Applied Ontology 0 (0) 1–38
IOS Press

The interplay between models and observations

Claudio Masolo [a,*], Alessander Botti Benevides [b] and Daniele Porello [c]
[a] *Laboratory for Applied Ontology, ISTC-CNR, Italy*
E-mail: masolo@loa.istc.cnr.it
[b] *Federal University of Espirito Santo, Brazil*
E-mail: alessanderbotti@gmail.com
[c] *Free University of Bozen/Bolzano, Italy*
E-mail: danieleporello@gmail.com

Abstract. We propose a formal framework to examine the relationship between (scientific) *models* and empirical *observations*. To make our analysis precise, models are reduced to first-order theories that represent both terminological knowledge—e.g., the laws that are supposed to regulate the domain under analysis and that allow for explanations, predictions, and simulations—and assertional knowledge—e.g., information about specific entities in the domain of interest. Observations are introduced into the domain of quantification of a distinct first-order theory that describes their nature and their organization and takes track of the way they are experimentally acquired or intentionally elaborated. A model mainly represents the theoretical knowledge or hypotheses on a domain, while the theory of observations mainly represents the empirical knowledge and the given experimental practices. We propose a precise identity criterion for observations and we explore different links between models and observations by assuming a degree of independence between them. By exploiting some techniques developed in the field of social choice theory and judgment aggregation, we sketch some strategies to solve inconsistencies between a given set of observations and the assumed theoretical hypotheses. The solutions of these inconsistencies can impact both the observations—e.g., the theoretical knowledge and the analysis of the way observations are collected or produced may highlight some unreliable sources—and the models—e.g., empirical evidences may invalidate some theoretical laws.

Keywords: Ontology, Epistemology, Scientific theories, Observations, Provenance, Data aggregation

1. Introduction

Conceptual modeling and knowledge representation mainly focus on characterizing how a given domain is structured, i.e., they identify a set of concepts and relations together with the constraints that hold for this domain. In these fields, the structure of the domain is usually intended to reflect the point of view of some experts, by endorsing a moderately subjective stance. Scientific theories have a similar goal but they usually embrace a more objective perspective: the aim is to capture how the world is, to explain and predict what happens in the world. In both cases, the aim is to produce a model (in the sense of conceptual modeling) or a theory (in the sense of model theory) of the domain—in the jargon of description logics, a *terminological box* (TBox). In this representational context, the *data* (the information about specific entities) usually reduce to instantiations of the model, an *assertional box* (ABox). Data sharing can then be achieved by integrating and aligning different models (of the same domain) while data analysis is grounded on logical inferences from the instantiated model (TBox+ABox).

*Corresponding author. E-mail: masolo@loa.istc.cnr.it.

Recently, an enormous volume of data collected by heterogenous sources or resulting from complex analyses is made available on the Internet. The homogeneity of the data taken into account and the understanding of their provenance critically impact the quality, reliability, validity, and trustworthiness of the analyses performed on these data. This is especially relevant for the e-science community and, in particular, for large-scale and distributed collaborative science where data-exchange is sometimes directed to ensure the reproducibility of scientific analyses and experiments. This leads to the need to explicitly represent the nature of data, the way they have been acquired, produced, modified, etc. The (sharing of the) model of the domain is not enough, one needs (to share) a model of the data. Measurements, observations, and analyses have a subjective nature that transcends the conceptual apparatus necessary to represent the domain and require a step towards an operationalist or constructivist stance about data. Furthermore, physics, economics, medicine, biology, psychology, cognitive science, and sociology are deeply founded on (statistical) data analysis and testing. A double subjectivity is present here: the acquisition of *raw data* and their transformation or aggregation into *indicators* or *scores*, i.e., pieces of information that are produced starting from raw data but that are not directly observable, what Bogen and Woodward (1988) calls *phenomena* (see Section 2 for more details).[1] Calibration, measurement procedures, and measurement standards inter-subjectively mediate and control the collection of raw data while mathematical techniques of analysis (e.g., procedures to manage measurement errors) smooth the difficulty to assure reliable analytical results.

Although the relevance and importance of these aspects is widely recognized, the Semantic Web, Applied Ontology, Conceptual Modeling, and Database communities started to pay attention to the nature and the provenance of data only quite recently with the intent to support the sharing and integration of data, so as to enable interoperability for sensors and sensing systems, and to produce detailed descriptions of scientific investigations. The approaches focused on provenance tend to introduce information about the life-cycle of data by means of annotations(-graphs). In this context, the Open Provenance Model[2] and the W3C PROV Data Model[3] result from standardization efforts aimed at establishing a reference provenance model. However, as recognized by part of the Database community, this strategy prevents a uniform approach where the information about the provenance of data is intrinsic to the schema rather than an external annotation. The approaches devoted to a conceptual analysis tend to extend foundational ontologies with notions able to characterize the nature and the provenance of data. Ontologies of observations and measurements mainly developed in the context of Geographical Information Systems (Kuhn (2009); Probst (2008)) explicitly refer to observations and measurement processes. A similar methodology guided the W3C Semantic Sensor Network Incubator group[4] in developing an OWL-2 ontology for describing sensors in terms of measurement processes, observations and deployments (Janowicz and Compton (2010)). The focus here is on the nature of raw data. The DataTop ontology, based on Batrancourt et al. (2010), and the Ontology for Biomedical Investigations, see Brinkman et al. (2010), address the need for the description of biological and clinical investigations, i.e., they are also concerned with the way raw data are elaborated.

Relying on the work done in Benevides and Masolo (2014), we follow this last kind of approaches by introducing what we call *states* into the domain of quantification of a first-order theory and by providing them with a precise identity criterion. States may be seen as (ontological) *facts* or *tropes* that

[1]We will use interchangeably the terms 'observation' and 'datum' to indicate both raw data (direct observations) and phenomena (indirect observations).

[2]`http://openprovenance.org/`

[3]`http://www.w3.org/TR/prov-overview/`

[4]`http://www.w3.org/2005/Incubator/ssn/ssnx/ssn`

capture how the world *is* and are the truth-makers of propositions. This perspective is briefly discussed in Section 6. Our main focus is, however, on (epistemological) *observations*. In this view, states describe how the world *appears*, they support propositions and may be unreliable or conflictual. For instance, a set of observations may violate some laws that are assumed to regulate the domain under consideration. Kuhn (2009), Probst (2008), and Janowicz and Compton (2010) model the provenance of observations by means of primitive relations that link the observations to the sensors that collected them, to the used measurement procedures, etc. Differently, our primitive of *data production* (see Section 7.1) keeps track of the observations used to *produce* (infer, estimate, etc.) more complex observations. In this way, both the provenance of the basic measurements and the one of the results of data-analyses can be uniformly represented. For instance, measurements may be seen as produced by an observation of an object being connected in a certain way with a sensor together with an observation of the position of the perceivable output of the sensor (see Section 7.2). Data production does not guarantee the correctness of the underlying production process. It is then an additional potential source of unreliable or conflicting data.

To try to logically manage these conflicts—by relying, for instance, on the production-chains at the origin of indirect observations—logical inconsistency must be avoided. We propose to separate the model of the data—the theory $\mathcal{T}_{\mathrm{ST}}$ presented in Section 3 and extended in Section 7—from the model of the domain—the theory $\mathcal{T}_{\mathrm{D}}$ presented in Section 4. $\mathcal{T}_{\mathrm{D}}$ formalizes the laws that regulate the kinds of objects taken into account and the way they are interconnected, but it does not directly refer to any specific observation. Vice versa, $\mathcal{T}_{\mathrm{ST}}$ collects all the available information on the data one disposes of, a sort of repository of *observation reports*, but it does not consider the laws that regulate the domain. Section 7.3 sketches some strategies to import data from $\mathcal{T}_{\mathrm{ST}}$ into $\mathrm{ABox}_{\mathrm{D}}$ and to deal with conflicting data. We shall introduce a methodology inspired by voting theory and judgment aggregation List and Puppe (2009) to pinpoint sources of data and to discuss a number of strategies to integrate them. These strategies rely on (*i*) filtering processes that find (on the basis of the information available in $\mathcal{T}_{\mathrm{ST}}$) subsets of observations that guarantee the consistency of $\mathcal{T}_{\mathrm{D}}$ once the correspondent $\mathrm{ABox}_{\mathrm{D}}$-assertions are introduced; (*ii*) revise $\mathcal{T}_{\mathrm{D}}$ to comply with the full set of observations in $\mathcal{T}_{\mathrm{ST}}$; or (*iii*) mix the two precedent strategies.

Before illustrating and discussing the technical details of our formal framework, in Section 2, we clarify our assumptions about the nature of states and observations.

2. Some clarifications about states and observations

Our framework is based on the distinction between *times*, *objects*, and (simple) *states*. Intuitively, *times*, are ordered atomic entities while *objects*—also called *substances*, *endurants*, or *continuants*—are particulars that persist through time by being wholly present at every time they exist, e.g., tables, persons, companies, customers, bits of stuff. A *state* corresponds to the *exemplification* of, or the *classification* under, a property (relation) by one (several) object(s) considered at a given time. E.g., Luca's being $2m$ high now, Luca's being enrolled in the University of Trento now. We leave open what is the exact nature of both properties (relations) and states. In an ontological perspective, see Section 6, states can be seen as real entities that are part of the world, e.g., *facts*, instantiations of *universals* (see Armstrong (1989)) by objects, or *tropes* (see Daly (1997); Maurin (2014)). In an epistemological perspective, see Section 7, states reduce to *observations*, i.e., empirical or cognitive classifications of objects under (qualitative or quantitative) *concepts* (see Margolis and Laurence (1999)) that result from measurement, testing, perceptive, cognitive, or analytical processes.

Despite their epistemological or cognitive nature, our observations do not depend on single observers, they are not private sensations or qualia (as intended by Goodman (1951)). Our observations have a *representational* nature: by abstracting from the observers and from the observing acts, observations are pieces of information that represent the world as being *seen* in a certain way or configuration. It folllows that observations are not necessarily veridical, truthful, or consistent. On the one hand, observations share some properties with *propositions*: (*i*) they are the objects of acts of statement making, perception, or thinking; (*ii*) they can be truthful or not, true or false, i.e., they are truth-bearers; and (*iii*) they abstract from specific languages or agent's minds. On the other hand, they are grounded on empirical practices, cognition, perception, etc., i.e., they can be understood as the result of an abstraction process from cognitive events of classification and categorization.

The classificatory nature of perception has been defended by Matthen (2005). Matthen insists on the fact that sensory awareness has a propositional form, it classifies distal objects as exemplifying certain properties, i.e., the "sensory systems are automatic sorting machines that come into direct contact with environmental objects and sort them into classes according to how they should be treated for the purposes of physical manipulation and investigation" (Matthen, 2005, p.8). Even though Matthen attributes to sensory classification an utilitarian role (to survive and adapt to the environmental niche of the organism), there is a strong commitment neither towards an epistemological view nor towards an ontological one.

In science, outside a purely positivist view, data are not always understood as observation-reports that are the result of direct measurements even though extended by means of given devices.[5] Data may be the result of complex (mental or non-mental) elaborations. As said, Bogen and Woodward (1988) distinguish *data* from *phenomena*: data are the result of direct observations while "[p]henomena are detected through the use of data, but in most cases are not [directly] observable in any interesting sense of that term" (Bogen and Woodward, 1988, p.306). According to Bogen and Woodward, scientific theories often refer to phenomena, i.e., the procedures for data analysis, data reduction, and for handling measurement errors play and important role in science. For instance, the claim that "lead melts at 327 degrees C" is based on the calculation of the mean of various temperature-measurements, "there is no reason why any observed reading must exactly coincide with this mean value" (Bogen and Woodward, 1988, p.308). Both direct and indirect observations—i.e., observations produced starting from (explained or justified in terms of) simpler ones—are explicitly taken into account in Sections 7.1 and 7.2, i.e., in terms of Bogen and Woodward, we include among observations both phenomena and data.

Recently, Soames (2010; 2015) defended the idea of grounding propositions on cognitive acts. According to Soames, propositions are "*repeatable, purely representational, cognitive act-types or operations*" (Soames, 2015, p.16) and to entertain a proposition "*is not to cognize it but to perform it*" (Soames, 2015, p.16). The idea is that agents *cognize* objects as being certain ways, they perform acts of predicating properties of objects, of representing objects as having certain properties. According to Soames, propositions *are* then *act-types*, i.e., specific properties that identify classes of events that have in common the *predicating* of a property to an object. In this perspective, "[i]nstead of deriving the intentionality of agents from independently representational propositions, we must explain the intentionality of propositions in terms of the conceptually prior ability of agents to represent the world in thought and perception" (Soames, 2015, p.14). Without entering the debate on the advantages and drawbacks of this theory,[6] we want just to underline that, as far as we can understand, act-types exist even when they are empty, i.e., even when the proposition is not entertained. While this makes sense in ontological terms, in our epistemological perspective we assume the existence of an observation only when *somebody* commits (through some act or operation) to the classification of an object under a property. The existence of

[5]We do not consider here data like analogical images or records.

[6]See, for instance, King et al. (2014), Caplan (2016), Schiffer (2016), and Soames (2016).

object a and property P seems to be enough to guarantee the existence of the act-type of predicating P to a, but, in our framework, it is not enough to guarantee the existence of an observation. The set of observations is then a subset of the set of propositions, it roughly corresponds to the entertained propositions. On the other hand, the set of observations is a superset of entertained *true* propositions. This means that neither all observations are truthful, nor all the true-propositions are observed. This also implies that our observations, but not the propositions of Soames, can be seen as the result of an abstraction process from acts or operations. However, we explicitly represent neither the classificatory acts the observations abstract from, nor the involved abstraction process. In addition, we assume the observations to be accessible without errors, i.e., to be shared by a community of agents. This sharing process, that can be very complex and can involve (linguistic) interactions between agents, is not taken into account in our framework. This does not prevent a future extension of the theory that explicitly addresses these aspects.

3. The formal framework of states

We have seen that our states (and observations) roughly correspond to entertained propositions. To represent them we observe that the traditional *structured conception* of propositions introduced by Frege and Russell sees propositions as decomposable into objects and properties (or relations), see King (2014) for a review. More precisely, propositions are *represented* as ordered pairs with form $\langle\langle x_1, \ldots, x_n\rangle, R^n\rangle$, where the x_i are objects and R^n is a relation of arity n.[7] Our states are submitted to identity criteria similar to the ones of tuples. The main differences are: (*i*) the existential conditions of states are more restrictive than the ones of pairs—e.g., given a and *Crimson* the pair $\langle a, Crimson\rangle$ exists, but without any classificatory act the corresponding state does not exist; (*ii*) a state can represent the classification under different properties—e.g., the state of a being *Crimson* is also a state of a being *Red* (see Section 3.2.3) while there are two different pairs, namely $\langle a, Crimson\rangle$ and $\langle a, Red\rangle$.

Furthermore, the possibility for a proposition to have different truth-values at different times has been deeply discussed in the philosophical literature, see Brogaard (2012) for a recent discussion. Temporalism claims that the same proposition can change truth-value while eternalism assumes that time 'is part' of the proposition, every proposition is true or false simpliciter. In the perspective of structured propositions, eternalism includes time into the representation of the proposition, e.g., $\langle\langle x_1, \ldots, x_n\rangle, R^n, t\rangle$ while for temporalism time is a sort of modality. We follow the eternalist view and we assume that t refers to the time at which the object has (is seen to have) the property,[8] i.e., t does *not* individuate when the classificatory statement is made (even though this act must exist), but it freezes the object, it establishes when the object is considered for the classification. However, an object can (be seen to) have the same property at different times. Even though this complicates a bit our framework (see Section 3.2.3 and Section 7.3), to minimize our commitment, we allow states to be linked to several times, i.e., they may represent the fact that an object has (is seen to have) the same property at several times.

States are formalized in a first-order setting that extends and systematizes our work published in Benevides and Masolo (2014) and Masolo (2016). Following the usual reification technique, for each tuple $\langle\langle a_1, \ldots, a_n\rangle, R^n, t\rangle$—where the maximal arity of relations is assumed to be finite—we introduce in the domain of quantification a state that (*i*) is linked to the a_is through different relations that aim at capturing the role of the object a_i in the relation R^n; (*ii*) it is an instance of a *unary* predicate that

[7] As observed by Schiffer (2016), this does not mean that propositions *are* ordered pairs, it just implies that a unique proposition composed by $\langle x_1, \ldots, x_n\rangle$ and R^n exists.

[8] In a four-dimensional perspective one would say that it is the *temporal slice* of the object at the time that has the property.

corresponds to R^n, and (*iii*) it is present at t. However, as we will see in Section 3.2.3, our states do not satisfy the standard existential and identity criteria for tuples.

We assume the following notational conventions: (*i*) the ι (inverted iota) is a definite description operator *à la Russell*, i.e., $\psi(\iota x(\phi(x)))$ is a schema for $\exists x(\phi(x) \wedge \forall y(\phi(y) \rightarrow y = x) \wedge \psi(x))$; (*ii*) predicates are noted in uppercase `typewriter` type, e.g., `PERSON`; (*iii*) logical functions and definite descriptions are noted in lowercase **bold** normal type, e.g., **mother_of**;[9] (*iv*) individual constants are noted in lowercase `typewriter` type, e.g., `luca`; (*v*) we write $\mathtt{P}_t x$ instead of $\mathtt{P}(x, t)$ to highlight the time-argument t; (*vi*) x^n is a shortcut for $x_1, \ldots, x_n$; and, finally, (*vii*) the symbol $\triangleq$ is used to introduce syntactic shortcuts.

