualities, such as a temperature or the number of objects in a community [10]. Measurement scales are reference systems that belong to a class of similar (homomorphic) numeric interpretations [32]. Note that object reference systems

are considered nominally scaled, yet it is not the case that all nominal scales refer to objects.

### 3.2. Data Items and Data Sets

A data item (instance of the class Data) is something that identifies and binds together a single observation with potentially many references. It corresponds to a record in a database table, which can have many attributes and is often identified by a primary key value. We require that data items cannot exist "alone": they need to be elements of some data set and they need to have certain kinds of references, namely a *measure* and a *support*:

$$Data \sqsubseteq (\exists elementOf.\top \sqcap \exists hasMeasure.\top \sqcap \exists hasSupport.\top) \quad (2)$$

### 3.3. Supports and Measures

The information contained in a data item is contained in its references, i.e., in information that is linked to the item, similar to table attribute values storing the information of a table record. In contrast to a table schema, we need to draw a conceptual distinction based on the role these references play in data analysis. A *measure* denotes an observed reference of a data item, whereas a *support* denotes the context of this observation which can be used to compare measures[6] [10]. Referents *of any kind of reference system* may play the role of support or of measure.

For example, the statistical data record expressed with the pattern in Fig. 4 has a population count as measure, and a spatial region and a time as its supports, meaning that the measure is valid (was observed) for this region and this time. Space and time, however, can also play the role of measures, e.g., a GPS trajectory has objects and time as supports and spatial points as measures [9,33].

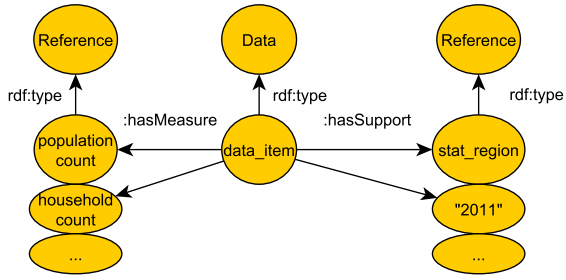



**Figure 4.** Distinction between data items, supports and measures.

[6]This distinction is similar to the one between *dimensions* and *measures* in OLAP data cubes. Thus, our vocabulary can be partially mapped to corresponding properties in the data cube vocabulary (`https://www.w3.org/TR/vocab-data-cube/`). Furthermore, we think the term "dimension" is easily confused with the dimensions of a multi-dimensional data space, which are measures and not supports. Compare also [33].

## 4. Inferring the Applicability of Analysis Procedures

How does the proposed design pattern help to infer whether a given analysis can be meaningfully applied to a data item or a data set? We propose a corresponding query expressed in terms of the pattern using SPARQL[7] and where possible, also DL for each analysis procedure discussed in Sect. 2. Examples of linked data graphs that satisfy these queries are given in Fig. 5.

### 4.1. Querying for Comparable Data Items

Data items can be meaningfully *compared* to a given example item (*ex*) if and only if the measured values are referenced by the same reference system:

$$Comp_{ex} \equiv \exists(hasMeasure \circ referencedBy \circ referencedBy^{-} \circ hasMeasure^{-}).\{ex\} \quad (3)$$

This corresponds to the following SPARQL query[8] (see also Fig. 5a), where *x* denotes data items comparable to *ex*, and *y e* are comparable references:

```
CONSTRUCT {?x ada:comparableTo ?ex}
WHERE {
?x ada:hasMeasure ?y. ?y ada:referencedBy ?r.
?ex ada:hasMeasure ?e. ?e ada:referencedBy ?r.
}
```

### 4.2. Querying for Datasets to Which a Statistic Is Applicable

We suggest that a (descriptive) statistic which requires scale level *l* can be applied to a data set *ds* if and only if for all data items in this set, all measures satisfy that scale level. This accounts also for multivariate data as a combination of different measures, where the requirement needs to be satisfied for each "dimension". For a given scale level, e.g. *ada:Ratio*, this pattern can be expressed in DL as:

$$statApplicable_{Ratio} \equiv \forall(elementOf \circ hasMeasure).(\exists referencedBy.Ratio) \quad (4)$$

