GFO-Data: Towards an Ontological Foundation of an Integrated Data Semantics

Heinrich Herre

In: Informatik und Gesellschaft
Eds. F. Fuchs-Kittowski, W. Kriesel
Peter Lang Internationaler Verlag der Wissenschaften
Frankfurt a. M., Bern, Bruxelles, New York, Oxford, Warszawa,
Wien 2016

Abstract

The capabilities to store and process data have been increasing exponentially during the past 10 such that today we face the problem of big data. Data are collected about anything which has a mode of existence; this can be objects, processes, pictures, verbal reports, and many other types of things. The final aim of data is not to collect more data, but to transform data into relevant applications.The current situation of data overload is caused by a lack of methods for abstraction and interpretation of data. In this paper we present the first version of a top level ontology for data, called GFO-Data (0) which is used to establish a semantic basis for a particular class of data. The current paper seems to be the first systematic ontological analysis of the notion of data which provides a basic classification of data.

1. Introduction

The term *data* occurs in various contexts: There are experimental data, textual data, visual data, heterogeneous data, and much more types or features of data. The capabilities to store and process data have been increasing exponentially during the past 10 such that today we face the problem of big data. The current situation of data overload is caused –in our opinion - by a lack of methods for abstraction and interpretation of data. There is a need for a language which provides a semantic basis for data and means for their correct representation. From this task four subtasks can be derived. First of all, we must clarify what data are and how they can be classified (*semantic problem*), then we need methods to acquire these data (*acquisition problem*), further we need means to correctly represent data in a formal framework (*representation problem*), and finally we need methods to evaluate and use these data (*utilization problem*). The current paper is devoted to the first task.

Data exhibit two basic aspects, a semantic and a syntactic one. The semantic aspect of data refers to its meaning, the syntactic one to the denotation or representation of the data's meaning by symbols or tokens. The symbol 1kg, for example, denotes an equivalence class of instances of the property *weight* which is established by a measurement process. The property *weight* is a concept, the specification of which presents the concept's intension. The semantics of the most elementary data refer to sense data. The color red, for example, is exemplified by a sense datum, which corresponds to a physical entity, namely, an electro-magnetic wave of a certain

frequency. The red as a sense datum can be understood as an interpretation of the corresponding physical entity by the mind.

Measuring instruments can be understood as artificial sense ogans which extend the class of naturally perceivable real entities of the world. For example, by instruments we may artificially perceive sounds of frequency much higher than sounds which can be perceived by animals, as for example bats. For physical or chemical entities, which are in the domain of our natural senses, we usually possess words, denoting the corresponding properties, for example the words *red*, *sweet*, or *sour*. An instrument "perceives" physical entities if it reacts to them, and transforms physical signals into a medium which can be accessed by Humans. In the simplest case such a medium is constituted by unit measures and scales which are specified by using mathematical entities, for example real numbers and other linear orderings.

We assume that data always existentially depend on bearers. *This blue*, for example, denotes an individual quality for which there exists a uniquely determined object as a bearer. We may speak of *this blue* of *this eye* or of *this red* of *this flower* etc. Data which are related to our senses belong to the realm of phenomenal data. There are types of data which are not related to sense data or mesasuring instruments, and, hence, which cannot be perceived or measured. The acquisition of non-measurable or non-perceivable data must be realized by higher cognitive procedures of the mind.

The communication, storing and processing of data need syntactic representations, being tokens of symbol structures. The meaning of data cannot be stored in a data base, but only their representations. Often we know only the syntax of data, for example, a natural language text, presented in some form; though the exact meaning, the semantics, remains unclear. In general, the semantics of data is provided by a system of inter-related concepts which is the basis for grasping the meaning, for interpretation, and for understanding.

The current paper is organized as follows. In section 2 we summarize the basics of GFO which is used as a framework for the subsequent investigation. GFO provides analytical basic principles and the most important general notions, on which this investigation is grounded.[1] Section 3 contains the first version of a top level ontology of data, called GFO-Data (0). This ontology describes the semantics of certain types of data. Section 4 is devoted to the representation of data, i.e. to the syntax of data. In section 5 we consider and discuss the relation between data and knowledge, and section 6 summarizes the results, collects various remarks and of open problems for future research.