3.1. *The theory of times and objects*

We assume a very weak theory of time (TM) and objects (OB), a minimal commitment that allows us to elaborate our theory on states. Times can be intended as punctual or extended atomic entities. Even though it is not essential to our goal, one can think *time* as linear and discrete (we will not formally take into account this aspect). Objects (continuants or endurants) may persist and change through time by possibly having, or being classified under, different properties at different times, i.e., they are the subjects of the temporally qualified predications.

The theory $\mathcal{T}_{\mathrm{TM+OB}}$ of times and objects we consider just assures that objects are disjoint from times (a1) but they necessarily exist at least at one time, where $\varepsilon_t x$ stands for "*x exists* at time *t*", see (a2) and (a3). A more realistic model would consider a set of object-kinds and some necessary relations among them. However, the rules and laws that regulate a specific domain constitute a form of contextual knowledge that cannot be easily generalized. $\mathcal{T}_{\mathrm{TM+OB}}$ has to be considered as a very minimal theory shared by both $\mathcal{T}_{\mathrm{D}}$ and $\mathcal{T}_{\mathrm{ST}}$ that can be extended with more complex constraints without undermining the work we will present below. $\mathcal{T}_{\mathrm{TM+OB}} = \mathrm{TBox}_{\mathrm{TM+OB}} \cup \mathrm{ABox}_{\mathrm{TM+OB}}$ is the first-order theory with (non-logical) vocabulary $\mathcal{V}_{\mathrm{TM+OB}} = \{\textsc{tm}, \textsc{ob}, \varepsilon\} \cup \mathcal{C}_{\mathrm{TM+OB}}$, where (*i*) $\mathcal{C}_{\mathrm{TM+OB}}$ is a set of individual constants for objects and times; (*ii*) $\mathrm{TBox}_{\mathrm{TM+OB}} = \{(\mathrm{a1}), (\mathrm{a2}), (\mathrm{a3})\}$; and (*iii*) $\mathrm{ABox}_{\mathrm{TM+OB}}$ is a set of atomic and closed $\mathcal{V}_{\mathrm{TM+OB}}$-formulas. Furthermore, syntactic definitions—to be intended as syntactic shortcuts—introduce some useful notions, e.g., *temporal inclusion* (d1) and *temporal overlap* (d2).

a1 $\textsc{tm}x \rightarrow \neg\textsc{ob}x$

a2 $\varepsilon_t x \rightarrow \textsc{tm}t \wedge \neg\textsc{tm}x$

a3 $\textsc{ob}x \rightarrow \exists t(\varepsilon_t x)$

d1 $x \circledcirc_{\varepsilon} y \triangleq \forall t(\varepsilon_t x \rightarrow \varepsilon_t y)$ (*temporal inclusion*)

d2 $x \otimes_{\varepsilon} y \triangleq \exists t(\varepsilon_t x \wedge \varepsilon_t y)$ (*temporal overlap*)

3.2. *The theory of states*

We specify our theory of states $\mathcal{T}_{\mathrm{ST}}$ by extending $\mathcal{T}_{\mathrm{TM+OB}}$ with additional primitives and constraints. (Simple) states (ST) are disjoint from objects and times (a4) and, like objects, they exist in time (a5).

a4 $\textsc{st}x \rightarrow (\neg\textsc{ob}x \wedge \neg\textsc{tm}x)$

a5 $\textsc{st}x \rightarrow \exists t(\varepsilon_t x)$

[9]In FOL, functions are total, so a function such as **mother_of** can be applied to every entity in the domain of quantification. Due to this aspect, we favor definite descriptions instead of functions.

Note that the existence of a state at a time represents the fact that an object is classified under a property depending on how the object is (seen) at that time, i.e., (according to some observers) the object has the property at that time (similarly for relations). Suppose that somebody, on the basis of a 1:1 photo taken at time t, measures, at time t', the height of an object a to be $1m$. In this case, the state of *'a being* $1m$ *high'* exists at time t, not at the time t' (when the measurement is done). Change can be represented by classifying objects, at different times, under incompatible properties.[10] Properties are here considered as static, they provide a reference system that allows diachronic comparisons. Still objects can be classified under the same property for several times (before changing), i.e., states may exist at different times.

3.2.1. The basic components of states

We have seen that the ontological and the epistemological views disagree on the nature of properties (universals vs. concepts) and of the classification relation (instantiation vs. classification). Despite that, they share the idea that the objects and the properties (relations) are the main components of states.

States are covered (see Section 3.2.2) by a *finite* set $\bar{\mathcal{P}}$ of unary predicates[11]—to be added to the vocabulary $\mathcal{V}_{\mathrm{ST}}$ of $\mathcal{T}_{\mathrm{ST}}$—that represent the *kinds* of states chosen by the user, i.e., intuitively, they collect all the states corresponding to the classification of an object (several objects) under the *same* property (relation). The atemporal primitives $\multimap_{\mathtt{i}}$ hold between objects and states (a6) and have the role of identifying *the i*th object involved in the state, the *i*th *participant* (a7). We assume that the properties and relations predicated of objects, that correspond to state-kinds, have the maximal arity α. This coincides with the number of the $\multimap_{\mathtt{i}}$ primitives that are necessary to distinguish the position (the role) the objects have in the relations. (d3) defines n-ary participation while the general participation abstracts from the position (role) (d4).[12] The participation relations as well as the $\bar{\mathcal{P}}$-predicates are not temporally qualified, i.e., both the participants and the property are constants components of states, states cannot vary their participants or migrate from a kind to a different one. The (temporal) necessity of participants is represented by (a8), where the temporal inclusion $\otimes_{\varepsilon}$ is defined in (d1).

d3 $x^n \multimap s \triangleq \bigwedge_{\mathtt{i}=1}^{n}(x_{\mathtt{i}} \multimap_{\mathtt{i}} s) \wedge \bigwedge_{\mathtt{i}=n+1}^{\alpha} \neg\exists x(x \multimap_{\mathtt{i}} s)$ (*n-ary participation*)

d4 $x \multimap\!\!\!\diamond\, s \triangleq \bigvee_{\mathtt{i}=1}^{\alpha} x \multimap_{\mathtt{i}} s$ (*general participation*)

a6 $x \multimap_{\mathtt{i}} s \rightarrow \mathrm{OB}x \wedge \mathrm{ST}s$

a7 $x \multimap_{\mathtt{i}} s \wedge y \multimap_{\mathtt{i}} s \rightarrow x = y$

a8 $x \multimap_{\mathtt{i}} s \rightarrow s \otimes_{\varepsilon} x$

Some illustrative examples: (*i*) the 'Luca's being enrolled in the University of Trento now' may be represented by (f1); (*ii*) a 'change' in the properties of an object requires at least two states, e.g., (f2); (*iii*) the same object can be characterized by several synchronous states, e.g., (f3).

f1 $\mathtt{luca} \multimap_1 \mathtt{s} \wedge \mathtt{unitn} \multimap_2 \mathtt{s} \wedge \overline{\mathtt{ENROLL}}\mathtt{s} \wedge \varepsilon_{\mathtt{now}}\mathtt{s}$

f2 $\mathtt{luca} \multimap_1 \mathtt{s} \wedge \mathtt{luca} \multimap_1 \mathtt{s}' \wedge \overline{\mathtt{80KG}}\mathtt{s} \wedge \overline{\mathtt{82KG}}\mathtt{s}' \wedge \varepsilon_{\mathtt{t}}\mathtt{s} \wedge \varepsilon_{\mathtt{t}'}\mathtt{s}' \wedge \mathtt{t} \neq \mathtt{t}'$

f3 $\mathtt{luca} \multimap_1 \mathtt{s} \wedge \mathtt{luca} \multimap_1 \mathtt{s}' \wedge \overline{\mathtt{80KG}}\mathtt{s} \wedge \overline{\mathtt{180CM}}\mathtt{s}' \wedge \varepsilon_{\mathtt{t}}\mathtt{s} \wedge \varepsilon_{\mathtt{t}}\mathtt{s}'$

---

[10]As discussed by Haslanger (2003) the temporal qualification of property instantiation is only a possible solution of the problem of representing change, a solution required if one assumes that properties do not change and that objects survive changes and are the direct subject of properties. Alternative solutions are possible if one assumes perdurantism, see Sider (2001). In this case one can read '*a* is *P* at *t*' as '*a*-at-*t* is *P*', i.e., the subject of the property *P* is the temporal slice of *a* at *t*.

[11]We represent the state-kinds by means of predicates with a bar, the motivation will be clear later. Note that $\mathrm{ST} \notin \bar{\mathcal{P}}$.

[12]By having properties and relations in the domain of quantification we could add them to the participants. To avoid the technical problem of property-reification, e.g., manage the instantiation relation as a new primitive, we prefer here to represent them by means of predicates.

First, (*iii*) shows that states only partially characterize their participants. In general a state regards only a specific *aspect* of the participants, e.g., the weight of Luca. Second, one could add, for instance, $\varepsilon_{\mathtt{yesterday}}\mathtt{s}$ in (f1). The possibility for a state to exist at multiple times motivates the distinction between existence and participation, see (a7).[13] Third, even though states may exist at different times, they always correspond to (synchronous) classifications of objects at every single time at which they exist, i.e., they concern properties that can be attributed only by examining the objects under observation during a single time. More complex states could refer to the history of the participants, to the way their properties vary through time. We will not consider this interesting aspect.

3.2.2. Taxonomical structure of state-kinds

The *user* may organize the predicates in $\bar{\mathcal{P}}$ by providing a taxonomy like the one depicted in Figure 1, where a vertical line between a bottom $\bar{\mathtt{P}}$ and a top $\bar{\mathtt{Q}}$ stands for "$\bar{\mathtt{P}}$ is *directly properly subsumed* by $\bar{\mathtt{Q}}$".[14] As a meta-condition, we assume that the provided taxonomy has no cycles, i.e., there are no necessarily co-extensive predicates in $\bar{\mathcal{P}}$. The idea is that every state needs to be classified under at least one state-kind in $\bar{\mathcal{P}}$. In an empirical setting, this kind is not necessarily a leaf of the taxonomy. For instance, not all the measurement devices have the same resolution, one can imagine devices able to distinguish green objects from red ones but not olive objects from emerald ones, thus, one could have green states that are neither olive nor emerald. This is the main role of the taxonomy that is intended to represent the different resolutions of the observations or, in an ontological perspective, a genus-species or determinate-determinables relation, see Sanford (2014).[15]

We assume that the user specifies a state taxonomy via a finite set of SUB-statements, where $\mathtt{SUB}(\bar{\mathtt{Q}},\bar{\mathtt{P}})$ represents that $\bar{\mathtt{Q}}$ is directly properly subsumed by $\bar{\mathtt{P}}$. For each $\mathtt{SUB}(\bar{\mathtt{Q}},\bar{\mathtt{P}})$ statement we add the axiom $\forall x(\bar{\mathtt{Q}}x \rightarrow \bar{\mathtt{P}}x)$ into the TBox of $\mathcal{T}_{\mathrm{ST}}$ (noted $\mathrm{TBox}_{\mathrm{ST}}$). As said before, $\mathrm{TBox}_{\mathrm{ST}}$ contains only these taxonomical axioms. To facilitate the formalization of $\mathcal{T}_{\mathrm{ST}}$, we introduce the following subsets of $\bar{\mathcal{P}}$ that, to be determined, need a simple (pre-)processing of the provided SUB-statements (where $\mathtt{SUB}^{+}$ is the transitive closure of SUB):

- $\bar{\mathcal{P}}_{\mathcal{L}} = \{\bar{\mathtt{P}} \in \bar{\mathcal{P}} \mid$ there are no $\bar{\mathtt{Q}} \in \bar{\mathcal{P}}$ such that $\mathtt{SUB}(\bar{\mathtt{Q}},\bar{\mathtt{P}})\}$; (*leaves* of $\bar{\mathcal{P}}$)
- $\bar{\mathcal{P}}_{\mathcal{R}} = \{\bar{\mathtt{P}} \in \bar{\mathcal{P}} \mid$ there are no $\bar{\mathtt{Q}} \in \bar{\mathcal{P}}$ such that $\mathtt{SUB}(\bar{\mathtt{P}},\bar{\mathtt{Q}})\}$; (*roots* of $\bar{\mathcal{P}}$)
- $\mathtt{ANC}(\bar{\mathtt{P}}) = \{\bar{\mathtt{Q}} \in \bar{\mathcal{P}} \mid \mathtt{SUB}^{+}(\bar{\mathtt{P}},\bar{\mathtt{Q}})\}$. (*ancestors* of $\bar{\mathtt{P}} \in \bar{\mathcal{P}}$)

In Figure 1, $\bar{\mathcal{P}}_{\mathcal{L}} = \{\mathtt{GL\bar{U}ED}, \mathtt{RO\bar{U}ND}, \mathtt{SQ\bar{U}ARE}, \mathtt{CRI\bar{M}SON}, \mathtt{OL\bar{I}VE}, \mathtt{EME\bar{R}ALD}\}$, $\bar{\mathcal{P}}_{\mathcal{R}} = \{\mathtt{PHYS\bar{I}CAL}, \mathtt{CONN\bar{E}CTED}\}$, and $\mathtt{ANC}(\mathtt{R\bar{E}D}) = \{\mathtt{COL\bar{O}RED}, \mathtt{PHYS\bar{I}CAL}\}$. Note that two different leaf-predicates can share instances (but, trivially, no subpredicate). This means that $\bar{\mathcal{P}}$ is not necessarily closed under the logical conjunction. Similarly for disjunction and negation. For instance, assume $\mathtt{SUB}(\bar{\mathtt{P}},\bar{\mathtt{Q}})$. Nothing assures the existence of the complement of $\bar{\mathtt{P}}$ with respect to $\bar{\mathtt{Q}}$; in particular, $\bar{\mathtt{P}}$ can be the only predicate subsumed by $\bar{\mathtt{Q}}$. It is also possible to have two different predicates with a single and shared subpredicate.

(a9) guarantees that all the states of the same root-kind have the same 'arity', i.e., they necessarily have the same number $n > 0$ of participants (where n can be different for different roots). This constraint (partially) characterizes the fact that the state-kinds correspond to properties or relations predicated of

[13] Alternatively one can introduce mereological sums of times, called *periods*, and assume that all states are linked to just one period, the maximal period at which they exist. The first approach is more parsimonious and it explicitly treats times differently from the other arguments of the relation.

[14] The fact that $\mathtt{PHYS\bar{I}CAL}$ and $\mathtt{CONN\bar{E}CTED}$ are subsumed by $\mathtt{ST}$ is directly captured by including $\mathtt{PHYS\bar{I}CAL}$ and $\mathtt{CONN\bar{E}CTED}$ in $\bar{\mathcal{P}}$, the dot-line in the Figure 1 represents this situation.

[15] Taxonomies are a basic way of organizing states-kinds. More sophisticated, topological or geometrical, methods can be considered like in the framework of conceptual spaces, see Gärdenfors (2000).

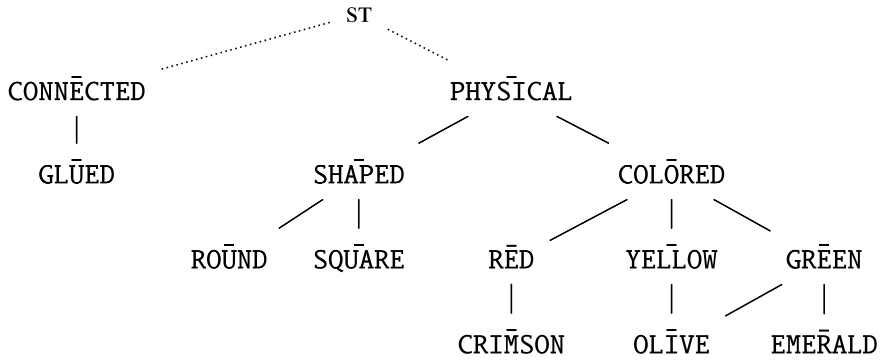

Fig. 1. Example of taxonomical organization of the predicates in $\bar{\mathcal{P}}$.

objects and not to generalizations on states, i.e., pure properties of states. For instance, in Figure 1, $\overline{\mathtt{PHYSICAL}}$ does not stand for 'being a physical state' (that would also include $\overline{\mathtt{CONNECTED}}$-states), but it individuates all the states that correspond to the predication of the property 'having a physical quality', like $\overline{\mathtt{COLORED}}$ corresponds to the property 'having a color quality'.

a9 $\bigwedge_{\bar{\mathtt{P}}\in\bar{\mathcal{P}}_{\mathcal{R}}}(\bar{\mathtt{P}}s \rightarrow \exists x^n(x^n \multimap s))$

Given the possibility to have multi-resolutions systems, we require the user to explicitly specify the resolution of states by assigning to each state s its *minimal kind* (a10), i.e., the kind $\bar{\mathtt{P}}$ such that s is an instance only of $\bar{\mathtt{P}}$ and of the ancestors of $\bar{\mathtt{P}}$ (d5). This trivially assures that states are covered by the state-kinds in $\bar{\mathcal{P}}$. In addition, it guarantees that the theory is able to claim if a state is or is not an instance of any $\bar{\mathtt{P}}\in\bar{\mathcal{P}}$, i.e., the theory is complete with respect to the belonging of states to state-kinds. This is in line with both an ontological and empirical view of states. In the second case, one always knows the resolution of the used device, there are no grey areas of knowledge concerning this aspect. Note that the minimal kind must be unique, a multiple minimal kind would trivially generate an inconsistency.

d5 $\mathtt{min}(s,\bar{\mathtt{P}}) \triangleq \bar{\mathtt{P}}s \wedge \bigwedge_{\bar{\mathtt{Q}}\notin \mathrm{ANC}(\bar{\mathtt{P}})}(\neg\bar{\mathtt{Q}}s)$
a10 $\mathrm{ST}s \rightarrow \bigvee_{\bar{\mathtt{P}}\in\bar{\mathcal{P}}}(\mathtt{min}(s,\bar{\mathtt{P}}))$

3.2.3. *Individuation of states*

The classical identity criterion for pairs with form $\langle\langle a_1,\dots,a_n\rangle, R^n\rangle$ may be captured by (f4) where (d6) defines when two states have the same participants.

d6 $s \bowtie s' \triangleq \bigwedge_{\mathtt{i}=1}^{\alpha} \forall x(x \multimap_{\mathtt{i}} s \leftrightarrow x \multimap_{\mathtt{i}} s')$ (*same participants*)
f4 $\bigvee_{\bar{\mathtt{P}}\in\bar{\mathcal{P}}}(\bar{\mathtt{P}}s \wedge \bar{\mathtt{P}}s') \wedge s \bowtie s' \rightarrow s = s'$

However, (f4) seems too strong for at least two reasons. First, consider the taxonomy in Figure 1 and assume $a \multimap_1 s \wedge a \multimap_1 s' \wedge \overline{\mathtt{ROUND}}s \wedge \overline{\mathtt{OLIVE}}s' \wedge s \otimes_{\varepsilon} s'$. The states s and s' have the same participants, different leaf-kinds, but a common ancestor, namely $\overline{\mathtt{PHYSICAL}}$. Because of that, (f4) would entail $s = s'$ even though s and s' regard two different aspects of a. Second, (f4), as well as the majority of philosophical discussions, do not consider the temporal dimension of states. For instance, assume $a \multimap_1 s \wedge a \multimap_1 s' \wedge \overline{\mathtt{ROUND}}s \wedge \overline{\mathtt{ROUND}}s'$ but $\varepsilon_t s \wedge \varepsilon_{t'} s' \wedge t \neq t'$. Again (f4) would entail $s = s'$ even though t and t' are two disjoint times, i.e., it requires the existence of a unique $\overline{\mathtt{ROUND}}$-state relative to a, what would allow for intermittent states.