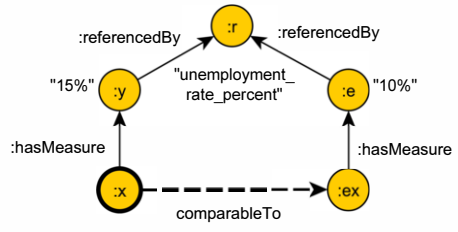
In SPARQL, this corresponds to two nested negated FILTER statements[9] (compare Fig. 5b):

```
CONSTRUCT {?z ada:statApplicableTo ?ds}
WHERE {
?ds a ada:DataSet.
FILTER NOT EXISTS {?x ada:elementOf ?ds. ?x ada:hasMeasure ?m.
FILTER NOT EXISTS{ ?m ada:referencedBy ?r. ?r rdf:type ?l.
FILTER (?l = ada:Ratio)}}
}
```

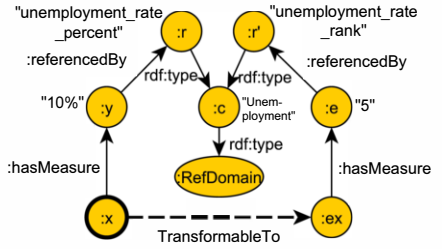
[7] https://www.w3.org/TR/sparql11-query/

[8] A CONSTRUCT query inserts new statements into the database, in our case applicability statements.

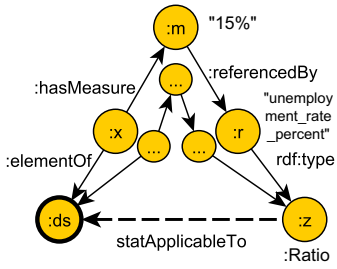
[9] Note that $\forall$ logically corresponds to $\neg\exists\neg$, and thus the allquantifier in DL can be translated as two nested FILTER NOT EXISTS statements.



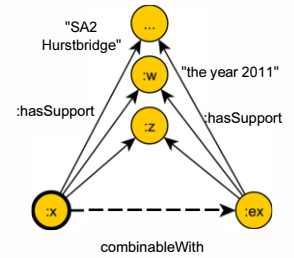

(a) Pattern CDI: for data items comparable to a given item (:ex).




(c) Pattern TDI: for data items transformable into the reference system of (:ex).




(b) Pattern DSLS: for data sets with applicable statistic on level *:z*.




(d) Pattern ADI: for data items strictly combinable with a given item (:ex).

**Figure 5.** Linked data graph patterns for comparability (a), applicability of statistics (b), transformability (c), and combinability (d). Examples explained in Sect. 5.

### *4.3. Querying for Transformable Data Items*

Data items can be meaningfully transformed into each other if and only if some measures are referenced by reference systems which denote a common domain. In SPARQL, we capture domains by mutually exclusive reference system classes $(c)$[10], where each class holds those reference systems that refer to a single domain. Note that this cannot be expressed in DL which cannot quantify over classes. The class *ada:RefDomain* contains all those domain classes:

```
CONSTRUCT {?x ada:transformableTo ?ex}
WHERE {
?x ada:hasMeasure ?w. ?w ada:referencedBy ?r. ?r rdf:type ?c.
?ex ada:hasMeasure ?y. ?y ada:referencedBy ?z. ?z rdf:type ?c.
?c rdf:type ada:RefDomain.
}
```

Note that comparability (Fig. 5a) is a special case of transformability (Fig. 5c), namely, when reference systems are identical.

### *4.4. Querying for Combinable Data Items*

Data items can be meaningfully *combined* if and only if they have common supports. The result of a combination is a new data item that contains the measures of the orig-

[10]These classes are mutually exclusive because a reference system must have a unique domain of interpretation.

inal ones, and which has common supports as its support. We distinguish two types of combinability: *strict* combinability requires all supports to be covered by the combinable dataset, *weak* combinability requires only a single common support. Data items *x* (weakly) combinable with *ex* are obtained by:

```
CONSTRUCT {?x ada:CombinableWith ?ex}
WHERE {
?ex ada:hasSupport ?support. ?x ada:hasSupport ?support.
        }
```

If a data item *x* is strictly combinable with *ex*, then additionally *x*'s supports must be part of *ex*'s supports:

$$Combinable_{ex} \equiv \exists hasSupport(\exists hasSupport^{-}.\{ex\}) \tag{5}$$

$$StrCombinable_{ex} \equiv \forall hasSupport(\exists hasSupport^{-}.\{ex\}) \tag{6}$$