2. **Basics of GFO**

In this section we summarize the basic categories of GFO, the details of which can be found in (Herre, 2010). The term *entity* refers to anything which has a mode of existence. Entities are classified into categories and individuals.. The basic entities of space and time are chronoids and topoids; these are considered as individuals. The ontology of space and time is inspired by ideas in (Brentano, 1976). The GFO-theory of time is presented in (Baumann, Loebe, & Herre, 2012). Individuals are divided into concrete and abstract ones. Concrete individuals exist in time or space, whereas abstract individuals are independent of time and space. According to their relations to time, concrete individuals are classified into objects and processes. Processes

[1] In GFO this method is called "ontological reduction" or "ontological analysis".

happen in time and are said to have a temporal extension. Objects[2] persist through time and have a lifetime, which is a chronoid. An object exhibits at any time point of its lifetime a uniquely determined entity, called presential, which is wholly present at this time-point.

Examples of objects are this ball and this tree, being persisting entities with a lifetime. Examples of presentials are this ball and this tree, any of them being wholly present at a certain time boundary t. Hence, the specification of a presential additionally requires the declaration of a time boundary. In contrast to a presential, a process cannot be wholly present at a time boundary. Examples of processes are particular cases of the tossing of a ball, a 100m run as well as a surgical intervention, the conduction of a clinical trial, etc. For any process p having the chronoid c as its temporal extension, each temporal part of p is determined by taking a temporal part of c and restricting p to this sub-chronoid. Similarly, p can be restricted to a time boundary t if the latter is a time boundary or an inner boundary of c. The resulting entity is called a process boundary, which does not fall into the category of processes.

Another dimension for classification of concrete individuals is their complexity which is based on the notion of existential dependence. The most elementary entites are called atttributives; objects are bundles of attributives, whereas objects are composed to facts which are embedded into situations. Attributives are individuals, which are connected to other entities, called bearers. There is a variety of types of attributives, among them, qualities, roles, relators, dispositions, functions, and structural features. The bearers of these attributives can be objects, and processes. But also attributives themselves may be bearers of attributives. Categories whose instances are attributives are called properties. According to the different types of attributives we distinguish quality-properties (or intrinsic properties) and role-properties (extrinsic properties), and the role-properties are classified into relational role properties (abr. relational properties), social role properties (social properties), see (Loebe, 2007).

We assume that the world is organized into strata, and that these strata are classified and separated into layers. The term *level* denotes both strata and layers. This approach is inspired by Hartmann (Hartmann, (1935–1950).) and by Poli (Poli, 2001). GFO distinguishes at least four ontological strata of the world: the material, the mental-psychological, the social stratum, and the region of ideal entities. Every entity of the world participates in certain strata and its levels. Among these levels specific forms of categorical and existential dependencies hold. For example, a mental entity requires an animate material object as its existential bearer. The strata to which categories should be placed must then be determined. Concepts are rooted in the psychological and social stratum, and the investigation of this ontological region must use results of cognitive science.

We distinguish at least three kinds of categories: universals, concepts, and symbol structures. We hold, that any fully developed foundational ontology must include at least these three types of categories. *Universals* are constituents of the real world, they are associated to invariants of the spatio-temporal real world, they are something abstract that is in real things. Concepts are categories that are expressed by linguistic expressions and which are represented as meanings in someone's mind. *Symbols* are signs or texts that can be instantiated by tokens. There is a close relation between these three kinds of categories: a universal is captured by a

[2] There various connotations of the term "object" in the literature. In the current paper this term is used equivalently to the term "continuant" or "endurant". Though, the interpretation of this notion in GFO strongly differs from the interpretation in DOLCE (Borgo, 2010) or in BFO (Spear, 2006)

concept which is individually grasped by a mental representation, and the concept and its representation is denoted by a symbol structure, being an expression of a language. Texts and symbolic structures may be communicated by their instances that are physical tokens.

A relation is a particular category, the instances of which are relators. Relators have parts, called roles, which inhere in players. The compositum of a relator together whith the players of the corresponding roles, are called facts. Throughout the paper we restrict to the case that the players of relators are spatio-temporal individuals, being objects and/or processes. We assume that concrete relational facts are always parts of more complex entities, called situations.