Regarding the first problem one needs to better qualify what are the predicates in $\bar{\mathcal{P}}$ that provide the identity criterion for states. The minimal kinds of states are a straightforward possibility: (f4) can then be replaced with (f5). The rationale behind this identity criterion is that, given the *resolution* one disposes of, it is not possible to have different states with the same content. E.g., it is not possible to have two minimally $\mathtt{GR\bar{E}EN}$-states with the same participant while it is possible to have a minimally $\mathtt{GR\bar{E}EN}$-state and a minimally $\mathtt{EME\bar{R}ALD}$-state both about the same object. The difference resides in the fact that they have a different resolution. For the second observation about time, we prefer here to weaken (f5) allowing for temporal disjoint states with the same content (a11), where the temporal overlap $\otimes_\varepsilon$ is defined in (d2).[16] Because all states have a minimal kind, (a11) applies to all the states, i.e., it is a true identity criterion for states. By accepting (a11) we can introduce the shortcut $\mathbf{p}_t x^n$ to denote the unique state that is identified by the time t, the participants x^n and the minimal kind $\bar{\mathtt{P}}$, see (d7). Note that states existing at multiple times can be described in different ways. For instance, in the case of a yellow state existing at both t and t' one has $\mathbf{yellow}_t x = \mathbf{yellow}_{t'} x$.

f5 $\bigvee_{\bar{\mathtt{P}} \in \bar{\mathcal{P}}}(\mathtt{min}(s, \bar{\mathtt{P}}) \wedge \mathtt{min}(s', \bar{\mathtt{P}})) \wedge s \bowtie s' \rightarrow s = s'$

a11 $\bigvee_{\bar{\mathtt{P}} \in \bar{\mathcal{P}}}(\mathtt{min}(s, \bar{\mathtt{P}}) \wedge \mathtt{min}(s', \bar{\mathtt{P}})) \wedge s \bowtie s' \wedge s \otimes_\varepsilon s' \rightarrow s = s'$

d7 $\mathbf{p}_t x^n \triangleq \iota s(\mathtt{min}(s, \bar{\mathtt{P}}) \wedge \varepsilon_t s \wedge x^n \multimap s)$, where n is the arity of $\bar{\mathtt{P}}$ (*state description*)

3.2.4. Complex states

The category of complex states (CST) is the closure of (simple) states under mereological sum. We introduce a primitive *parthood* relation on CST—$x \sqsubseteq y$ stands for "x is part of y"—on the basis of which one can define the classical mereological notions of *proper part* (d8), *overlap* (d9), *sum* (d10), *atom* (d11), and *atomic part* (d12). Sometimes we write $x+y$ to refer to the z such that $z\Sigma xy$. The parthood relation satisfies the axioms for a classical atomic extensional mereology closed under the mereological sum: parthood holds only between complex states (a12), it implies temporal inclusion (a13), it is reflexive (a14), antisymmetric (a15), transitive (a16), and atomic (a17), it satisfies the *strong supplementation principle* (a18), and it is closed under sum (a19). In this theory, complex states are uniquely decomposable into atoms, see Casati and Varzi (1999). (a20) enforces states and mereological atoms to coincide, thus complex states are uniquely decomposable into simple ones.[17]

d8 $x \sqsubset y \triangleq x \sqsubseteq y \wedge x \neq y$ (*proper part*)

d9 $x \between y \triangleq \exists z(z \sqsubseteq x \wedge z \sqsubseteq y)$ (*overlap*)

d10 $x\Sigma y^n \triangleq \forall w(w \between x \leftrightarrow (w \between y_1 \vee \cdots \vee w \between y_n))$ (*sum*)

d11 $\Lambda x \triangleq \mathrm{CST}x \wedge \neg\exists y(y \sqsubset x)$ (*atom*)

d12 $x {}_\Lambda\!\sqsubseteq y \triangleq \Lambda x \wedge x \sqsubseteq y$ (*atomic part*)

a12 $x \sqsubseteq y \rightarrow \mathrm{CST}x \wedge \mathrm{CST}y$

a13 $x \sqsubseteq y \rightarrow x \oslash_\varepsilon y$

a14 $x \sqsubseteq x$

a15 $x \sqsubseteq y \wedge y \sqsubseteq x \rightarrow x = y$

a16 $x \sqsubseteq y \wedge y \sqsubseteq z \rightarrow x \sqsubseteq z$

a17 $\exists y(y {}_\Lambda\!\sqsubseteq x)$

a18 $\neg x \sqsubseteq y \rightarrow \exists z(z \sqsubseteq x \wedge \neg z \between y)$

[16] Stronger constraints can be added to, for instance, enforce the convexity of states.

[17] Complex states correspond to conjunctions of classifications under simple properties not to classifications under complex properties. For instance, one could distinguish $\mathbf{p}_t(a)+\mathbf{q}_t(a)$—a conjunction of classifications—from $[\mathbf{p}\wedge\mathbf{q}]_t(a)$—a classification under a conjunction of properties, see Masolo and Porello (2016).

a19 $\mathrm{CST}x \wedge \mathrm{CST}y \rightarrow \exists s(s\Sigma xy)$

a20 $\Lambda x \leftrightarrow \mathrm{ST}x$

With a slight abuse of notation, (d13) extends the existence ε to complex states. A complex state e is *completely existent* at t, $\hat{\varepsilon}_t e$, if all its (atomic) parts exist at t (d14). At a time t, the participants in a complex state are the participants in its atomic parts that exist at t (d15) (where $x \multimap\!\!\diamond\, s$ is defined in (d4)). Again with a slight abuse of notation, (d16) just abstracts from time. Hence, the participants in the part participate also in the whole. In the following, we will often use the term 'state' to indicate both simple and complex states. In addition, seeing states as facts, these complex states are quite similar to what Galton (2005) call *eventualities*.

d13 $\varepsilon_t e \triangleq \exists s(s \,{}_{\Lambda}{\sqsubseteq}\, e \wedge \varepsilon_t s)$

d14 $\hat{\varepsilon}_t e \triangleq \forall s(s \,{}_{\Lambda}{\sqsubseteq}\, e \rightarrow \varepsilon_t s)$

d15 $x \multimap\!\!\diamond_t\, e \triangleq \exists s(s \,{}_{\Lambda}{\sqsubseteq}\, e \wedge x \multimap\!\!\diamond\, s \wedge \varepsilon_t s)$

d16 $x \multimap\!\!\diamond\, e \triangleq \exists t(x \multimap s \wedge \varepsilon_t s)$

3.2.5. The theory $\mathcal{T}_{\mathrm{ST}}$

We are now in the condition of precisely defining our *theory of states*:

– $\mathcal{V}_{\mathrm{ST}} = \mathcal{V}_{\mathbf{TM+OB}} \cup \{\mathrm{ST}, \mathrm{CST}, \multimap_1, \ldots, \multimap_\alpha, \sqsubseteq\} \cup \bar{\mathcal{P}} \cup \mathcal{C}_{\mathrm{ST}}$
  where $\mathcal{C}_{\mathrm{ST}}$ is a set of individual constants for simple and complex states, and

– $\mathcal{T}_{\mathrm{ST}} = \mathcal{T}_{\mathbf{TM+OB}} \cup \mathrm{TBox}_{\mathrm{ST}} \cup \mathrm{ABox}_{\mathrm{ST}}$, where
  $\mathrm{TBox}_{\mathrm{ST}}$ ={(a4), (a5), (a6), (a7), (a8), (a9), (a10), (a11), (a12), (a13), (a17), (a18), (a19), (a20)} ∪ {all the material implications that correspond to SUB-statements}
  and $\mathrm{ABox}_{\mathrm{ST}}$ is a set of atomic and closed $\mathcal{V}_{\mathrm{ST}}$-formulas that concern at least one symbol in $\mathcal{V}_{\mathrm{ST}} \backslash \mathcal{V}_{\mathbf{TM+OB}}$.

3.2.6. A short comparison

States vs. tuples. First, states but not tuples are linked with times. Second, states are only partially identified by their participants: states of different kinds but with the same participants are possible, i.e., $x^n \multimap s \wedge x^n \multimap s' \wedge \bar{\mathtt{P1}}s \wedge \bar{\mathtt{P2}}s' \wedge s \neq s'$ is consistent. Consider now the tuples with form $\langle\langle a_1, \ldots, a_n\rangle, R^n, t\rangle$ discussed in Section 2 in the context of the representation of propositions. First, if $t \neq t'$, then $\langle\langle a_1, \ldots, a_n\rangle, R^n, t\rangle$ is necessarily different from $\langle\langle a_1, \ldots, a_n\rangle, R^n, t'\rangle$. Differently, states may exist at different times. Second, a state with minimal kind $\bar{\mathtt{P}}$ is an instance also of all the state-kinds in $\mathtt{ANC}(\bar{\mathtt{P}})$. Vice versa, if $R \neq S$, then $\langle\langle a_1, \ldots, a_n\rangle, R^n, t\rangle$ is different from $\langle\langle a_1, \ldots, a_n\rangle, S^n, t\rangle$. Third, given n objects $a_1, \ldots, a_n$, an n-ary relation R^n, and a time t, the tuple $\langle\langle a_1, \ldots, a_n\rangle, R^n, t\rangle$ always exists while the state exists only when the objects $a_1, \ldots, a_n$ are (appear to be) in the relation R^n at t.

States vs. Kim's events. Intuitively our states satisfy the Kim's *existence condition* for events: "event $[x, P, t]$ exists just in case substance x has property P at time t", see Kim (1976).[18] However, the Kim's *identity condition*—"$[x, P, t] = [y, Q, t']$ just in case $x = y$, $P = Q$, and $t = t'$" does not hold in general for our states. First, our states can be linked with several times. Second, the same state can have several (leaf- or non leaf-) kinds. For instance, when $\mathtt{SUB}^+(\bar{\mathtt{CRIMSON}}, \bar{\mathtt{RED}})$, $\bar{\mathtt{CRIMSON}}s \wedge \bar{\mathtt{RED}}s$ is consistent. What distinguishes Kim's events from the tuples with form $\langle\langle a_1, \ldots, a_n\rangle, R^n, t\rangle$ is the existence condition.

[18]While Kim's properties are in the domain of quantification, we represent them through $\bar{\mathcal{P}}$-predicates.

States vs. RTLs' fluents or states. In *reified temporal logics* (RTLs), e.g., the *situation calculus* (see Reiter (2001)) or the *event calculus* (see Mueller (2006)), *fluents* and *states* are represented as total functions applied to objects, e.g., respectively, **tired**(john) and **tired**(john, t). States were introduced to account for the criticisms originally raised in Galton (1991) regarding the fact that fluents are reifications of types instead of tokens. States are then represented as total functions where at least one of the arguments is a time point or an interval. These total functions correspond to (some of) our state-kinds. In RTLs, the totality guarantees that the existence of luca entails the one of **tired**(luca, t) for each time t. Vice versa, in our framework, the existence of luca and the inclusion of the predicate $\mathtt{TI\bar{R}ED}$ in $\bar{\mathcal{P}}$ implies the existence neither of a $\mathtt{TI\bar{R}ED}$-state with the participant luca, nor, more generally, of a state involving luca. Again RTLs' states seem closer to reifications of propositions than to facts or observations. This claim is supported by the fact that RTLs' states are in the domain of quantification even if they do not hold; e.g., the state **tired**(luca, t) is in the domain of quantification even though we have ¬HOLDS(**tired**(luca, t)).[19] Actually, for every time and object in the domain, there will be a **tired**-state independently of its holding. This seems to contradict the Kim's *existence condition*.

States vs. DB-tables vs. Neo-Davidsonian states. In Section 3.2.1 we have seen that time is not considered as a participant ($\multimap$) of a state mainly because states may be ε-linked with several times. One could apply this abstraction process also to participation. More strongly, one could replace the $\multimap_{\mathtt{i}}$ and ε primitives with a single (several) participation relation(s) PC that does not satisfy the critical constraints (a7), (a9), and (a11). In this case the number of participants of states of a given kind is not explicitly represented in the model.[20] This move is close to the one followed by some Neo-Davidsonian approaches that, similarly to what done for events, introduce (optional) *thematic roles* (e.g., subject and object) also for states (see, for instance, Parsons (1990) and Landman (2000)). An intermediate position consists in establishing a maximal arity for each state-kind. This is particularly interesting in the case of databases where a table with fixed columns could have some empty cells, i.e., one has partial information. Halpin and Morgan (2008) analyze the case of optional roles in the context of students that score a rate for a given subject, claiming that in this way "it is possible to record the fact that a person enrolls in a subject before knowing what rating the student gets for that subject" (Halpin and Morgan, 2008, p.708), i.e., *nesting*[21] with *optional* roles (instead of mandatory ones) allows to encode partial information, tables with incomplete entries. The notion of *partial reification* introduced in Olivé (2007) is quite similar to nesting with optional roles, i.e., the user is free to establish if the roles are mandatory or optional.

This is a flexible approach to represent partial information. However, first, it really complicates the establishment of an identity criterion for states (and events), as proved by all the failed attempts of providing a clear identity criterion for events in the Davidsonian approaches. Second, from a data-centric viewpoint, it is difficult to understand what a partial or incomplete datum is. One can have partial information about an object a. For instance, one knows the color of a (at a given time) but not its shape, or one knows that a is red without knowing the exact shade of color. Similarly, while in $\mathtt{ENROLL}_t xy$, where x is a person and y is a subject, the score is unknown, in $\mathtt{ENROLL}_t xyz$ the score z is explicitly taken

[19] HOLDS($\mathbf{f}x^n$, t) means that the fluent $\mathbf{f}x^n$ holds (is true) at t. Concerning states, as the state-function **f** has already a temporal argument, there is no need to temporally qualify HOLDS, e.g., HOLDS(**tired**(luca, t)). Some approaches consider a time interval or a pair of time points as arguments of HOLDS, see Vila and Reichgelt (1996).

[20] One could also distinguish mandatory from optional participants.

[21] Nesting is a mechanism to reduce, for instance, a ternary relation where a student scores a rate for a given subject to a binary enrolling relation defined between students and subjects with a *mandatory* rating role.

into account. However, according to us, $\mathtt{ENROLL}_t xy$ and $\mathtt{ENROLL}_t xyz$, or $\mathtt{RED}_t x$ and $\mathtt{SCARLET}_t x$, are all full-fledge data even though they have a different resolution or precision.[22]

States vs. individual qualities and relators. The *individual qualities* in the DOLCE ontology (see Masolo et al. (2003)) are more abstract than our states. For instance, 'the color of the rose r', the individual quality $\mathtt{c_r}$ specific to r, can change from scarlet to crimson, i.e., it survives movements in the color-space. Colors are here in the domain of quantification and the fact that r changes from scarlet to crimson is represented by the shift from the scarlet position of $\mathtt{c_r}$ to the crimson one in the color-space. In our framework, individual qualities can be seen as complex states, i.e., similarly to classical trope theory, change reduces to state (relative to the same object and dimension) substitution.