In this case, *x* can be considered at least as general as *ex*, which is one requirement for a part-whole ratio in which *ex* is part of *x* and their measures form a ratio. In SPARQL, this can be expressed using two nested negated FILTER statements:

```
CONSTRUCT {?x ada:strCombinableWith ?ex}
WHERE {
?ex ada:hasSupport ?support. ?x ada:hasSupport ?support.
FILTER NOT EXISTS {?x ada:hasSupport ?w. FILTER NOT EXISTS{?ex ada:hasSupport ?w}}
        }
```

## 5. Application of the Pattern to AURIN Dataset and Results

We encoded part of the AURIN metadata with our ontology design pattern to show how datasets can be queried regarding the analysis concepts proposed in this article. AURIN is an Australian e-research effort providing online access to over 2000 datasets from federated data sources about the Australian urban environment. A Web portal enables data search, retrieval, integration, analysis and visualization [34]. The visual environment allows fast identification of areas of interest based on administrative partitioning and the retrieval of spatially related datasets. Linked visualizations of the data enable exploratory analysis and the discovery of patterns across datasets [5]. At the heart is a metadata system annotating datasources, contained datasets and their values, e.g., with the measurement level of each value. Dataset comparability is assured through a careful vetting process of the datasets and a tool-assisted metadata augmentation [35].

Fig. 6 shows an excerpt of five datasets that we have fully annotated using the proposed pattern, including data about socio-demographic status, education, indigenous status, health and quality of life, over different Australian statistical regions (Statistical Areas Level 2 (SA2) and Local Government Areas (LGA)) and for different reference times[11]. Since it was not possible to integrate with the operational infrastructure of the portal directly, we have translated only the metadata into our ontology and described the dataset itself with a minimal set of "blank noded"[12] data items.

[11] Note: in these statistical datasets there are no examples of supports other than space or time.

[12] https://www.w3.org/TR/2014/REC-rdf11-mt-20140225/#blank-nodes

| dataset | description |
|---|---|
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B40-sa2 | "SA2-based B40 Non-School Qualification: Level of Education by Age by Sex as at 2011-08-11" |
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B07-sa2 | "SA2-based B07 Indigenous Status by Age by Sex as at 2011-08-11" |
| http://data.aurin.org.au/datasource/dataset-UQeresearch-uq_polygons_lgas_socio_economic_variables-lga | "socio-economic, LGA, Local Government Area, 2006 census, 2001 census, 1996 census" |
| http://data.aurin.org.au/datasource/dataset-VicDOH-Health_lga_profilesdata2011-LGA | "LGA, Health, Well being, Quality of Life, Demographic, Population, Employment, Social" |
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B04-sa2 | "Age (individual years and 5-year age groups), and Sex" |

**Figure 6.** An excerpt of the datasets described with our pattern.

When running our queries against these metadata, we can establish that census data and health data are comparable because they both measure unemployment rates in the same way (see Fig. 7). Further, several census data sets are comparable because they all measure parts of the population. These datasets could then be visualized together, e.g., in a histogram or in a choropleth map. Further, a ratio-scaled (multivariate) statistic could

| datasetx | refSys | rlabel | datasetb |
|---|---|---|---|
| http://data.aurin.org.au/datasource/dataset-UQeresearch-uq_polygons_lgas_socio_economic_variables-lga | http://data.aurin.org.au/dataset/percent_unemployment_rate | "The percentage of the labour force which is unemployed." | http://data.aurin.org.au/datasource/data VicDOH-Health_lga_profilesdata2011-LG |
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B04-sa2 | http://data.aurin.org.au/dataset/persons_count | "Number of people" | http://data.aurin.org.au/datasource/data ABS-ABS_CENSUS2011_B07-sa2 |

**Figure 7.** AURIN datasets comparable via measures in a common reference system (refSys).

be applied to four of these five datasets, since all their measures are on a ratio scale level. Socio-economic and health datasets are transformable into each other because they both measure the common domain "unemployment" in terms of ranks and, alternatively, in terms of percent (Fig. 8).

| dataset | r | domain | rx | datasetx |
|---|---|---|---|---|
| n.org.au/datasource/dataset-as_socio_economic_variables-lga | http://data.aurin.org.au/dataset/percent_unemployment_rate | http://data.aurin.org.au/dataset/Unemployment | http://data.aurin.org.au/dataset/rank_unemployment_rate | http://data.aurin.org.au/datasource VicDOH-Health_lga_profilesdata201 |
| n.org.au/datasource/dataset-lga_profilesdata2011-LGA | http://data.aurin.org.au/dataset/rank_unemployment_rate | http://data.aurin.org.au/dataset/Unemployment | http://data.aurin.org.au/dataset/percent_unemployment_rate | http://data.aurin.org.au/datasource UQeresearch-uq_polygons_lgas_socio_economic |