3. Constituents of an ontology-based data semantics

Data depend on bearers, and we assume for the current version of GFO-Data that the bearers are concrete individuals. Data are classified with respect to three dimensions, specified by the bearer and connecting relations, by the level of abstraction, and by complexity. Atomic data are covered in GFO-Data (0) by attributives and the corresponding properties; they are constituents for complex data. The atomic data in GFO-Data(0) are restricted to the following types: qualities, relators and relational roles. According to the level of abstraction we distinguish three basic levels, phenomenal data, relational data, and relational propositions. Subsequently, these data levels are considered in more detail.

3.1 Phenomenal data.

The elementary form and the origin of these data are sense data, but also data wich can be measured by instruments. These data correspond to qualities. With respect to the bearers we distinguish between object-data and processual data. The term phenomenal datum has its origin in the theory of phenomenalism. Phenomenalism defends the view that whatever is finally meaningful can be expressed in terms of our own sense experience; hence, reference to objects is always finally a reference to sense experience.

3.1.1 Object-Data

3.1.1.1 Presentic object-data

The most important relation, connecting attributives to bearers, is the inherence relation; a quality inheres in an object. For example, the quality blue inheres in this eye or in this flower. Object-data exhibit at any time point of the object's life time a presentic entity, being wholly present at this time point. This means that an individual quality of an object, say an individual red, can be wholly accessed at time points. Categories of object data are: visual data, forms, color etc. Others are captured by measuring instruments, for example weight and size. The composition of an object with some of its qualities exhibits more complex data, called object-facts. Examples are the *person Hans Müller together with its weight of 70 kg*. Object facts which are constructed around an object lead to a bundle of qualities. If the object is a bundle of visual data, then the set of facts creates a visual whole, namely the object itself.

3.1.1.2 Non-Presentic object-data

Any object has a life time of non-zero duration. Formally, we could consider the life time of an object as a quality; another example is an electro-cardiogram of a patient. We assume that genuine qualities of an object are wholly present at every time-point of the object's life time. In 3D-Ontologies, as BFO and DOLCE, processes depend on objects, they are – so to say - qualities of objects. This approach leads to several difficulties. First of all, the notion of an object, being an enduring entity which is the bearer of processes, leads to inconsistencies, see (Barker, 2005), (Wahlberg, 2008); secondly, global processual properties are incompatible with the presentic nature of the object's qualities, and, finally, a reconstruction of processes in the framework of a 3D-ontology, which takes into account all relevant features of processes, seems to be impossible. In the GFO-approach, based on the integration law, for any object O there exists a process Proc(O), such that the process boundaries of Proc(O) coincide with presentials, exhibited by O. Then, we hold that the global non-presentic qualities of an object O should be associated to the underlying process Proc(O) and, hence, are borrowed from it.[3]

3.1.2 Processual Data

The bearers of processual data are processes. Processual data are classified into presentic and global.

3.1.2.1 Presentic processual data

These are data, associated to process boundaries. They must be wholly accessible at time points.

3.1.2.1.1 Isolated presentic data

The isolated presentic data of process boundaries do not need any reference to a process.They can be completely reduced to object qualities. These are typically qualities of objects, participating in the process. For example, consider the movement of a thrown red ball. Any boundary of this process contains a presentic red ball, and this quality can be accessed without reference to a process. Such object qualities are, so to say, independent of the process in which the ball participates.

3.1.2.1.2 Non-isolated presentic data

Presentic data of a process are qualities associated to process boundaries. Investigating the movement of the red ball B, we may consider the quality of the velocity of B at a time-point t. This quality cannot be specified without a preceding process. Hence, such qualities are called presentic non-isolated. Other examples are qualities of a presentic event which refer to a process. For example if a train comes to an end; this can be unterstood only if there was a movement, i.e. if there was a preceding process. Discrete changes within a process also belong to this kind of processual data, see (Baumann, Loebe, & Herre, 2012). Other types of such qualities can be found in physics, for example in fluid or aero-dynamics.