Guarino and Guizzardi (2015) extend the idea of individual qualities to time-varying *relationships* that are represented as *relators*, i.e., entities constituted by *relational* qualities that have some time-varying attributes, e.g., the attribute 'credits' for the enrollment of a student in an university. Again, we could modify our framework to consider optional participants but, from a data-centric perspective, this seems a worthless complication. For instance, the change of the score relative to a course-subject may be represented as in (f6) by relying on two different states (with the same person and course but different scores, namely r_1 and r_2) where no optional participants are involved. Relationships with time-varying arguments can again be reconstructed through complex states, i.e., by summing up states that share some participants (i.e., same person and course in the previous example).

f6 $\bar{\mathtt{ENROLL}} s_1 \wedge \mathtt{luca} \multimap_1 s_1 \wedge \mathtt{math} \multimap_2 s_1 \wedge r_1 \multimap_3 s_1 \wedge \varepsilon_{t_1} s_1 \wedge$
$\bar{\mathtt{ENROLL}} s_2 \wedge \mathtt{luca} \multimap_1 s_2 \wedge \mathtt{math} \multimap_2 s_2 \wedge r_2 \multimap_3 s_2 \wedge \varepsilon_{t_2} s_2$

4. The model of the domain of interest

We introduce the theory $\mathcal{T}_{\mathrm{D}}$, which models (the rules regulating) the domain of interest. Technically, $\mathcal{T}_{\mathrm{D}}$ extends $\mathcal{T}_{\mathrm{TM+OB}}$ with a *finite* set $\mathcal{P}$ of temporally qualified (contingent) predicates that apply to objects, i.e., $\mathcal{P}$-atomic formulas have the form $\mathtt{P_t c}^n$, where $\mathtt{P} \in \mathcal{P}$ has arity n+1, t is an individual constant of kind TM, and the $\mathtt{c}_i$ are individual constants of kind OB such that they are in the relation ε with t (a21).[23]

a21 $\bigwedge_{\mathtt{P} \in \mathcal{P}} (\mathtt{P}_t x^n \rightarrow \mathrm{OB} x_1 \wedge \varepsilon_t x_1 \wedge \ldots \wedge \mathrm{OB} x_n \wedge \varepsilon_t x_n)$ (where n+1 is the arity of P)

The predicates $\mathtt{P} \in \mathcal{P}$ are *temporally contingent* in the sense that $\mathtt{P}_t x \wedge \varepsilon_{t'} x \wedge t \neq t' \rightarrow \mathtt{P}_{t'} x$ does not necessarily hold. *Kinds* like 'being a person' or 'being an electron' are then excluded from $\mathcal{P}$. If needed, these kinds can be introduced in $\mathcal{T}_{\mathrm{TM+OB}}$ (without any temporal argument).[24] $\mathcal{T}_{\mathrm{D}}$ typically contains the terminological axioms (the $\mathrm{TBox_D}$) that characterize the way the predicates in $\mathcal{P}$ are interlinked. Similarly to the case of states, we assume that the $\mathcal{P}$-predicates are taxonomically structured by the user via a finite set of SUB(P, Q) statements that are translated into axioms with form $\mathtt{P}_t x^n \rightarrow \mathtt{Q}_t x^n$ where

[22]Note that, intuitively, $\mathtt{ENROLL}_t xyz \rightarrow \mathtt{ENROLL}_t xy$. However this generalization is not captured by the SUB relation.

[23]See Varzi (2003) for a discussion about modal alternatives to represent the temporal qualification of propositions.

[24]Some philosophers support the idea that the truth-makers of *kinds* are their instances. For example, for Armstrong (1997), the truth-makers of 'being a person' are the persons themselves. However, if necessary, our framework can be easily adapted to accept temporally necessary predicates in $\mathcal{T}_{\mathrm{D}}$. In addition, $\varepsilon \notin \mathcal{P}$ because finding the counterpart of existential propositions in terms of states would cause an infinite regress (in their turn, existence-states exist) that is unmanageable in a standard representational setting. Indeed, the existence of truth-makers of existential propositions has also been debated in philosophy. Armstrong (2004) himself claims that it is a mistake to recognize states of affairs with the form "*a*'s existence", because this will turn existence into a property of *a*.

both P and Q have arity n+1. Again, these SUB-statements are intended to represent generalization or determinable/determinate links, e.g., the link between SCARLET and RED. Theoretical or empirical laws (or causal correlations) have a different nature. For instance, to capture the situation where, in a given domain, all the round-objects are necessarily red, one can introduce the axiom $\mathtt{ROUND}_t x \rightarrow \mathtt{RED}_t x$ into TBox_D, but this is not a pure taxonomical statement representable by means of $\mathtt{SUB}^+$.[25]

We will see that, in a scientific perspective, it is interesting to distinguish the initial ABox_D—intuitively, the set of data directly grounded on experimental outputs, let us say—from its deductive closure under $\mathcal{T}_\text{D}$—the data that can be inferred from the experimental data by using the theoretical or empirical laws and that, in their turn, need to be experimentally evaluated. This distinction allows to highlight discrepancies between the theory and the experimental data one disposes of.

The theory $\mathcal{T}_\text{D}$ of *temporally qualified predicates* is defined as in the following:

- $\mathcal{V}_\text{D} = \mathcal{V}_\mathbf{TM+OB} \cup \mathcal{P}$
  i.e., there are no new individual constants;
- $\mathcal{T}_\text{D} = \mathcal{T}_\mathbf{TM+OB} \cup \text{TBox}_\text{D} \cup \text{ABox}_\text{D}$, where
  TBox_D is a set of axioms that link the predicates in $\mathcal{P}$ (and, eventually, with the predicates for kinds of objects in $\mathcal{V}_\mathbf{TM+OB}$) including (a21) and the axioms that correspond to SUB-statements, and ABox_D is a set of atomic and closed $\mathcal{P}$-formulas ($\mathcal{T}_\text{D}$ does not introduce additional $\mathcal{V}_\mathbf{TM+OB}$-assertions).

Note that $\mathcal{T}_\text{D}$ does not introduce any new (kind of) object with respect to $\mathcal{T}_\mathbf{TM+OB}$. This seems to go against the assumption, largely shared among philosophers of science, that theories may refer to non directly observable entities, e.g., electrons, bosons, etc.[26]

Vice versa, some (kinds of) observable objects could not be directly considered by theories. In particular, some raw data used to produce complex data about theoretical entities can be hidden at the level of the theory. This shows that, in principle, the objects involved in $\mathcal{T}_\text{D}$ and the ones involved in $\mathcal{T}_\text{ST}$ may properly overlap. However, in Section 5, we will see that our framework does not assume a sharp boundary between theoretical and experimental objects and laws. Some empirical laws can be represented at the level of $\mathcal{T}_\text{D}$ while some theoretical laws can be partially captured at the level of $\mathcal{T}_\text{ST}$ by the primitive of data production. We prefer then to have a unique vocabulary for objects without clearly distinguishing directly observable objects from indirectly observable ones. Indeed, both $\mathcal{T}_\text{ST}$ and $\mathcal{T}_\text{D}$ can involve only a subclass of the objects in $\mathcal{T}_\mathbf{TM+OB}$, i.e., it is possible to have objects that do not participate in any state in $\mathcal{T}_\text{ST}$ or that are not the subject of any predication in $\mathcal{T}_\text{D}$.

5. Linking models with states

We analyze some relations between the model of the domain, the theory $\mathcal{T}_\text{D}$, and the model of data, the theory $\mathcal{T}_\text{ST}$. The link between theories and observations has been deeply debated in philosophy of science. The so called *Syntactic View*, mainly developed by Carnap (1956), conceives scientific theories as set of sentences in predicate logic with the (non-logical) vocabulary partitioned into theoretical and

[25]In an empirical perspective, round-data and red-data are collected in completely different ways and with different instruments. Vice versa, scarlet-data are just classified also as red-data without any additional empirical activity.

[26]Events or changes are interesting cases of entities that are not *synchronously* observable: to observe a change one needs at least two diachronic observations. The existence of a change (and its properties) is then grounded on patterns of diachronic observations. One could also undertake a radical position where all the objects are only indirectly observable, i.e., their existence is determined on the basis of patterns of observations, see Bottazzi et al. (2012).

observational terms. Empirical laws regulate observations, they help "to explain a fact that has been observed and to predict a fact not yet observed" (Carnap, 1966, p.229), while theoretical laws have a sort of hypothetical role, they help "to explain empirical laws already formulated, and to permit the derivation of new empirical laws" (Carnap, 1966, p.229). Theoretical terms are connected with observational terms via *correspondence rules*, e.g., as illustrated by Carnap: "If there is an electromagnetic oscillation of a specified frequency, then there is a visible greenish-blue color of a certain hue". Only the existence of correspondence rules allows to derive empirical laws that, in their turn, allow to empirically evaluate theoretical laws. In this perspective, there is not necessarily a realistic stance on theories. Theories can be seen as "language tools for organizing the observational phenomena of experience into some sort of pattern that will function efficiently in predicting new observables. The theoretical terms are convenient symbols. The postulates containing them are adopted because they are useful, not because they are "true"" (Carnap, 1966, p.254). The *Semantic View*, mainly developed by Suppes (2002) and van Fraassen (1989), criticizes logical empiricists principally for the linguistic nature of scientific theories. The Semantic View refuses the use of predicate logic and commits to set-theory, it describes both theories and observations as classes of mathematical structures that abstract from the formal language used to reconstruct them. Theories and observations are then linked by means of embeddings that play the role of the correspondence rules of Carnap.

Without entering this complex debate,[27] our first-order framework is quite close to the Syntactic View but some important differences exist. First, the Syntactic View assumes a sharp separation between theoretical and empirical (observational) terms and laws that, in their turn, are distinct from correspondence rules. To address the possibility to have partial information about the data that may be produced through disparate kinds of processes or collected by heterogeneous sources, we do not assume this sharp separation. Both the axioms in $\mathcal{T}_{\mathrm{D}}$ and the data production links in $\mathcal{T}_{\mathrm{ST}}$ (see Section 7.1) may represent empirical, theoretical, experimental laws, or correspondence rules. In addition, to simplify the framework, we assume the link between $\mathcal{T}_{\mathrm{ST}}$-observations and $\mathrm{ABox}_{\mathrm{D}}$-assertions to be direct, i.e., the imported $\mathrm{ABox}_{\mathrm{D}}$-assertions just mirror a (possibly produced) observation (see below and Section 7 for the formalization).

This does not mean that our framework is incompatible with the Syntactic View. For instance, one may assume that $\mathcal{T}_{\mathrm{D}}$ contains the theoretical laws, $\mathcal{T}_{\mathrm{ST}}$ contains the empirical laws, while the correspondence rules are represented by data production chains (in $\mathcal{T}_{\mathrm{ST}}$) together with direct links between (produced) observations and $\mathrm{ABox}_{\mathrm{D}}$-assertions. Second, in the Syntactic View, the data and the laws must be consistent. Vice versa, $\mathcal{T}_{\mathrm{ST}}$ is intended to collect all the available information about the data even when they conflict with some empirical or theoretical law. This is a challenging scenario that however allows to investigate some techniques to solve these conflicts on the ground of the reliability of data and laws, avoiding then to discard the whole set of data, see Section 7.3 for a preliminary analysis.

In Section 5.1 we analyze how the propositions in $\mathrm{ABox}_{\mathrm{D}}$ can be *grounded* in (supported by) the states in $\mathcal{T}_{\mathrm{ST}}$, i.e., we start from the theory $\mathcal{T}_{\mathrm{D}}$ and we study what atomic propositions have a corresponding state in $\mathcal{T}_{\mathrm{ST}}$. In Section 5.2 we take into account the reverse link, i.e., we start from the states in $\mathcal{T}_{\mathrm{ST}}$ and we study what states are *covered* by $\mathcal{T}_{\mathrm{D}}$, what states have an assertional counterpart in $\mathcal{T}_{\mathrm{D}}$.

We have seen that $\mathcal{T}_{\mathrm{D}}$ and $\mathcal{T}_{\mathrm{ST}}$ may use partially overlapping sets of individual constants for objects (introduced in $\mathcal{T}_{\mathrm{TM+OB}}$). They could also consider predicates that do not have a direct correspondence. For instance, (*i*) $\mathcal{T}_{\mathrm{D}}$ may talk of specific relations between electrons that could be nonobservable (in

[27]The interested readear can refer to Lutz (2015) for a recent overview. Note that, as showed by Lutz (2014), the syntactic and the semantic methods are not in opposition, they represent two perspectives on theories that can co-exist and can support each other. In particular, it seems that an improved version of the Syntactic View is somewhat more general than the Semantic View, see Halvorson (2013) and Lutz (2015).

the sense of Carnap); (*ii*) the concept of *density* may be introduced in $\mathcal{T}_{\text{D}}$ while only volume-states and mass-states are present in $\mathcal{T}_{\text{ST}}$; (*iii*) $\mathcal{T}_{\text{ST}}$ may consider some state-kinds (and objects) that are not relevant for the domain of interest of $\mathcal{T}_{\text{D}}$; or (*iv*) $\mathcal{T}_{\text{D}}$ and $\mathcal{T}_{\text{ST}}$ may have different resolutions, for instance, the available devices could not resolve all the properties considered in $\mathcal{T}_{\text{D}}$, or, vice versa, $\mathcal{T}_{\text{D}}$ could abstract from some detailed observations.

The example (*iv*) about resolution allows us to clarify what we intend with *direct* correspondence. Assume $\texttt{RED} \in \mathcal{P}$ and $\overline{\texttt{CRIMSON}} \in \bar{\mathcal{P}}$. Intuitively, the association of $\texttt{RED}$ with $\overline{\texttt{CRIMSON}}$ makes sense for the grounding but not for the covering: $\texttt{RED}$-propositions are supported by $\overline{\texttt{CRIMSON}}$-states, but $\overline{\texttt{CRIMSON}}$-states are not covered by $\texttt{RED}$-propositions because $\texttt{RED}$-propositions could be grounded in $\overline{\texttt{SCARLET}}$-states. Dually, $\overline{\texttt{RED}}$-states are intuitively covered by $\texttt{CRIMSON}$-propositions, but $\texttt{CRIMSON}$-propositions are not grounded in (too general) $\overline{\texttt{RED}}$-states. To capture this asymmetry, two different mappings are required: one for groundedness from $\mathcal{P}$ to $\bar{\mathcal{P}}$ and one—that, in general, is not the inverse mapping—for covering from $\bar{\mathcal{P}}$ to $\mathcal{P}$. Here we are interested in studying the groundedness of $\mathcal{T}_{\text{D}}$ and the covering of $\mathcal{T}_{\text{ST}}$ once a unique correspondence between predicates has been established, i.e., we want to know if a given correspondence guarantees both the groundedness and the covering. We then require that the user provides (*i*) the subsets of predicates $\mathcal{P}^* \subseteq \mathcal{P}$ and $\bar{\mathcal{P}}^* \subseteq \bar{\mathcal{P}}$ with a direct correspondent together with (*ii*) a $\texttt{SUB}^+$-preserving[28] bijection $\gamma : \mathcal{P}^* \to \bar{\mathcal{P}}^*$, i.e., γ is as one-to-one relation between the predicates in $\mathcal{P}^*$ and the ones in $\bar{\mathcal{P}}^*$ such that $\texttt{SUB}^+(\gamma(\texttt{P}), \gamma(\texttt{Q}))$ if and only if $\texttt{SUB}^+(\texttt{P}, \texttt{Q})$. The predicate associated with $\texttt{P} \in \mathcal{P}^*$ is noted $\bar{\texttt{P}} \in \bar{\mathcal{P}}^*$, i.e., $\bar{\texttt{P}} = \gamma(\texttt{P})$. At the end, we assume that the user provides as input the tuple $\langle \mathcal{T}_{\text{D}}, \mathcal{P}^*, \mathcal{T}_{\text{ST}}, \bar{\mathcal{P}}^*, \gamma \rangle$ on the basis of which groundedness and covering can be studied.

The correspondence is direct in the sense that the grounding of $\texttt{P}$-propositions depends on the existence of $\bar{\texttt{P}}$-states (see Section 5.1) and the covering of $\bar{\texttt{P}}$-states depends on the presence of $\texttt{P}$-propositions (see Section 5.2). However, on the one hand, states may encapsulate or explicitly capture—via the data production primitive—dependencies on other states. This allows to ground ABox_{D}-assertions on states produced (by someone) through (not necessarily truthful) cognitive or mathematical operations. On the other hand, sometimes, the assumed bijection requires an explicit representation of an abstraction process. Consider the previous example of theories with different resolution. Intuitively, neither $\gamma(\texttt{RED}) = \overline{\texttt{CRIMSON}}$ nor $\gamma(\texttt{CRIMSON}) = \overline{\texttt{RED}}$ make sense because, in the first case, $\overline{\texttt{CRIMSON}}$-states are not covered by $\texttt{RED}$-propositions and, in the second case, $\texttt{CRIMSON}$-propositions are not grounded in $\overline{\texttt{RED}}$-states. To represent the fact that 'being red' is coarser than 'being crimson', the user needs to add a predicate and a $\texttt{SUB}$-assertion: in the first case, it is enough to introduce $\overline{\texttt{RED}} \in \bar{\mathcal{P}}^*$ and $\texttt{SUB}(\overline{\texttt{CRIMSON}}, \overline{\texttt{RED}})$; in the second case, $\texttt{RED} \in \mathcal{P}^*$ and $\texttt{SUB}(\texttt{CRIMSON}, \texttt{RED})$. The claim $\gamma(\texttt{RED}) = \overline{\texttt{RED}}$ establishes a direct correspondence that maintains the previous intuitions about groundedness and covering: $\texttt{RED}$-propositions are still grounded in $\overline{\texttt{CRIMSON}}$-states—because $\overline{\texttt{CRIMSON}}$-states are $\overline{\texttt{RED}}$-states—and $\overline{\texttt{RED}}$-states are still covered by $\texttt{CRIMSON}$-propositions—because $\texttt{CRIMSON}$-propositions are $\texttt{RED}$-propositions—but not the vice versa—not all the $\overline{\texttt{RED}}$-states ($\texttt{RED}$-propositions) are $\overline{\texttt{CRIMSON}}$-states ($\texttt{CRIMSON}$-propositions).