**Figure 8.** AURIN datasets transformable into each other via a common domain.

Last but not least, unemployment rates can be weakly combined (but not compared!) with other indicators of social well being (health data), e.g., the number of family offences, for the purpose of further processing via common spatial LGA supports. Further-

more, census data on the level of SA2 regions can be strictly combined (Fig. 9) with descriptors of the educational and indigenous status, yielding a ratio of the amount of people with indigenous status across those regions in 2011.

| dataset | supportl | exdataset |
|---|---|---|
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B04-sa2 | "The year 2011" | http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B40-sa2 |
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B04-sa2 | "SA2 Sunbury" | http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B40-sa2 |
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B04-sa2 | "SA2 Hurstbridge" | http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B40-sa2 |
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B04-sa2 | "SA2 Sunbury" | http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B07-sa2 |
| http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B04-sa2 | "The year 2011" | http://data.aurin.org.au/datasource/dataset-ABS-ABS_CENSUS2011_B07-sa2 |

**Figure 9.** AURIN datasets strictly combinable via spatio-temporal supports.

## 6. Discussion and Conclusion

In this article, we introduced an ontology design pattern together with a number of competency questions that can be used to query for datasets that are comparable, transformable, combinable, and to which summary statistics are applicable. We suggest these fundamental forms of analysis as a first basis for more complex assessments of dataset applicability. The advantage of publishing datasets as linked data in this form is the ability to apply methods to data sets originating from different sources and provided across different infrastructures. We have illustrated this using an operational Australian data portal with datasets of diverse origins.

Assessing the applicability of particular analysis tools and methods, however, requires substantial future work. First, the analysis tasks we identified in this article are not unambiguously linkable to tools yet, as they capture only *necessary* conditions of applicability. For example, in order to generate a meaningful scatter plot, comparability and combinability information need to be integrated: the former needs to be satisfied for all measures of each dimension of the plot, and the latter for each data point in the plot. Second, some requirements for analysis are not reflected in this pattern. For example, a choropleth map requires spatial supports that tessellate space without overlaps. Other plots have similar requirements about orderings for their axes. Third, fundamental analysis procedures and data types are missing in this study, including *aggregation* and *association* of objects in space and time (e.g., associating trajectories with places or part-whole relations). Based on this, combinability of datasets can be inferred based on spatially related supports (which is a common practice of geographical analysis). Lastly, we have only investigated thematic measures over spatial and temporal supports.

In future work, the applicability of the patterns over datasets with different combinations of supports across theme, space and time [33] will be investigated. Furthermore, we will research methods to automatically annotate datasets with our ontology at the

point of creation and to propagate these annotations throughout the analytical chain [9], in order to reduce the need for manual description.

## References


[1] J. Gray, D. Liu, M. Nieto-Santisteban, A. Szalay, D. DeWitt, and G. Heber. Scientific data management in the coming decade. *CoRR*, abs/cs/0502008, 2005.

[2] B. Brodaric and M. Gahegan. Ontology use for semantic e-science. *Semantic Web*, 1(1-2):149–153, 2010.

[3] B. Javadi, M. Tomko, and R. Sinnott. Decentralized orchestration of data-centric workflows in cloud environments. *Future Generation Computer Systems*, 29:1826–1837, 2013.

[4] M. A. Parsons, Ø. Godøy, E. LeDrew, T. F. de Bruin, B. Danis, S. Tomlinson, and D. Carlson. A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, 37(6):555–569, 2011.

[5] I. Widjaja, P. Russo, C. Pettit, R. Sinnott, and M. Tomko. Modeling coordinated multiple views of heterogeneous data cubes for urban visual analytics. *International Journal of Digital Earth*, 8(7):558–578, 2014.

[6] D. Weinberger. The machine that would predict the future. *Scientific American*, 305(6):323–340, 2011.

[7] K. Janowicz, F. van Harmelen, J. A. Hendler, and P. Hitzler. Why the data train needs semantic rails. *AI Magazine*, 36(1):5–14, 2015.

[8] Y. Shoham. Why knowledge representation matters. A personal story: From philosophy to software. *Communications of the ACM*, 59(1), 2016.

[9] S. Scheider, B. Gräler, E. Pebesma, and C. Stasch. Modeling spatiotemporal information generation. *International Journal of Geographical Information Science*, pages 1–29, 2016.