3.1.2.2 Global processual data

The global qualities of processes is the richest class of qualities of processes. A systematic classification of these qualities is in its initial stage. The main feature of them is that does not make any sense to specify them at a process boundary. One type of such qualities is abstracted

[3] We emphasize that the object is not the same entity as the underlying process. This point was extensively discussed in (Herre, 2015).

from time series. Time series can be understood as selections of qualities of process boundaries, linearly ordered by time. Such entities are called in GFO histories, see (Herre, et al., 2007). We may consider the values of a fixed property, say the temperature of a patient. Since the patient himself can be considered as a process Proc(P), the temperature can be measured at selected time-points and the corresponding set of values can be transformed into a curve. Such curves can be evaluated to draw conclusions, for example, that the fever curve is typical for the disease Malaria. Other examples are electro-cardiograms, or a long term blood pressure measurement. Such curves are global qualities of the underlying process which are abstracted from the set of presentic values of a property. The visualization of the pattern is an indirect global property, associated to the process. On the top level of a process' property some basic patterns can be established, see (Herre, et al., 2007). These include continuous changes, discret change, states, and a manifold of combinations constructed out of them. There are many other global qualities of a process which are not derived from time series. Examples are the duration of a process, its temporal extension, or its occupied space. Physics provides many examples of this kind, for example the average velocity of moving body

3.2 Relational data

Relational data are based on relations. A relation is a category, the instances of which are relators. A relator is an attributive which is composed of (relational) roles. Throughout this section we use an illustrating example, the expression G := "John's drinking a beer". The subterm "drink" denotes a relation, denoted by Rel(drink). Let p be an instance of Rel (drink), then from this we may derive two roles, the role q1 of the drinker, and the role q2 of the drunken. John plays the role of the drinker and the beer plays the role of the drunken. These constituents are composed to a complex entity, a relational fact, expressed by "John`s drinking a beer"; the fact, denoted by this expression G, is denoted by Fact(G). The bearers of a relator are determined/specified by the players, which play the corresponding roles. The roles themselves occur as unary attributives, though, they cannot be separated from the relator of which they are a part.

Relators and roles are considered as attributives, being more abstract than phenomenal data, as, for example qualities. These data cannot be accessed by perception and measuring instruments. Relators can be classified with respect to the bearers; the role players may be objects or processes. The relation, connecting the roles to the players, is the inherence relation.

3.3 Relational propositions

We hold that propositions are more abstract parts of the world than facts. We restrict this question to what we call elementary relational propositions. Elementary relational propositions correspond to relational facts. We hold that the bearers of propositions are parts of the world which act as truthmakers. Let us consider the fact Fact(G), associated tot the expression G := "John's drinking a beer." By an operation of abstraction the mind transforms the fact Fact(G) into the proposition Prop(Fact(G)) := "John is drinking a beer." The modes of existence of Fact(G) and Prop(Fact(G)) are different: Fact(G) is a part of spatio-temporal reality, whereas Prop(Fact(G)) is an abstract entity the relation of which to reality is indirect, mediated by the

corresponding fact. Propositions can be satisfied or disproved, hence, they can be true or false. Relational propositions can be made true by corresponding relational facts.
Figure 1 summarizes the system of basic categories of GFO-Data(0).