Finally, note that the deductive closure of $\mathcal{T}_{\text{D}}$ and $\mathcal{T}_{\text{ST}}$ could introduce new propositions in the ABoxes of, respectively, $\mathcal{T}_{\text{D}}$ and $\mathcal{T}_{\text{ST}}$ whose groundedness and covering is not always assured by the groundedness and covering of the initial ABoxes (ABox_{D} and ABox_{ST}). We will illustrate the importance of this distinction in the following. For the moment it is important to clarify that we understand theories as usually done in knowledge representation. While logicians usually reduce a theory to a set of sentences in a given language, in knowledge representation a theory is usually intended as a set of axioms (in a given language) explicitly introduced by the developer, i.e., there is a distinction between the theory, the

[28]Remember that $\texttt{SUB}^+$ is the transitive closure of $\texttt{SUB}$.

set of axioms, and its deductive closure, the set of formulas deducible from these axioms. Our theories (our ABoxes and TBoxes) always refer to the lists of propositions introduced by the user and not to their deductive closure.

5.1. Groundedness

The general idea is that the proposition $\mathtt{P_t c}^n \in \mathrm{ABox_D}$ is grounded on a $\bar{\mathtt{P}}$-state that exists at time $\mathtt{t}$ and has participants $\mathtt{c}^n$. Consider the tuple $\langle \mathcal{T}_D, \mathcal{P}^*, \mathcal{T}_{ST}, \bar{\mathcal{P}}^*, \gamma \rangle$ provided by the user. Definition 1 introduces the notion of *(simple) groundedness* for a subset of the $\mathcal{P}^*$-propositions in $\mathrm{ABox_D}$. We say that $\mathcal{T}_D$ is simply grounded on $\mathcal{T}_{ST}$ if and only if the set of all the $\mathcal{P}^*$-propositions in $\mathrm{ABox_D}$—i.e., the set of all the propositions that involve predicates on which γ is defined—is grounded in $\mathcal{T}_{ST}$.

Definition 1. *(Simple Groundedness) $A \subseteq \mathrm{ABox_D}$ of $\mathcal{P}^*$-propositions is (simply)* grounded *in $\mathcal{T}_{ST}$ if and only if for every $\mathtt{P_t c}^n \in A$ there exists a $\mathtt{s} \in \mathcal{C}_{ST}$ such that $\mathcal{T}_{ST} \vdash \bar{\mathtt{P}}\mathtt{s} \wedge \varepsilon_\mathtt{t}\mathtt{s} \wedge \mathtt{c}^n \multimap \mathtt{s}$.*

Definition 1 considers the deductive closure of $\mathrm{ABox_{ST}}$, i.e., the taxonomical structure of states is used. This implies that the same state can ground several propositions. Consider the taxonomy in Figure 1. The state **crimson**$_\mathtt{t}\mathtt{c}$ can ground both $\mathtt{CRIMSON_t c}$ and $\mathtt{RED_t c}$ because $\mathtt{SUB}(\overline{\mathtt{CRIMSON}}, \overline{\mathtt{RED}})$. Furthermore, a proposition may have several grounding states. Consider, for instance, **crimson**$_\mathtt{t}\mathtt{c}$ and **red**$_\mathtt{t}\mathtt{c}$. These two states are different (see (d7) and (a11)) but both of them ground $\mathtt{RED_t c}$. An ontological perspective on states could motivate a criticism about the grounding of propositions on multiple facts, the idea that a proposition has multiple truth-makers. Vice versa, in an epistemological context, the previous example captures the situation where one disposes of two data, with different resolution, both supporting $\mathtt{RED_t c}$. Note that if $\mathtt{CRIMSON_t c} \in A$ the $\mathtt{SUB}^+$-preserving assumption on the correspondence γ avoids clashes among the taxonomical information in $\mathcal{T}_D$ and the one in $\mathcal{T}_{ST}$.

Definition 2 introduces the notion of *strong groundedness* that assures that all the $\mathcal{P}^*$-propositions one can infer in $\mathcal{T}_D$ (more precisely, in $\mathcal{T}_D \setminus \mathrm{ABox_D}$) from the propositions in A are grounded (where $cl(X)$ identifies the deductive closure of the set of formulas X). We say that $\mathcal{T}_D$ is strongly grounded in $\mathcal{T}_{ST}$ if the set of all the $\mathcal{P}^*$-propositions in $\mathrm{ABox_D}$ is strongly grounded in $\mathcal{T}_{ST}$.

Definition 2. *(Strong Groundedness) $A \subseteq \mathrm{ABox_D}$ of $\mathcal{P}^*$-propositions is* strongly grounded *in $\mathcal{T}_{ST}$ if and only if the set of $\mathcal{P}^*$-propositions in $cl(\mathcal{T}_{\mathbf{TM+OB}} \cup \mathrm{TBox_D} \cup A)$ is simply grounded in $\mathcal{T}_{ST}$.*

Consider the previous example of correlation, and assume that $\mathtt{ROUND}_t x \to \mathtt{RED}_t x$ is in $\mathrm{TBox_D}$ (and neither $\mathtt{SUB}^+(\mathtt{ROUND}, \mathtt{RED})$ nor $\mathtt{SUB}^+(\overline{\mathtt{ROUND}}, \overline{\mathtt{RED}})$). Assume now that $\mathtt{ROUND_t c} \in \mathrm{ABox_D}$ and that $\mathtt{ROUND} \in \mathcal{P}^*$. If $\mathtt{RED} \notin \mathcal{P}^*$, both simple and strong groundedness do not guarantee the existence of red-states, the user just ignored evidences for $\mathtt{RED}$-propositions.[29] If $\mathtt{RED} \in \mathcal{P}^*$, according to strong groundedness, both a $\overline{\mathtt{ROUND}}$-state and a $\overline{\mathtt{RED}}$-state must be found in $\mathrm{ABox_{ST}}$ even though $\mathtt{RED_t c} \notin \mathrm{ABox_D}$. According to simple groundedness, if $\mathtt{RED_t c} \notin \mathrm{ABox_D}$, a $\overline{\mathtt{ROUND}}$-state would be enough, i.e., it is possible that no $\overline{\mathtt{RED}}$-state relative to $\mathtt{c}$ at $\mathtt{t}$ exists. While this is not acceptable in ontological terms, it is quite usual in an empirical perspective where we have only partial experimental measurements concerning $\mathtt{c}$ at $\mathtt{t}$. Strong groundedness requires all the inferred $\overline{\mathtt{RED}}$-propositions to be grounded, while for simple groundedness only the $\overline{\mathtt{RED}}$-proposition in $\mathrm{ABox_D}$ must be grounded.

[29] Actually $\mathcal{T}_{ST}$ could contain red states, it is the user that did not introduce any link between $\mathtt{RED}$ and $\overline{\mathtt{RED}}$.

Note that when both ROUND and RED belong to $\mathcal{P}^*$, in $\mathcal{T}_{\mathrm{ST}}$ the correlation $\mathtt{ROUND}_t x \rightarrow \mathtt{RED}_t x$ could be captured by the strong (generic) existential dependence $\mathtt{SED}(\bar{\mathtt{ROUND}}, \bar{\mathtt{RED}})$, where SED is defined in (d17). In a data-centric perspective, SED-relations can be intended as integrity constraints, i.e., a way to check if our data are logically consistent. Note that $\mathtt{SED}(\bar{\mathtt{ROUND}}, \bar{\mathtt{RED}})$ can be satisfied by the existence of several states, e.g., by the presence, in addition to $\mathbf{round}_{\mathtt{t}}\mathtt{c}$, of $\mathbf{crimson}_{\mathtt{t}}\mathtt{c}$, $\mathbf{red}_{\mathtt{t}}\mathtt{c}$, or both of them. One could also assume that the user, as in the case of SUB, introduces SED-relations to be used to close or complete the data, i.e., to add new (indirect) states into $\mathcal{V}_{\mathrm{ST}}$. In this case, starting from $\mathbf{round}_{\mathtt{t}}\mathtt{c}$, $\mathtt{SED}(\bar{\mathtt{ROUND}}, \bar{\mathtt{RED}})$ only introduces $\mathbf{red}_{\mathtt{t}}\mathtt{c}$, one has no information about the exact shade of red of c.[30]

d17 $\mathtt{SED}(\bar{\mathtt{P}}, \bar{\mathtt{Q}}) \triangleq \forall t x^n(\exists s(\bar{\mathtt{P}}s \wedge \varepsilon_t s \wedge x^n \multimap s) \rightarrow \exists r(\bar{\mathtt{Q}}r \wedge \varepsilon_t r \wedge x^n \multimap r))$

Notice that $\mathtt{SUB}^+$ is a specific case of SED where s and r coincide. Hence, material implications between object-properties may be grounded on different types of existential dependence between the corresponding state-kinds.[31] As said, here we consider only SUB-relations between state-kinds. The production of data is represented only via the data production primitive defined on individual states (not on state-kinds, i.e., it captures punctual dependencies not general ones) or by correlation rules in $\mathrm{TBox}_{\mathrm{D}}$. Section 7.3 takes into account data production dependencies to weaken the notions of groundedness and covering.

5.2. Covering

In the previous section we considered the assertions in $\mathcal{T}_{\mathrm{D}}$ that are supported by (the existence of) states in $\mathcal{T}_{\mathrm{ST}}$. Here we take into account the opposite direction, i.e., how and when the states in $\mathcal{T}_{\mathrm{ST}}$ have a correspondent assertion in $\mathcal{T}_{\mathrm{D}}$. We have already seen that the (and only the) $\bar{\mathcal{P}}^*$-predicates have an unique correspondent $\mathcal{P}^*$-predicate, i.e., $\gamma^{-1}(\bar{\mathtt{P}}) = \mathtt{P}$. The definitions of *covering* we propose are then a sort of *dual* of the ones of groundedness.

Definition 3 introduces the notion of (simple) covering for a set A of $\bar{\mathcal{P}}^*$-states in $\mathcal{T}_{\mathrm{ST}}$. We say that $\mathcal{T}_{\mathrm{D}}$ simply covers $\mathcal{T}_{\mathrm{ST}}$ if A is the set of all the $\bar{\mathcal{P}}^*$-states in $\mathcal{T}_{\mathrm{ST}}$. Like grounding, the same state can be covered by different propositions and the same proposition can cover different states. Strong covering, see Definition 4, requires the covering proposition to belong to $\mathrm{ABox}_{\mathrm{D}}$.

Definition 3. *(Simple Covering)* $\mathcal{T}_{\mathrm{D}}$ simply covers *the set* $A \subseteq \mathcal{C}_{\mathrm{ST}}$ *of* $\bar{\mathcal{P}}^*$*-states in* $\mathcal{T}_{\mathrm{ST}}$ *if and only if for every* $\mathtt{s} \in A, \mathtt{t}, \mathtt{c}^n \in \mathcal{C}_{\mathbf{TM+OB}}, \bar{\mathtt{P}} \in \bar{\mathcal{P}}^*$ *such that* $\mathcal{T}_{\mathrm{ST}} \vdash \bar{\mathtt{P}}\mathtt{s} \wedge \varepsilon_{\mathtt{t}}\mathtt{s} \wedge \mathtt{c}^n \multimap \mathtt{s}$ *then* $\mathcal{T}_{\mathrm{D}} \vdash \mathtt{P}_{\mathtt{t}}\mathtt{c}^n$.

Definition 4. *(Strong Covering)* $\mathcal{T}_{\mathrm{D}}$ strongly covers *the set* $A \subseteq \mathcal{C}_{\mathrm{ST}}$ *of* $\bar{\mathcal{P}}^*$*-states in* $\mathcal{T}_{\mathrm{ST}}$ *if and only if for every* $\mathtt{s} \in A, \mathtt{t}, \mathtt{c}^n \in \mathcal{C}_{\mathbf{TM+OB}}, \bar{\mathtt{P}} \in \bar{\mathcal{P}}^*$ *such that* $\mathcal{T}_{\mathrm{ST}} \vdash \bar{\mathtt{P}}\mathtt{s} \wedge \varepsilon_{\mathtt{t}}\mathtt{s} \wedge \mathtt{c}^n \multimap \mathtt{s}$ *then* $\mathtt{P}_{\mathtt{t}}\mathtt{c}^n \in \mathrm{ABox}_{\mathrm{D}}$.

Note that the notions of covering and groundedness are independent, i.e., covering does not rule out the possibility to have ungrounded $\mathcal{P}$-propositions in (the deductive closure of) $\mathrm{ABox}_{\mathrm{D}}$.

In a scientific scenario, one would like to import as much data as possible into $\mathcal{T}_{\mathrm{D}}$, i.e., to populate $\mathrm{ABox}_{\mathrm{D}}$ by assuring that $\mathcal{T}_{\mathrm{D}}$ strongly covers $\mathcal{T}_{\mathrm{ST}}$ and $\mathcal{T}_{\mathrm{D}}$ is simply grounded in $\mathcal{T}_{\mathrm{ST}}$. In this case, all the $\bar{\mathcal{P}}^*$-states in $\mathrm{ABox}_{\mathrm{ST}}$ have at least a correspondent $\mathrm{ABox}_{\mathrm{D}}$-assertion, and all the $\mathrm{ABox}_{\mathrm{D}}$-assertions have at least a supporting state. Strong groundedness guarantees that all the propositions in the deductive closure of $\mathrm{ABox}_{\mathrm{D}}$ have a supporting state, i.e., there exist empirical evidences of everything $\mathcal{T}_{\mathrm{D}}$ can infer from the initial ABox. Vice versa, in the case of simple groundedness, some propositions inferred

[30] Actually, this is another reason to allow for the existence of states of non leaf-kinds.

[31] Implications with form $\neg \mathtt{P}_t x^n \rightarrow \mathtt{Q}_t x^n$ are more problematic because it is not clear what guarantees the existence of $\bar{\mathtt{Q}}$-states.

from the initial data could lack an empirical support. However, even in this weaker scenario, since the states are not necessarily truthful, the importing of states could end up in a logically inconsistent $\mathcal{T}_{\mathrm{D}}$. In order to maintain the grounding of $\mathcal{T}_{\mathrm{D}}$ on data by preserving its consistency, one can consider (*i*) to avoid to import some states, i.e., states need to be filtered, aggregated, or harmonized before transferring them into $\mathrm{ABox}_{\mathrm{D}}$; (*ii*) to revise the theory $\mathcal{T}_{\mathrm{D}}$ to comply with the whole set of data; or (*iii*) a mix of harmonization of data and revision of the theory. Section 7.3 takes into account some of these general scenarios in the case of (weak) measurement.

6. States as facts or tropes

We briefly make precise a realist (ontological) view where states are entities that exist in the world like *facts* or holding *states of affairs*, see Armstrong (1997), or (relational) *tropes*, see Daly (1997) and Maurin (2014). In the following we will talk about facts, but everything holds also for tropes.

In this realist view, states are always completely determined and never disputable. The complete determination of facts can be partially characterized by adding to $\mathcal{T}_{\mathrm{ST}}$ the axioms (a22) and (a23) that assure states to be partitioned by leaf-state-kinds, i.e., the theory of facts $\mathcal{T}_{\mathrm{FC}} = \mathcal{T}_{\mathrm{ST}} \cup \{(a22), (a23)\}$. It follows that the identity criterion for states has the form (t1), i.e., for every $\bar{\mathtt{P}} \in \bar{\mathcal{P}}_{\mathcal{L}}$ ($\bar{\mathcal{P}}_{\mathcal{L}}$ is the set of leaves of $\bar{\mathcal{P}}$, see Section 3.2.2) we have a unique fact s such that $\bar{\mathtt{P}}s \wedge \varepsilon_t s \wedge x^n \multimap s$ (noted $\mathbf{p}_t x^n$). With reference to the taxonomy in Figure 1, this move implies, for example, that $\bar{\mathtt{RED}}$ is equivalent to $\bar{\mathtt{CRIMSON}}$, $\bar{\mathtt{OLIVE}}$ is equivalent to $\bar{\mathtt{YELLOW}}$, while $\bar{\mathtt{SHAPED}}$ is equivalent to the disjunction of $\bar{\mathtt{ROUND}}$ and $\bar{\mathtt{SQUARE}}$.

a22 $\mathrm{ST}s \rightarrow \bigvee_{\bar{\mathtt{P}} \in \bar{\mathcal{P}}_{\mathcal{L}}} \bar{\mathtt{P}}s$
a23 $\bigwedge_{\bar{\mathtt{P}} \neq \bar{\mathtt{Q}} \in \bar{\mathcal{P}}_{\mathcal{L}}} (\bar{\mathtt{P}}s \rightarrow \neg \bar{\mathtt{Q}}s)$
t1 $\bigvee_{\bar{\mathtt{P}} \in \bar{\mathcal{P}}_{\mathcal{L}}} (\bar{\mathtt{P}}s) \wedge s \bowtie s' \wedge s \otimes_{\varepsilon} s' \rightarrow s = s'$

Because facts are part of the world they cannot be wrong, the set of facts is an indisputable reference for any theory. If, once imported into $\mathcal{T}_{\mathrm{D}}$, some facts generate an inconsistency then only a revision of $\mathcal{T}_{\mathrm{D}}$ makes sense. However, note that neither $\mathcal{T}_{\mathrm{D}}$ is necessarily grounded on facts, nor facts are necessarily covered by $\mathcal{T}_{\mathrm{D}}$. In the first case, $\mathcal{T}_{\mathrm{FC}}$ represents only a subset of the whole facts. In the second case, $\mathcal{T}_{\mathrm{D}}$ focuses just on some aspects of the world, on some specific domains of interest. In addition, $\mathcal{T}_{\mathrm{D}}$ may be partially underdeterminate with respect to the world. For instance, $\mathtt{RED} \in \mathcal{P}^*$ may be a leaf in $\mathcal{T}_{\mathrm{D}}$ while, in $\mathcal{T}_{\mathrm{FC}}$, $\bar{\mathtt{RED}}$ may subsume some leaves, e.g., $\bar{\mathtt{SCARLET}}$ and $\bar{\mathtt{CRIMSON}}$.