[10] C. Stasch, S. Scheider, E. Pebesma, and W. Kuhn. Meaningful spatial prediction and aggregation. *Environmental Modelling & Software*, 51:149–165, 2014.

[11] P. A. Burrough and R. A. Mcdonnell. *Principles of Geographical Information Systems*. Oxford University Press, USA, 1998.

[12] J. Hughes. Why functional programming matters. *Comput. J.*, 32(2):98–107, 1989.

[13] K. A. Borges, A. H. Laender, and C. A. Davis Jr. Spatial data integrity constraints in object oriented geographic data modeling. In *Proceedings of the 7th ACM international symposium on Advances in geographic information systems*, pages 1–6. ACM, 1999.

[14] H.-J. Lenz and A. Shoshani. Summarizability in OLAP and statistical data bases. In *SSDBM*, pages 132–143. IEEE Computer Society, 1997.

[15] R. Devillers, Y. Bédard, R. Jeansoulin, and B. Moulin. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *Int. J. Geogr. Inf. Sci.*, 21(3):261–282, 2007.

[16] P. Ahonen-Rainio and M.-J. Kraak. Deciding on fitness for use: evaluating the utility of sample maps as an element of geospatial metadata. *Cartography and geographic information science*, 32(2):101–112, 2005.

[17] T. Niemi and M. Niinimäki. Ontologies and summarizability in OLAP. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1349–1353, New York, NY, USA, 2010. ACM.

[18] J. Zhao and O. Hartig. Towards interoperable provenance publication on the linked data web. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012*, volume 937 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[19] B. Sundgren. Classifications of statistical metadata. In *METIS 2008*, 2008.

[20] H. Stuckenschmidt and F. Harmelen. Sharing statistical information. In L. Jain and X. Wu, editors, *Information Sharing on the Semantic Web*, Advanced Information and Knowledge Processing, pages 143–165. Springer Berlin Heidelberg, 2005.

[21] W. Kuhn and A. Ballatore. Designing a language for spatial computing. In F. Bação, M. Y. Santos, and M. Painho, editors, *AGILE 2015 - Geographic Information Science as an Enabler of Smarter Cities and Communities, Lisboa, Portugal, 9-12 June 2015*, Lecture Notes in Geoinformation and Cartography, pages 309–326. Springer, 2015.

[22] B. Hofer. Uses of online geoprocessing technology in analyses and case studies: a systematic analysis of literature. *International Journal of Digital Earth*, 8(11):901–917, 2015.

[23] A. Gangemi and V. Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.
[24] S. Stevens. On the theory of scales of measurement. *SCIENCE*, 103(2684), 1946.
[25] N. Chrisman. Reference systems for measurement. In *Exploring Geographic Information Systems, 2nd Edition*, pages 15–35. Wiley, 2002.
[26] J. Iliffe and R. Lott. *Datums and map projections for remote sensing, GIS, and surveying. Whittles Pub*. CRC Press, Scotland, UK, 2008.
[27] W. Kuhn. Semantic reference systems. *International Journal of Geographical Information Science*, 17(5):405–409, 2003.
[28] L. Breiman et al. The 1991 census adjustment: Undercount or bad data? *Statistical Science*, 9(4):458–475, 1994.
[29] K. Hornsby and M. Egenhofer. Qualitative representation of change. *Spatial Information Theory A Theoretical Basis for GIS*, pages 15–33, 1997.
[30] M. Krötzsch, F. Simancik, and I. Horrocks. A description logic primer. *arXiv preprint arXiv:1201.4089*, 2012.
[31] S. Scheider, K. Janowicz, and W. Kuhn. Grounding geographic categories in the meaningful environment. In *Spatial Information Theory*, pages 69–87. Springer, 2009.
[32] P. Suppes and J. Zinnes. *Handbook of mathematical psychology*, volume 1, chapter Basic Measurement Theory, pages 1–76. Wiley, 1967.
[33] D. F. Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard papers on geographic information systems*, 6:1–17, 1978.
[34] M. Tomko and et al. The design of a flexible web-based analytical platform for urban research. In *ACM SIGSPATIAL '12 Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, 2012.
[35] M. Kalantari, A. Rajabifard, H. Olfat, C. Pettit, and A. Keshtiarast. Automatic spatial metadata systems the case of australian urban research infrastructure network. *Cartography and Geographic Information Science*, to appear 2016.