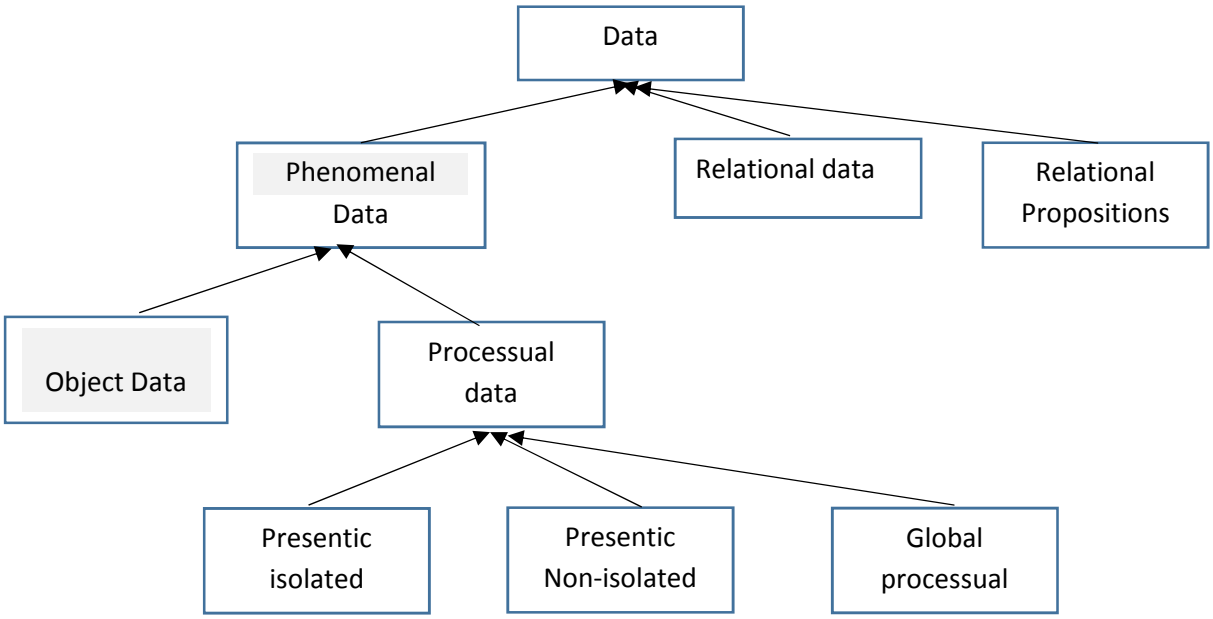

Fig 1. Categorial basic structure of GFO(0).

4. Symbolic Representation of Data

The pure semantic description of data is not sufficient for an ontology of data, because data must be represented, symbolically denoted. For this purpose we need a denotation relation. We use for this purpose the notion of a data element, as presented in the ISO-standard 11179. An ontological analysis uncovers the essense of the notion of data element, see (Uciteli, Groß, Kireyev, & Herre, 2011).

A data element has two constituents/components, a semantic one, called data element concept, and a syntactic one, called representation. Since the data element itself is a category in GFO we must clarify how its components relate to the whole. For this purpose we introduce a suitable part-of relation, called constituent-part. Data elements are GFO-categories with certain constituents.

A data element concept DEC includes an object class ObC(DEC) and a property P(DEC). An object class is a category whose instances are entities of the real world. The property being a constituent of the data element concept (DEC) can be attributed to all instances of the class ObC, i.e. every instance of ObC(DEC) has the property P(DEC)[4]. To a data element concept there is associated a uniquely determined *conceptual domain*. This conceptual domain is a set of entities which serve as the values meanings of the property P. At this place we must clarify what value meanings of a property are. Let us consider as an example the property *weight* denoted by W. The instances of this property are qualities being individual properties that inhere

[4] It should be clarified whether the selected property P(DEC) is included in the set of attributes specifying the object class ObC(DEC).

in objects which are instances of ObC. We may partition the instances of W by a measure, say, g, kg. Then, for example, 70 kg represents an equivalence class which exhibits the set of all instances of W which are measured as 70 kg. Hence, W(70kg) may be considered as a sub-property of W. But, the main point is that by using a measuring process we get a natural partition of the instances of W into equivalence classes.

The value domain is the most important part of the representation. A value domain is a set of permissible values that are represented by relators consisting of value meaning/value-pairs. The values denote the value meanings. We consider now in more detail the relations between the *conceptual domain* and the *value domain*. A conceptual domain may be represented by different/several value domains whereas every value domain represents a uniquely determined conceptual domain. Hence, if we introduce a relation Repr(VD, CD) with the meaning that VD represents CD than this relation is a many-one relation, i.e. for every VD there exists exactly one CD. Furthermore, there is also another relation repr(x,y), where x is an element of VD, and x represents an element y of CD.

The relation between value meanings and values can be ontologically specified by using the notion of the relator. We introduce a (value, meaning)-relation, briefly denoted by R_{mv}, whose instances are relators. The relators of R_{mv} are individuals with two parts, called roles, the value-role and the meaning-role. These roles inhere in the players, and the player of the meaning-role is a member of the conceptual domain, and the value-role is played by token. A token is considered in the current context as an instance of a symbol structure.
()

5. Relation between Data and Knowledge

There is a difference between data and knowledge. The boundary between both is defined by the transformation from factual data to relational propositions. A proposition has a truth-value and can be verified or satisfied by a fact. A knowledge system, also called theory, can be represented as a network of concepts which are connected by relations. On the other hand, concepts themselves might be specified by theories, which are sets of propositions. Hence, there is an interrelation between concepts and theories, they are mutually dependent.