In this realist perspective, one could think that the perfect matching between theory and facts occurs when (*i*) $\mathcal{T}_{\mathrm{D}}$ is strongly grounded in $\mathcal{T}_{\mathrm{FC}}$ and (*ii*) $\mathcal{T}_{\mathrm{D}}$ strongly covers $\mathcal{T}_{\mathrm{FC}}$. Still the $(\mathcal{P} \setminus \mathcal{P}^*)$-propositions are not necessarily grounded in $\mathcal{T}_{\mathrm{FC}}$, even though, in some cases, they are deducible from grounded $\mathcal{P}^*$-propositions. In addition to the conditions (*i*) and (*ii*) one could then add that $\mathcal{P}^* = \mathcal{P}$, i.e., $\mathcal{T}_{\mathrm{D}}$ and $\mathcal{T}_{\mathrm{FC}}$ are strictly aligned. The facts may then be seen as the *truth-makers* of $\mathcal{P}$-propositions, what exists in the world that makes $\mathcal{P}$-propositions true. Armstrong (1997) assumes ($\langle x \text{ exists} \rangle$ indicates the proposition 'that x exists' and p entails q if and only if there is no possible world where p holds but q fails):

(A) "If a proposition p is true then there exists an entity x such that $\langle x \text{ exists} \rangle$ entails p." (Armstrong, 1997, p.115).

Slightly differently, Rodriguez-Pereyra (2005) considers:

(TM) "Necessarily, if $\langle p \rangle$ is true, then there is some entity in virtue of which it is true," i.e., "if a proposition is true there must be an entity that would not have existed if the proposition in question had been false" (Rodriguez-Pereyra, 2005, p.19). Without entering this debate,[32] we just highlight that truthmaking has an *explicative* power: the truth of p is explained in terms of the existence of the truth-maker x, p is true in virtue of x. This explicative aspect is only partially captured by our framework, which represents a simple correspondence between true propositions and facts.

7. States as observations

In this section we focus on a cognitive or empirical interpretation of states, we talk of simple and complex observations rather than states. In Section 2 we have seen that observations represent both direct classifications like sensations or simple measurements, and classifications resulting from high-level cognitive or empirical processes like reasoning, testing, reporting, etc. These processes rely on simple observations to build explicit, concise, meaningful, and cognitively effective classifications. For instance, in data analysis, one starts from raw data to build indicators that can be further elaborated. In metrology, one can represent the way data are collected, their origin or provenance, the involved devices or observers, etc.

One possible way to capture this generative aspect is to consider all the observations in $\mathcal{T}_{\mathrm{ST}}$ as raw and perform all the elaborations at the level of $\mathcal{T}_{\mathrm{D}}$ by introducing some axioms that represent the needed data-aggregation rules. This strategy is perfectly acceptable but it has at least two issues. First, the theory $\mathcal{T}_{\mathrm{D}}$ would mix the realm of data (analysis) and the one of the laws that are supposed to govern the world, the realm of the empirical practice and the realm of purely theoretical knowledge. Second, and more importantly, this strategy would imply that (*i*) for all the complex data taken into account, the way they are calculated and the raw data from which they are produced are explicitly represented in $\mathcal{T}_{\mathrm{D}}$;[33] and (*ii*) no wrongly-produced complex data exist, every complex datum is perfectly well defined and all the elaborations are free of mistakes. We prefer to follow a weaker approach that lacks the explanatory power of this ideal approach but is able to manage *partial* information about the *origin* (the *provenance*) of complex data. Rather than representing *how* complex data are produced, we just introduce a mechanism (namely, the *data production* relation introduced in the next section) to keep track of the data on which they depend, of the data that has been used to produce them. This seems a more realistic position in a scenario where data are made available by different sources without necessarily a complete information on the way they have been achieved and where the possibility to have false data or wrong elaborations is not negligible. This weak framework cannot be used to certify the (dis)alignment of the empirical or experimental data with the theoretical framework under test. However, as we will see in Section 7.3, it allows to address some simple kinds of mismatches between $\mathcal{T}_{\mathrm{D}}$ and $\mathcal{T}_{\mathrm{ST}}$.

7.1. The primitive of data production

Data production—or, simply, production—is a (cognitive) process that induces a *simple* observation to emerge from a (possibly) complex one: $x \lessdot_t s$ stands for "at time t, the simple observation s is *directly produced* starting from the (complex) observation x" (a24). We consider here only *direct* production, i.e., productions with no intermediate steps (a25), the basic blocks for building production-chains, see

[32]The interested reader can refer to Beebee and Dodd (2005) for a deep discussion.

[33]This also impacts the required expressive power of the logic behind $\mathcal{T}_{\mathrm{D}}$, given the fact that often the indicators are the result of complex statistical analysis, etc.

Section 7.3. Data production is a sort of (temporally qualified) *specific* dependence. In particular, here we focus on a *synchronic* dependence, i.e., both the 'producer' and the 'product' must *completely* exist when a production relation holds (a26). Productions that contemplate temporal patterns of diachronic observations, forms of historical dependence, are not considered in this work.

a24 $x \prec_t y \rightarrow \mathrm{CST}x \wedge \mathrm{ST}y \wedge \mathrm{TM}t$
a25 $x \prec_t y \rightarrow \neg\exists z(x \prec_t z \wedge z \prec_t y)$
a26 $x \prec_t y \rightarrow \hat{\varepsilon}_t x \wedge \varepsilon_t y$

Logically, $x \prec_t s$ can be seen as a form of inference, a dependence of the information contained in s on the one in x. However, our production primitive is not necessarily truth-preserving, it just allows to take track of the data advocated for the introduction of a new datum. Neither the truth of the starting data nor the validity of the elaboration of these data are guaranteed, e.g., one can encounter faulty productions. To check the *validity* of a given data production and then to filter out incorrect production-statements, additional information is required. Note that we represent neither who or what performs the data analysis nor the way the raw data are elaborated. These are interesting extensions of our framework whose usefulness will be briefly discussed in Section 7.3.

The temporal qualification of the production-primitive is necessary because observations can exist at several times, therefore, the data that produce them can change through time, i.e., (f7) does not hold in our framework (even in the case $\hat{\varepsilon}_{t'} x$). For instance, the weight of an object o can be measured at several times by using different devices even when the weight of o does not change. Second, even at a single time, an observation can rely on different raw data. For instance, both $\mathbf{r}_t(x,d)+\mathbf{p}_t(d) \prec_t \mathbf{9.109 \times 10^{-31}kg}_t(x)$ and $\mathbf{1.602 \times 10^{-19}coulomb}_t(x) \prec_t \mathbf{9.109 \times 10^{-31}kg}_t(x)$ are possible. In the first case the mass of an electron x is determined by observing a P-configuration of the scale d when R-connected to x, while, in the second case, by applying a physical law. Third, there are observations, called *primitive* observations or raw data, that are not produced. Primitive observations supply a starting point to production-chains. Intuitively, they include phenomenological conscious sensations, simple readings of the outputs of the technical devices by the operators, or, simple measurements (in the Carnap sense previously discussed). However, in our framework, they are observations that have been endorsed by someone (not explicitly represented in the theory) without providing any information about their origin. Fourth, one is tempted to assume that if $x \prec_t y$ holds then, at any time x completely exists, x continues to be linked with y by the production-primitive (f8). We do not commit to this view because the holding of the production-relation has to be intended as an *explicit* statement resulting from underlying cognitive processes. The production $x \prec_t y$ represents an explicit commitment to the fact that, at t, y has been produced starting from x. Both x and y could exist at t' without any production-relationship (at t') between them, without any explicit elaboration of (the observations that compose) x to obtain y.

f7 $x \prec_t y \wedge \varepsilon_{t'} y \rightarrow x \prec_{t'} y$
f8 $x \prec_t y \wedge \hat{\varepsilon}_{t'} x \rightarrow x \prec_{t'} y$

Note that the production-primitive could also represent ontological information. Consider, for instance, $\mathbf{p}_t(x) \prec_t \mathbf{q}_t(x)$. At least two readings are possible: (*i*) we are in presence of an ontological correlation, e.g., all the objects with mass 9.109×10^{-31}kg have an electric charge of 1.602×10^{-19} coulomb, or (*ii*) we are in presence of a generalization, e.g., all the scarlet objects are red. Consider now $(\mathbf{inheres}_t(x,q)+\mathbf{p}_t(q)) \prec_t \mathbf{p}_t(x)$. Here, q can be seen as an *individual quality* or *trope* that inheres in x. The classification of x under P depends on the classification of its trope under P, e.g., x appears red because a red-trope inheres in it.

Our theory of data production $\mathcal{T}_{PR}$ is then an extension of $\mathcal{T}_{ST}$ such that:

- $\mathcal{V}_{PR} = \mathcal{V}_{ST} \cup \{\lessdot\}$, and
- $\mathcal{T}_{PR} = \mathcal{T}_{ST} \cup \{(a24), (a25), (a26)\} \cup \text{ABox}_{PR} \cup \text{TBox}_{PR}$
  where ABox_{PR} and TBox_{PR} extend, respectively, ABox_{ST} and TBox_{ST} with statements that involve the primitive $\lessdot$ of direct production.

7.2. Weak measurement

We focus here on one special kind of data production called *evaluation* that has the form in (f9) where $\mathbf{q}_t(x)$ is produced from (*i*) a relational observation concerning both x and m and (*ii*) a classification of m. The observation $\mathbf{p}_t(m)$ is a sort of *proxy* for $\mathbf{q}_t(x)$, i.e., m is a *mediator* able to transduce, by R-interacting with x, a property of x into a property of m. By connecting the object x to the mediator m in a qualified way, some observations about x can be indirectly obtained by observing m. *Measurements* may be seen as specific evaluations, e.g., the weight of x is obtained by observing that x is on the plate of the scale m and by observing the position of the pointer of m.

f9 $(\mathbf{r}_t(x,m)+\mathbf{p}_t(m)) \lessdot_t \mathbf{q}_t(x)$ (*evaluation*)

d18 $\sigma_t^m s \triangleq \bigvee_{\mathbf{r},\mathbf{p}\in\mathbf{P}}(\exists x s_1 s_2(x \multimap s \wedge s_1{=}\mathbf{r}_t(x,m) \wedge s_2{=}\mathbf{p}_t(m) \wedge s_1{+}s_2 \lessdot_t s))$ (*source*)

For evaluations, the *source* relation σ can be defined as in (d18)[34]: $\sigma_t^m s$ stands for "at time t, m is a *source* or *origin* of the observation s". The source m coincides with the mediator, *what* or *who* interacted with the participant of s during the evaluative process. When the mediator is a person we talk of pure evaluation, opinion, sensation, expertise, etc. Note that some observations may have several sources that can be further characterized by information about its kind, its reliability, etc., see Janowicz and Compton (2010). We do not consider this additional information that, however, can be easily integrated in our framework.

As said before, the production primitive does not guarantee the validity of the production process. The theory of measurement allows to introduce some constraints to rule out *material* errors in the data production. More specifically, here we focus on the case of *qualitative* evaluations, a perspective explicitly addressed by the theory of *weak measurement* introduced by Finkelstein (2003) and further elaborated by Mari (2013). In this weak perspective, measurement does not necessarily involve *quantities* (interval or ratio properties) but also *qualities* (nominal or ordinal properties). Qualitative classification plays a fundamental role in disciplines like psychology, medicine, or sociology where non-physical properties can be attributed to the subjects via the administration of *tests*. Measurement becomes "uncorrelated with quantification: the measurability of a property is a feature derived from experiment, not algebraic constraints" (Mari, 2013, p.2894). The basic formula of the quality calculus is $p = \{p\}\,\mathbf{in}\,[p]$ where: (*i*) p is a property of an object (e.g., the color or the shape of an object); (*ii*) $[p]$ is a *classification system*, a system of properties all related to the same quality (e.g., the color-properties); and (*iii*) $\{p\}$ is an element of $[p]$, it individuates the position of the object under measurement in the system $[p]$ (e.g., scarlet for colors), see Mari and Giordani (2012) for more details.[35] (Weak) measurement commits to *individual properties* or *tropes* of objects. The value $\{p\}$ is attributed to an individual property of the object under measurement, e.g., it is the *color of the object* that is scarlet. The object is (indirectly) scarlet just because

[34]$\mathbf{P}$ is the set of functions that is in a one-to-one correspondence with the set of state kinds $\bar{\mathcal{P}}$.

[35]Usually, the classification systems are structured. The *domains* of *conceptual spaces* (see Gärdenfors (2000)) and the *quality spaces* of DOLCE-CORE (see Borgo and Masolo (2009)) can then be seen as classification systems.

its color is scarlet (see the end of Section 7.1). However, as shown in Masolo (2010), by explicitly taking into account the measurement devices, one can avoid this commitment and consider measurement as a (partial) mapping from objects to properties in the system $[p]$. Given the fact that the evaluation form (f9) explicitly refers to the mediators, here we embrace this less committed approach.

According to Mari (2013), the distinction between *measurement* and *evaluation* is based on the *objectivity* and *inter-subjectivity* of the results: measurements must be sharable by different subjects at different times and in different places. To achieve this, measurement relies on *measurement standards* and *calibration*. A device transduces the interaction with an object into an internal state (of the device connected to the object) that is empirically accessible via the pointer. Calibration provides a meaning to the positions of the pointers. Once a physical or theoretical measurement standard isomorphic to $[p]$ is established, calibration—see Masolo (2010) and Frigerio et al. (2009) for details—establishes a one-to-one correspondence between the positions of the pointers and the properties in $[p]$, i.e., the positions of the pointers stand for properties, the device physically embodies the classification system.

Let us analyze the conditions evaluations must satisfy to be classified as measurements. First, the mediators must be measurement devices, i.e., objects with the design characteristics previously discussed. Here we do not explicitly consider these characteristics, we simply introduce a finite number of kinds of stable and calibrated devices all subsumed by OB. Each kind of devices is characterized in terms of (*i*) the possible positions of the pointers, and (*ii*) the possible ways an object can be connected to the devices. Given a device-kind $\mathtt{D}$, we assume that $\bar{\mathcal{D}} \subseteq \bar{\mathcal{P}}$ characterizes the configurations (of the pointers) of the devices while $\bar{\mathcal{RD}} \subseteq \bar{\mathcal{P}}$ specifies how the input objects must be connected to the devices.

Second, we need to represent the classification system $[p]$. To account for the possibility to classify objects at different *resolutions*, we allow $[p]$ to contain multi-resolution properties, i.e., the properties in $[p]$ can be *taxonomically* structured. In our framework, a multi-resolution system can be represented by a taxonomically structured $\bar{\mathcal{S}} \subseteq \bar{\mathcal{P}}$. For instance, in Figure 1, one can consider $\bar{\mathcal{S}} = \{\overline{\mathtt{COLORED}}, \overline{\mathtt{RED}}, \overline{\mathtt{CRIMSON}}, \overline{\mathtt{YELLOW}}, \overline{\mathtt{OLIVE}}, \overline{\mathtt{GREEN}}, \overline{\mathtt{EMERALD}}\}$ that contains predicates at different levels of resolution, e.g., $\overline{\mathtt{COLORED}}$, $\overline{\mathtt{RED}}$, and $\overline{\mathtt{CRIMSON}}$. Taxonomically unstructured flat systems can also be considered, e.g., $\bar{\mathcal{S}} = \{\overline{\mathtt{CRIMSON}}, \overline{\mathtt{OLIVE}}, \overline{\mathtt{EMERALD}}\}$.

Third, to represent calibration, we need to individuate the properties in $\bar{\mathcal{S}}$ the observation-kinds in $\bar{\mathcal{D}}$ stand for. We need then an embedding of $\bar{\mathcal{D}}$ into $\bar{\mathcal{S}}$. This embedding represents the calibration of $\mathtt{D}$-devices with respect to the system $\bar{\mathcal{S}}$, i.e., we explicitly represent neither the calibration process nor the measurement standard that allows for the calibration. Given a device-kind $\mathtt{D}$, $\bar{\mathtt{P}} \in \bar{\mathcal{D}}$, $\bar{\mathtt{R}} \in \bar{\mathcal{RD}}$, and a classification system $\bar{\mathcal{S}}$ the device refer to, calibration constraints have the form in (f10) where $\bar{\mathtt{Q}} \in \bar{\mathcal{S}}$.

f10 $(\mathtt{D}m \wedge (\mathbf{r}_t(x,m){+}\mathbf{p}_t(m)) \lessdot_t s \wedge x \multimap s) \rightarrow \bar{\mathtt{Q}}s$

Once the calibration constraints are available, it becomes trivial to filter out the evaluations that are not measurements. If the observations $\mathbf{r}_t(x,m)$ and $\mathbf{p}_t(m)$ are not primitive, the calibration and filtering process can be applied also to them. Vice versa, the checking of the correctness of pure evaluations—e.g., when the mediator is a person, group, or institution—could involve social or historical behaviors of the mediator that are very difficult to be analyzed and represented.

Let us extend $\mathcal{T}_{\text{TM+OB}}$ with n device kinds $\mathtt{D}_i$ (such that $\mathtt{D}_i x \rightarrow \text{OB}x$) and a set of individual constants for devices. In this way the theories $\mathcal{T}_{\text{ST}}$ and $\mathcal{T}_{\text{PR}}$ contain individual constants and predicates also for states that concern devices. The theory $\mathcal{T}_{\text{MS}}$ of *(weak) measurement* is defined as in the following:

- $\mathcal{V}_{\text{MS}} = \mathcal{V}_{\text{PR}}$
- $\mathcal{T}_{\text{MS}} = \mathcal{T}_{\text{PR}} \cup \text{TBox}_{\text{MS}}$, where

TBox_{MS} is a set of axioms with form (f10) that refer to m measurement standards $\bar{\mathcal{S}}_j \subseteq \bar{\mathcal{P}}$, n sets $\bar{\mathcal{D}}_i \subseteq \bar{\mathcal{P}}$ (one for each device-kind) that characterize the configurations of the $\mathtt{D}_i$-devices, and n sets $\bar{\mathcal{RD}}_i \subseteq \bar{\mathcal{P}}$ that specify how the input objects must be connected to the $\mathtt{D}_i$-devices.

7.3. Linking models with measurements

We have seen that the calibration constraints help in discovering material mistakes. Consider now two devices d_1 and d_2 of kind D1 and D2, both calibrated with the classification system $\bar{\mathcal{S}}$. Furthermore, assume that the object x is correctly connected to d_1 by $\bar{\mathtt{R1}} \in \bar{\mathcal{RD}}_1$ and to d_2 by $\bar{\mathtt{R2}} \in \bar{\mathcal{RD}}_2$ and that the calibration constraints are satisfied. In the situation represented in (f11), all these conditions do not guarantee the identity of $\mathbf{q1}_t(x)$ and $\mathbf{q2}_t(x)$, it is still possible to have two different synchronous measurements of x both relative to $\bar{\mathcal{S}}$.

f11 $(\mathbf{r1}_t(x, d_1)+\mathbf{p1}_t(d_1)) \prec_t \mathbf{q1}_t(x) \wedge (\mathbf{r2}_t(x, d_2)+\mathbf{p2}_t(d_2)) \prec_t \mathbf{q2}_t(x)$

This difference can be due to the resolution of the devices, e.g., $\mathbf{q1}_t(x) = \mathbf{crimson}_t(x)$ and $\mathbf{q2}_t(x) = \mathbf{red}_t(x)$, or to the kind of receptors the devices are equipped with. For instance, in Figure 1, $\mathtt{OL\bar{I}VE}$ is subsumed by both $\mathtt{YEL\bar{L}OW}$ and $\mathtt{GR\bar{E}EN}$, therefore $\mathbf{q1}_t(x) = \mathbf{yellow}_t(x)$ and $\mathbf{q2}_t(x) = \mathbf{green}_t(x)$ can be justified by the lack of information about the exact shade of x: d_1 classifies an olive shade as yellow, while d_2 as green. Disagreements like $\mathbf{q1}_t(x) = \mathbf{olive}_t(x)$ and $\mathbf{q2}_t(x) = \mathbf{crimson}_t(x)$ are less easy to be justified because, intuitively, *being olive* and *being crimson* are incompatible properties, no calibrated devices should, in principle, produce these results. However, in the scientific and ordinary practice, sometimes devices are used in a wrong way, in extreme environmental conditions, or they are just malfunctioning. Thus, an epistemological approach cannot exclude the previous kind of conflictual observations.

Our framework does not contain disjointness constraints that concern the leaves of the $\bar{\mathcal{P}}$-taxonomy, therefore, for instance, the existence of both $\mathbf{olive}_{\mathtt{t}}(\mathtt{c})$ and $\mathbf{crimson}_{\mathtt{t}}(\mathtt{c})$ in ABox_{MS} does not generate a logical inconsistency. This is a prerequisite to allow $\mathcal{T}_{\text{MS}}$ to contain observations that may produce inconsistencies when provided by a domain theory. To represent the fact that *being olive* and *being crimson* are, in a given theoretical (or ontological) perspective, incompatible properties, one needs to constrain the predicates OLIVE and CRIMSON. If ABox_{D} (simply or strongly) covers $\mathcal{T}_{\text{MS}}$ and $\mathtt{OLIVE}, \mathtt{CRIMSON} \in \mathcal{P}^*$ then, in the previous example, both $\mathtt{OLIVE}_{\mathtt{t}}\mathtt{c}$ and $\mathtt{CRIMSON}_{\mathtt{t}}\mathtt{c}$ are in ABox_{D}. Therefore, by adding $\mathtt{OLIVE}_t x \rightarrow \neg\mathtt{CRIMSON}_t x$ into TBox_{D}, $\mathcal{T}_{\text{D}}$ becomes inconsistent. To avoid the inconsistency one can (*i*) filter and clean ABox_{MS}, i.e., identify the *most plausible* observations among a set of (possibly) contradictory ones, and assure ABox_{D} covers only the filtered ABox_{MS};[36] (*ii*) revise $\mathcal{T}_{\text{D}}$ to comply with the full set of observations in $\mathcal{T}_{\text{MS}}$; or (*iii*) mix the two precedent strategies.[37]

One first possibility is to follow a brute procedure: start from ABox_{D} that simply covers and it is strongly grounded in $\mathcal{T}_{\text{MS}}$ and check if $\mathcal{T}_{\text{D}}$ is consistent. If it is consistent we are done. Otherwise start to randomly avoid to cover some observations and recheck if the new $\mathcal{T}_{\text{D}}$ is consistent. After some

[36]Alternatively one could import only states that have sources and contextualize all the ABox_{D}-assertions to their sources, i.e., the propositions that correspond to observations have the form $\mathtt{P}^d_t x$ and correspond in $\mathcal{T}_{\text{D}}$ to $\exists s(\bar{\mathtt{P}}s \wedge \sigma^d_t s \wedge x \multimap s)$. At this point the inconsistency is present only when the same source has contrasting outputs. The disagreement between sources can then be resolved at the level of $\mathcal{T}_{\text{D}}$ by introducing specific axioms that aggregate source-dependent propositions into source-independent ones. This would mean that (*i*) a subset of the ABox_{D} has an epistemological nature, it reflects the point of view of the devices on the world; (*ii*) that the source-independent propositions do not correspond to any observation; and (*iii*) that the empirical and theoretical levels are mixed up.

[37]By decoupling ABox_{MS} from ABox_{D}, i.e., observations from true $\mathcal{P}$-propositions, we shift towards a *verificationist* approach to truth: propositions must be verifiable, they are true only if they are verified, and *truth* "is constrained by our abilities to verify, and is thus constrained by our epistemic situation" Glanzberg (2014).

iterations one could identify a subset of states that can be safely imported into $\mathcal{T}_{\rm D}$. Alternatively, one can start to delete some axioms from $\mathcal{T}_{\rm D}$ and recheck if the new theory is consistent with the import of all the observations. A procedure that deletes both axioms and observations is also possible. These procedures are clearly inappropriate both from a theoretical and a computational perspective.

Let us focus on procedures that filter observations. One can try to follow a sort of *divide and conquer* strategy. Rather than trying to find a maximal set of states that are consistent with $\mathrm{TBox}_{\rm D}$[38] one could try to individuate n subsets $S_1, \ldots, S_n$ of $\bar{\mathcal{P}}^*$-states such that $S_1, \ldots, S_n$ contains, but not necessarily partitions, all the $\bar{\mathcal{P}}^*$-states in $\mathcal{T}_{\rm MS}$. Let us define $T_i = \mathcal{T}_{\rm TM+OB} \cup \mathrm{TBox}_{\rm D} \cup \mathrm{ABox}^i_{\rm D}$, where $\mathrm{ABox}^i_{\rm D}$ is the ABox that covers and is strongly grounded in S_i. If all the theories T_i are consistent, the standard hypothesis of the framework of judgment aggregation, cf. List and Puppe (2009), are met, and one can rely on aggregation techniques—like the ones analyzed in Porello and Endriss (2014)—to check and possibly solve potential inconsistencies and find a single integrated theory. If some T_i is inconsistent, it is necessary to apply a recursion step to S_i (and T_i). The attempt is to divide the consistency check by focusing on more simple and manageable sets of observations. However, we have the problem of identifying the sets S_i of observations.

Porello and Endriss (2014) assume the $T_1, \ldots, T_n$ as given. However, they conceive ontology aggregation (explicitly distinguished from ontology merging and integration) in the context of the theory of *judgement aggregation*: each ontology is seen as a voter, as an individual judgment about the truth of a list of propositions, and the goal is to aggregate the single views into a collective one.[39] In $\mathcal{T}_{\rm MS}$, a measurement can be defined as a state s for which there exist a time t and a device d such that $\sigma^d_t s$ where d represents the voter. Given a finite number of devices $\mathtt{d}_1, \ldots, \mathtt{d}_n$, $\mathcal{M}_i$ is the set of $\bar{\mathcal{P}}^*$-observations with source $\mathtt{d}_i$ (at least at one time of their existence). The set $\mathcal{M}_i$ of measurements offers a sort of agent-oriented perspective on observations where each device individually contributes to the whole information about the world, each device accesses and investigates the world in a peculiar way.

With *data aggregation* we refer to the aggregation of measurements taken by different devices. The theories T_i are then built on the basis of the sets $\mathcal{M}_i$ following what we have done with the S_i. The case of measurements is however more complex. In Porello and Endriss (2014), an ontology is a finite set of (closed) formulas in a given language (namely, for the sake of example, the description logic $\mathcal{ALC}$) and an ABox is a finite set of propositions with form $P(a_1, \ldots, a_n)$. In our framework, the $\mathrm{ABox}_{\rm D}$-propositions have an instantaneous temporal qualification, i.e., they have the form $\mathtt{P}_t x^n$. Vice versa, the states in the sets $\mathcal{M}_i$ are not necessarily linked to a single time and therefore they can change their source(s) or, at some times, they can lack a source. It follows that, to say that the proposition $\mathtt{P}_\mathtt{t}\mathtt{c}^n \in \mathrm{ABox}_{\rm D}$ corresponds to an observation collected by the device $\mathtt{d}_i$, it is not enough to find a state $\mathtt{s} \in \mathcal{M}_i$ such that $\mathcal{T}_{\rm PR} \vdash \bar{\mathtt{P}}\mathtt{s} \wedge \varepsilon_\mathtt{t}\mathtt{s} \wedge \mathtt{c}^n \multimap \mathtt{s}$ because $\mathtt{d}_i$ could not be the source of $\mathtt{s}$ at $\mathtt{t}$, i.e., $\mathtt{s}$ should belong to $\mathcal{M}_i$ because $\sigma^{\mathtt{d}_i}_{\mathtt{t}'}\mathtt{s}$ holds at $\mathtt{t}' \neq \mathtt{t}$. To guarantee that the propositions really correspond to observations collected by $\mathtt{d}_i$, we modify the notions of groundedness and covering by adding the condition $\sigma^{\mathtt{d}_i}_{\mathtt{t}}\mathtt{s}$ as done in Definition 5 and in Definition 6.

[38]When we talk of consistency of a set of states with $\mathrm{TBox}_{\rm D}$ we mean that, given the $\mathrm{ABox}_{\rm D}$ that simply covers and it is strongly grounded in this set of states, then $\mathcal{T}_{\rm D}$ is consistent.

[39]Porello and Endriss (2014) clarify some important differences between standard judgment aggregation and *ontology* aggregation. Only in the latter case we have that: (*i*) the set of propositions considered by the individuals is not closed under complementation, i.e., if a proposition p is available this does not imply also $\neg p$ is available; (*ii*) an open world assumption is necessary because agents cannot provide a judgment on each proposition about the world; and (*iii*) in logical terms it is possible to distinguish TBox-propositions from Abox-propositions and exploit this distinction in the aggregation of ontologies.

Definition 5. *(Strong M-Covering)* ABox_D strongly M-covers $\mathcal{M}_i$ *if and only if for every* $\mathsf{s} \in \mathcal{M}_i$*, every* $\mathsf{t}, \mathsf{c}^n \in \mathcal{C}_{\text{TM+OB}}$*, and every* $\bar{\mathsf{P}} \in \bar{\mathcal{P}}^*$ *such that* $\mathcal{T}_{\text{PR}} \vdash \sigma_{\mathsf{t}}^{\mathsf{d}_i}\mathsf{s} \wedge \bar{\mathsf{P}}\mathsf{s} \wedge \varepsilon_{\mathsf{t}}\mathsf{s} \wedge \mathsf{c}^n \multimap \mathsf{s}$ *then* $\mathsf{P}_{\mathsf{t}}\mathsf{c}^n \in \text{ABox}_\text{D}$.

Definition 6. *(Simple M-Groundedness) A set* $A \subseteq \text{ABox}_\text{D}$ *of* $\mathcal{P}^*$*-propositions is (simply)* M-grounded *in* $\mathcal{M}_i$ *if and only if for every* $\mathsf{P}_{\mathsf{t}}\mathsf{c}^n \in A$ *there exists a* $\mathsf{s} \in \mathcal{M}_i$ *such that* $\mathcal{T}_{\text{PR}} \vdash \sigma_{\mathsf{t}}^{\mathsf{d}_i}\mathsf{s} \wedge \bar{\mathsf{P}}\mathsf{s} \wedge \varepsilon_{\mathsf{t}}\mathsf{s} \wedge \mathsf{c}^n \multimap \mathsf{s}$.

Now, as in the case of the S_i, denote by ABox^i_D the set of $\mathcal{P}^*$-propositions that (*i*) are simply M-grounded in $\mathcal{M}_i$ and that (*ii*) strongly M-cover $\mathcal{M}_i$. Given a *data set* $\mathcal{DS}^n = \langle \mathcal{M}_1, \ldots, \mathcal{M}_n \rangle$, we can introduce $\langle \text{ABox}^1_\text{D}, \ldots, \text{ABox}^n_\text{D} \rangle$ and—using the terminology in Porello and Endriss (2014)—the *profile* $\mathcal{T}^n = \langle T_1, \ldots, T_n \rangle$ where $T_i = \mathcal{T}_{\text{TM+OB}} \cup \text{TBox}_\text{D} \cup \text{ABox}^i_\text{D}$.[40] This makes explicit in which sense data aggregation can be seen as a specific case of ontology aggregation where all the ontologies share $\mathcal{T}_{\text{TM+OB}} \cup \text{TBox}_\text{D}$ (and then the TBox) but they can have conflictual ABoxes. As said, in ontology aggregation, all the theories T_i are supposed to be consistent. By embracing this constraint, if $\mathcal{T}_{\text{TM+OB}} \cup \text{TBox}_\text{D}$ is consistent, then the data collected by a single device need to be consistent, i.e., the problems arise only when we put together data with different provenance. The inconsistency of the theory T_i, assuming the TBox_D is a reference theory that is not under test or revision, would imply the unreliability of the device d_i and the consequent discarding of (part of) the measurements collected by d_i. In the following we explore only the simple case where the sets of measurements collected by single devices that are inconsistent with TBox_D are totally discarded. This means that the previous devices $\mathsf{d}_1, \ldots \mathsf{d}_n$ represent all the consistent (with respect to $\mathcal{T}_\text{D}$) sources in $\mathcal{T}_{\text{MS}}$.

Given these premises, we can follow ontology aggregation strategies to define and study different *aggregators*, i.e., functions F that map a profile into a theory. The *union aggregator* F is defined by $F(\mathcal{T}^n) = T_1 \cup \ldots \cup T_n$ and it corresponds to an unfiltered import of all the measurements in $\mathcal{M} = \mathcal{M}_1 \cup \ldots \cup \mathcal{M}_n$. The *unanimity* aggregator F_u is defined by $F_u(\mathcal{T}^n) = T_1 \cap \ldots \cap T_n$ and it preserves the measurements unanimously collected by all the devices. The *absolute majority* aggregator F_m collects in $F_m(\mathcal{T}^n)$ all the measurements that are contained in more than $n/2$ T_i-theories. One can also adapt the properties of aggregators considered in Porello and Endriss (2014) to data aggregation: *anonymity* (the aggregator is impartial with respect to the devices), *neutrality* (the aggregator is impartial with respect to the measurements), *monotonicity* (the aggregator is sensitive to increasing the number of the devices that support a certain measurement), *independence* (the method of aggregation is invariant in each profile), *groundedness* (the aggregated theory does not introduce new formulas, i.e., $F(\mathcal{T}^n) \subseteq T_1 \cup \ldots \cup T_n$), *exhaustivity* (the aggregated theory is maximal, i.e., for no profile $\mathcal{T}^n$ there exists a formula $\psi \in T_1 \cup \ldots \cup T_n \backslash F(\mathcal{T}^n)$ such that $F(\mathcal{T}^n) \cup \{\psi\}$ is consistent). In particular, groundedness and exhaustivity seem reasonable requirements for data aggregation (at least in the case the theories T_i can be modified only by deleting some axiom).

No one of these properties or aggregators guarantee the consistency of $F(\mathcal{T}^n)$. Strategies to solve potential inconsistencies must then be taken into account. Notice that the procedures analyzed in Porello and Endriss (2014) are not intended to be directly usable in applications, they just "provide a catalogue of basic aggregators that can serve as building blocks for constructing more sophisticated procedures in the future" (Porello and Endriss, 2014, p.1241). In this perspective, the *support-based procedure* seems especially interesting for data aggregation. The support-based procedure works by (1) ordering measurements in terms of the number of devices that support them, and (2) by accepting formulas in decreasing order, but dropping the formulas that introduce an inconsistency. This procedure basically

[40]One could assume that the $\mathcal{M}_i$s contain also non $\bar{\mathcal{P}}^*$-states. In this case the ABox^i_D can be introduced by considering the maximal set of $\mathcal{P}$-propositions grounded in $\mathcal{M}_i$.

accepts measurements one after the other starting from the most supported one and guaranteeing the consistency. Porello and Endriss (2014) show that this procedure is anonymous, monotone, grounded and exhaustive—therefore it satisfies the requirements for data aggregation—but both neutrality and independence are violated. The *sequential* procedure is a variation of the previous one that orders devices according to, for instance, their reliability and then progressively includes into $F(\mathcal{T}^n)$ the ABox^i_{D} relative to a device starting from the most reliable device and discarding devices whose measurements would introduce an inconsistency with what is already in $F(\mathcal{T}^n)$. In our framework the support-based and sequential procedures may be modified to take into account the deductive closures. One starts by importing, for instance, the ABox^i_{D} relative to the most reliable device and consider $cl(\text{TBox}_{\text{D}} \cup \text{ABox}^i_{\text{D}})$. If the ABox of $cl(\text{TBox}_{\text{D}} \cup \text{ABox}^i_{\text{D}})$ contains new propositions (with respect to the starting ABox^i_{D}), one can assume that these data are reliable and then revise the ordering of devices taking into account the sources of the measurements that ground the inferred ABox-assertions. In this case we are using both the knowledge about the reliability of sources and the terminological one in TBox_{D} to identify new reliable sources. A similar mechanism can be exploited in the case of the support-based procedure: one can import the most shared measurements and identify through the deductive closure new measurements to be imported into the collective theory. As stated at the beginning of this section, all these procedures assure that $\mathcal{T}_{\mathbf{TM+OB}} \cup \text{TBox}_{\text{D}} \subseteq F(\mathcal{T}^n)$, i.e., they filter the measurements, the assertional knowledge, without impacting the terminological one.