In the general approach to concepts as theories, the instances can be defined as follows. Let T ($\sum$) be a theory, based on the signature $\sum$ (predicate- and relational symbols) which refers to a certain domain D, being a part of reality. The symbols in $\sum$ have in D an interpretation, and hence the extensions of these predicates and relations correspond to the instances of the elements of the signature. If T(Σ) is true in D then we stipulate that D is an instance of T(Σ). A theory or concept in isolation is neither true nor false, but only a proposition saying that "D is a model of T(Σ)", or "T(Σ) is true in D". We call such sentences judgements, they are meta-sentences, proposing a truth relation between a theory and a domain, being a part of reality.

6. Conclusion, remarks, and further research

In the present paper we expounded an approach for developing a semantic basis for data. These are classified with respect to three dimensions; the level of abstraction, the bearer and connecting relation, and the complexity. The current approach, to the best of our knowledge, seems to be the first systematic attempt to establish a top level ontology of data. Though, there is some related work in the field of phenotypes in biology, among them (Gkoutos, Green, &

Mallon, 2005), and (Robinson & Mundlos, 2010), and, the important and broad work by Hoehndorf on phenotypes, from which we select (Hoehndorf, 2011).

GFO-Data (0) is only the first version of an intended top level ontology of data. The next step consists in analyzing and integrating other types of data, not yet sufficiently investigated. We need, for example, a theory of visual data, an ontology of pictures etc. An unsolved problem is the development of a semantic basis for natural languages texts. A complete top level ontology of data must take into consideration the development of further ontologies, in particular, an ontology of data acquisition, an ontology of perception and observations, an ontology of measuring instruments, an ontology of verbal reports, and an ontology of scales.

Literaturverzeichnis

Barker, S. a. (2005). Endurance is paradoxical. *Analysis*, pp. 69-74.

Baumann, R., Loebe, F., & Herre, H. (2012). Ontology of time in GFO. In M. Donnelly, & G. Guizzardi, *FOIS 2012* (pp. 293–306). Amsterdam: IOS Press.

Borgo, S. &. (2010). Ontological foundations of DOLCE. In M. H. R. Poli, *Theory and Application of Ontology.* Springer.

Brentano, F. (1976). *Philosophische Untersuchungen zu Raum, Zeit und Kontinuum.* Hamburg: Felix-Meiner Verlag.

Gkoutos, G. V., Green, E. C., & Mallon, A. M. (2005). Using ontologies to describe mouse phenotypes. *Genome Biol.*

Hartmann, N. ((1935–1950).). *Ontologie (Vol. 4). Berlin: Walter de Gruyter.* Berlin: Walter de Gruyter.

Herre, H. (2010). General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In R. Poli, M. Healy, & A. Kameas, *Theory and Applications of Ontology: Computer Applications* (pp. pp. 297–345). Heidelberg: Heidelberg: Springer.

Herre, H. (2015). Change, and the Integration of Objects and Processes in the Framework of the General Formal Ontology:. In *Dynamic Being: Essays in Process-Relational Ontology.* Cambridge: Cambridge Scholars Publ.

Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F., & Michalek, H. (2007). *General Formal Ontology (GFO) – A foundational ontology integrating objects and processes [Version 1.0.1].* Leipzig: University of Leipzig, IMISE.

Hoehndorf, R. S. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*.

Loebe, F. (2007). Abstract vs. Social roles- Towards a general theoretical account of roles. *Applied Ontology*, pp. 127-158.

Poli, R. (2001, 12 (3,4)). The basic problem of the theory of levels of reality. *Axiomathes*, pp. 261–283.

Robinson, P., & Mundlos, S. (2010). The human phenotype ontology. *Clinical genetics*, pp. 525–34.

Spear, A. (2006). *Ontology for the Twenty First Century: An Introduction with Recommendations (Manual). Saarbrücken, Germany: , .* Saarbrücken: Institute for Formal Ontology and Medical Information Systems (IFOMIS), University of Saarbrücken.

Uciteli, A., Groß, S., Kireyev, S., & Herre, H. (2011). An ontologically founded architecture for information systems in clinicalo and epidemiological research. *Journal of Biomedical Semantics*, p. 2 (Suppl): S1.

Wahlberg, T. H. (2008). Can I be an instantaneous stage and yet persist through time? *Metaphysica*, pp. 235-239.