In $\mathcal{T}_{\text{MS}}$ one may abstract from specific devices by considering the device-kinds $\mathtt{D}_j$. One can then group the measurements that are collected by devices of the same kind, obtaining a new data set $\langle \mathcal{M}^{\mathtt{D}}_1, \dots, \mathcal{M}^{\mathtt{D}}_k \rangle$ of the $\bar{\mathcal{P}}^*$-states and a new profile $\langle T^{\mathtt{D}}_1, \dots, T^{\mathtt{D}}_k \rangle$. One can then use knowledge about the reliability of the kinds of devices to apply support-based or sequential procedures. In this case, the hypothesis on the consistency of the ontologies $T^{\mathtt{D}}_j$ would be quite critical because the malfunctioning of few devices of a given kind would cause the discarding of all the $\mathcal{M}^{\mathtt{D}}_j$ data. One can then assume a two steps procedure that divides again the problem in hopefully easier problems, i.e., instead of starting from $\langle \mathcal{M}^{\mathtt{D}}_1, \dots, \mathcal{M}^{\mathtt{D}}_k \rangle$ one may consider the data set $\langle F_1(T_{11}, \dots, T_{1n_1}), \dots, F_k(T_{k1}, \dots, T_{kn_k}) \rangle$ where the profiles $\langle T_{i1}, \dots, T_{in_i} \rangle$ are determined by taking into account, as done before, the measurements collected by all the n_i devices of kind $\mathtt{D}_i$. The two steps aggregation is explicit because the output ontology is $F(F_1(T_{11}, \dots, T_{1n_1}), \dots, F_k(T_{k1}, \dots, T_{kn_k}))$.

Another interesting modification relies on the use of the *partial* information represented by the data production relation to extend the set of states imported into the theories T_i. Each set $\mathcal{M}_i$ can be closed under a production chain of a given length, a sort of 'production closure', the correspondent of the deductive closure in terms of $\lessdot$. More formally, $cl^m_{\lessdot}(\mathcal{M}_i)$ is the set of states that can be produced in at most m-steps from $\mathcal{M}_i$-states:

- $cl^0_{\lessdot}(\mathcal{M}_i) = \mathcal{M}_i$;
- for $n > 0$, $cl^n_{\lessdot}(\mathcal{M}_i)$ is $cl^{n-1}_{\lessdot}(\mathcal{M}_i)$ union all the states that are directly produced from complex states whose atomic components are all in $cl^{n-1}_{\lessdot}(\mathcal{M}_i)$.

We can then mimic everything done in the case of $\langle \mathcal{M}_1, \dots, \mathcal{M}_n \rangle$ by considering the new data set $\langle cl^m_{\lessdot}(\mathcal{M}_1), \dots, cl^m_{\lessdot}(\mathcal{M}_n) \rangle$ where m is the number of steps taken into account. Informally this means that we consider not only the measurements directly collected by the devices but also all the observations explicitly produced from these measurements. This strategy could be further generalized by taking into account observations produced by using measurements collected by different devices. Note that by adding a parameter to the data production primitive, one could take track of *who* produced the data or of the *method* used. This extension allows to refine the previous analysis by individuating subsets of the $cl^{m}_{\lessdot}(\mathcal{M}_i)$ that contain observations produced by the same agent or by using the same method.

The information about the reliability of devices could be also used to modify TBox_D. For instance, one could import all the observations taken by the most reliable device d_i into ABox_D. If the resulting $\mathcal{T}_\text{D}$ is inconsistent, instead of discarding the data collected by d_i, one may delete some rules in TBox_D. The previously discussed support-based and sequential procedures (and their modifications) can be easily adapted to the case where the observations in $\mathcal{T}_\text{ST}$ are more reliable than the model $\mathcal{T}_\text{D}$.

7.4. Data production as an observable

We briefly sketch the idea that the cognitive production processes underlying the relation $\lessdot$, can be, in their turn, observed, i.e., we introduce observations about the way observations are produced in terms of other observations. The data production relation must then be moved from the 'ontological' to the 'epistemological' realm, i.e., the primitive relation $\lessdot$ must be replaced by a new kind of observations. We then extend $\bar{\mathcal{P}}$ with the new $\mathtt{PR\bar{O}D}$-kind that have two participants: $\mathbf{prod}_t(s, s')$ is the observation about the fact that, at time t, the simple observation s' is directly produced by the complex observation s. Because $\mathtt{PR\bar{O}D}$-observations are about observations, (a6) must be modified to allow observations to participate (in the sense of the primitives $\multimap_{\mathtt{i}}$) in observations. The definition of provenance (d18) needs to be modified as in (d19). We can then replicate what done in Section 7.3 by just considering (d19).

d19 $\sigma^m_t s \triangleq \bigvee_{\mathbf{r},\mathbf{p}\in\mathbf{P}}(\exists x s_1 s_2 s_3 (x \multimap s \wedge s_1{=}\mathbf{r}_t(x, m) \wedge s_2{=}\mathbf{p}_t(m) \wedge s_3{=}\mathbf{prod}_t(s_1{+}s_2, s)))$

It is then possible to talk about the provenance inside $\mathcal{T}_\text{MS}$, i.e., one could have observations about the plausibility of a state s that are produced by taking into account the way s has been produced.

8. Conclusion

The development of suitable foundational theories about the interplay between theoretical models and empirical observations is an important step towards the definition of precise semantics and sound methodological principles for analytical investigations, where the reproducibility of analyses and experiments is of major importance. From an engineering point of view, there is a need for a formal representation of the activity of empirical research, in a suitable level of abstraction. A step in this direction is offered by the present work that is founded on two main distinctive features. First, observations are introduced into the domain of quantification of a first-order theory that precisely describes their nature and their organization and takes track (by means of the data production primitive) of the way they are experimentally acquired or intentionally elaborated. In this way, the observations and their provenance are uniformly represented in a single formal framework rather than as external meta annotations. Second, the proposed framework is based on the decoupling between the model of the domain of interest, which mainly represents the theoretical knowledge or hypotheses on the domain, and the model of the observations, which mainly represents the empirical knowledge and the given experimental practices. In this way, clean theoretical models and chaotic sets of observations with heterogeneous provenance can coexist making possible to formally manage the conflicts between theoretical and empirical knowledge. In particular, we explored the possibility to solve inconsistencies between a given set of observations and the assumed theoretical hypotheses by exploiting some techniques developed in the field of social choice theory and judgment aggregation. These solutions, which may rely on the information provided by the data-production primitive, may impact both the observations—e.g., the theoretical knowledge and the analysis of the way observations are collected or produced may highlight some unreliable sources—and the model of the domain—e.g., empirical evidences may invalidate some theoretical laws.

There are at least three main directions to be pursued for extending the present work. First, the coverage of our framework could be broadened (*i*) by including data productions that contemplate temporal patterns of diachronic observations, in order to deal with historical dependence; and (*ii*) by extending our theory with observations about events and processes, what requires an ontological analysis on the nature of events, a task we already sketched in Benevides and Masolo (2014). Second, one could explore the possibility to develop aggregation techniques specifically based on the information provided by data production. Even though this direction is quite interesting and promising, still it is necessary to better understand how to use the information about who produced a new observation and the method used. Third, the communities dealing with huge amounts of empirical data are also interested in reasoning on this data in an effective and efficient way. One can then explore alternative ways of formalizing the proposed framework, for example using description logics, that deal with different aspects of the trade-off between expressivity and computability/tractability.

References

Armstrong, D.M. (1989). *Universals: An opinionated introduction*. Westview Press.

Armstrong, D.M. (1997). *A world of states of affairs*. Cambridge University Press.

Armstrong, D.M. (2004). *Truth and truthmakers*. Cambridge University Press.

Batrancourt, B., Dojat, M., Gibaud, B. & Kassel, G. (2010). A core ontology of instruments used for neurological, behavioral and cognitive assessments.. In A. Galton and R. Mizoguchi (Eds.), *Proceedings of the 6th International Conference on Formal Ontology in Information Systems (FOIS 2010)*. Frontiers in artificial intelligence and applications (Vol. 209, pp. 185–198). IOS Press.

Beebee, H. & Dodd, J. (Eds.) (2005). *Truthmakers: the contemporary debate*. Oxford University Press.

Benevides, A.B. & Masolo, C. (2014). States, events, and truth-makers. In P. Garbacz and K. Oliver (Eds.), *Proceedings of the 8th International Conference on Formal Ontology in Information Systems (FOIS 2014)*. Frontiers in artificial intelligence and applications (Vol. 267, pp. 93–102). IOS Press.

Bogen, J. & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, *97*(3), 303–352.

Borgo, S. & Masolo, C. (2009). Foundational choices in DOLCE. In S. Staab and R. Studer (Eds.), *Handbook on Ontologies* (2nd ed., pp. 361–381). Springer.

Bottazzi, E., Ferrario, R. & Masolo, C. (2012). The mysterious appearance of objects. In M. Donnelly and G. Guizzardi (Eds.), *Proceedings of the 7th International Conference on Formal Ontologies in Information Systems (FOIS 2012)*. Frontiers in artificial intelligence and applications (Vol. 239, pp. 59–72). IOS Press.

Brinkman, R.R., Courtot, M., Derom, D., Fostel, J.M., He, Y., Lord, P., Malone, J., Parkinson, H., Peters, B., Rocca-Serra, P., et al. (2010). Modeling biomedical experimental processes with obi. *Journal of biomedical semantics*, *1*(1), 1–11.

Brogaard, B. (2012). *Transient truths: an essay in the metaphysics of propositions*. Oxford University Press.

Caplan, B. (2016). Soames's new conception of propositions. *Philosophical Studies*, *173*(9), 2533–2549.

Carnap, R. (1956). The methodological character of theoretical concepts. In H. Feigl and M. Scriven (Eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis* (pp. 38–75). University of Minnesota Press.

Carnap, R. (1966). *An introduction to the philosophy of science*. Basic Books.

Casati, R. & Varzi, A. (1999). *Parts and places*. MIT Press.

Daly, C. (1997). Tropes. In D.H. Mellor and A. Oliver (Eds.), *Properties* (pp. 140–159). Oxford University Press.

Finkelstein, L. (2003). Widely, strongly and weakly defined measurement. *Measurement*, *34*, 39–48.

Frigerio, A., Giordani, A. & Mari, L. (2009). Outline of a general model of measurement. *Synthese*, *175*(2), 123–149.

Galton, A. (1991). Reified temporal theories and how to unreify them. In J. Mylopoulos and R. Reiter (Eds.), *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI91)* (pp. 1177–1182). Morgan Kaufmann.

Galton, A. (2005). Eventualities. In M. Fisher, D. Gabbay and V. L. (Eds.), *Handbook of Temporal Reasoning in Artificial Intelligence* (pp. 25–58). Elsevier.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Bradford Books.

Glanzberg, M. (2014). Truth. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2014 ed.). http://plato.stanford.edu/archives/fall2014/entries/truth/.

Goodman, N. (1951). *The structure of appearance*. Harvard University Press.

Guarino, N. & Guizzardi, G. (2015). "We need to discuss the relationship": Re-visiting relationships as modeling constructs. In J. Zdravkovic and M. Kirikova (Eds.), *Proceedings of the 27th International Conference on Advanced Information Systems Engineering (CAISE 2015)*.

Halpin, T. & Morgan, T. (2008). *Information modeling and relational databases* (2nd ed.). Morgan Kaufmann.

Halvorson, H. (2013). The semantic view, if plausible, is syntactic. *Philosophy of Science*, *80*(3), 475–478.

Haslanger, S. (2003). Persistence through time. In M.J. Loux and D.W. Zimmerman (Eds.), *The Oxford Handbook of Metaphysics* (pp. 315–354). Oxford: Oxford University Press.

Janowicz, K. & Compton, M. (2010). The stimulus-sensor-observation ontology design pattern and its integration into the semantic sensor network ontology. In K. Taylor, A. Ayyagari and D.D. Roure (Eds.), *Proceedings of the 3rd International workshop on Semantic Sensor Networks (SSN10)*. CEUR.

Kim, J. (1976). Events as property exemplifications. In M. Brand and D. Walton (Eds.), *Action Theory* (pp. 159–177). Reidel.

King, J.C. (2014). Structured propositions. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014 ed.). http://plato.stanford.edu/archives/spr2014/entries/propositions-structured/.

King, J.C., Soames, S. & Speaks, J. (2014). *New thinking about propositions*. Oxford University Press.

Kuhn, W. (2009). A functional ontology of observation and measurement. In K. Janowicz, M. Raubal and S. Levashkin (Eds.), *GeoSpatial Semantics - Third International Conference (GeoS 2009)*. Lecture notes in computer science 5892 (pp. 26–43). Springer.

Landman, F. (2000). *Events and plurality: The jerusalem lectures*. Kluwer Academic.

List, C. & Puppe, C. (2009). Judgment aggregation. In P. Anand, P. Pattanaik and C. Puppe (Eds.), *Handbook of Rational and Social Choice*. Oxford University Press.

Lutz, S. (2014). Empirical adequacy in the received view. *Philosophy of Science*, *81*(5), 1171–1183.

Lutz, S. (2015). What was the syntax-semantics debate in the philosophy of science about? *Philosophy and Phenomenological Research,* doi:10.1111/phpr.12221.

Margolis, E. & Laurence, S. (Eds.) (1999). *Concepts: Core readings*. MIT Press.

Mari, L. (2013). A quest for the definition of measurement. *Measurement*, *46*, 2889–2895.

Mari, L. & Giordani, A. (2012). Quantity and quantity value. *Metrologia*, *49*, 756–764.

Masolo, C. (2010). Founding properties on measurement. In A. Galton and R. Mizoguchi (Eds.), *Proceedings of the 6th International Conference on Formal Ontology in Information Systems (FOIS 2010)*. Frontiers in artificial intelligence and applications (Vol. 209, pp. 89–102). IOS Press.

Masolo, C. (2016). Observations and their explanations. In *Proceedings of the 9th International Conference on Formal Ontology in Information Systems (FOIS 2016)*. Frontiers in artificial intelligence and applications (Vol. 283, pp. 197–210). IOS Press.

Masolo, C. & Porello, D. (2016). Understanding predication in conceptual spaces. In R. Ferrario and W. Kuhn (Eds.), *Proceedings of the 9th International Conference on Formal Ontology in Information Systems (FOIS 2016)*. Frontiers in artificial intelligence and applications (Vol. 283, pp. 139–152). IOS Press.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003). *Wonderweb deliverable d18 ontology library (final)*. Technical report.

Matthen, M. (2005). *Seeing, doing, and knowing. a philosophical theory of sense percetion.* Oxford Univeristy Press.

Maurin, A.-S. (2014). Tropes. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2014 ed.). http://plato.stanford.edu/archives/fall2014/entries/tropes/.

Mueller, E.T. (2006). *Commonsense reasoning*. Morgan Kaufmann.

Olivé, A. (2007). *Conceptual modeling of information systems*. Springer.

Parsons, T. (1990). *Events in the semantics of english: a study in subatomic semantics*. MIT Press.

Porello, D. & Endriss, U. (2014). Ontology merging as social choice: judgment aggregation under the open world assumption. *Journal of Logic and Computation*, *24*(6), 1229–1249.

Probst, F. (2008). Observations, measurements and semantic reference spaces. *Applied Ontology*, *3*(1-2), 63–90.

Reiter, R. (2001). *Knowledge in action*. MIT Press.

Rodriguez-Pereyra, G. (2005). Why truthmakers. In H. Beebee and J. Dodd (Eds.), *Truthmakers: the contemporary debate* (pp. 17–31). Oxford University Press.

Sanford, D.H. (2014). Determinates vs. determinables. In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2014 ed.). http://plato.stanford.edu/archives/sum2014/entries/determinate-determinables/.

Schiffer, S. (2016). Cognitive propositions. *Philosophical Studies*, *173*(9), 2551–2563.

Sider, T. (2001). *Four-dimensionalism. an ontology of persistence and time*. Clarendon Press.

Soames, S. (2010). *What is meaning?* Princeton University Press.

Soames, S. (2015). *Rethinking language, mind, and meaning*. Princeton University Press.

Soames, S. (2016). Yes, the search for explanation is all we have. *Philosophical Studies*, *173*(9), 2565–2573.

Suppes, P. (2002). *Representation and invariance of scientific structures*. CSLI Publications.

van Fraassen, B. (1989). *Laws and symmetry*. Clarendon Press.

Varzi, A.C. (2003). Riferimento, predicazione, e cambiamento. In C. Bianchi and A. Bottani (Eds.), *Significato e ontologia* (pp. 221–249). Franco Angeli.
Vila, L. & Reichgelt, H. (1996). The token reification approach to temporal reasoning. *Artifical Intelligence*, *83*, 59–74.