

Foundations of the Cell Tracking Ontology

Patryk BUREK^{a,1}, Nico SCHERF^{b,c} and Heinrich HERRE^d

^a*Institute of Computer Science, Faculty of Mathematics, Physics and Computer Science, Marii Curie-Skłodowskiej University, pl. Marii Curie-Skłodowskiej 5, 20-031 Lublin, Poland*

^b*Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstr. 1a, 04103 Leipzig, Germany*

^c*Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, TU Dresden, Fetscherstr. 74, 01307 Dresden, Germany*

^d*Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Haertelstr. 16-18, 04107 Leipzig, Germany*

Abstract. Time-lapse microscopy gives us insight into the dynamic behaviour of cells as they develop into tissues, organs, and entire organisms. Being able to image and analyse these multi-cellular dynamics might provide the key to understand the underlying organising principles. There is an ever increasing amount of available cell tracking data that could be systematically integrated to increase our knowledge of the organizing biological processes. This integration of heterogeneous data demands the creation of frameworks suitable for rendering extracted data searchable and interoperable between experiments and even data from complementary measurements such as single cell sequencing. In the current paper we discuss the fundamental concept of a cellular genealogy, a basic knowledge structure forming conceptual and formal foundations of the Cell Tracking Ontology (CTO) suitable for describing, querying and integrating data from complementary experimental techniques in the domain of single cell analysis. Cellular genealogies allow integration of cell tracking data, which typically consist of sets of observations of cells at single time points organized into time indexed sequences mapping the behaviour (e.g. cell division, cell death or differentiation) of individual cells and their offspring into a common reference frame. The current paper discusses the main ontological components of cellular genealogies: cells, considered as material objects, having a lifetime and persisting through time; temporal situations, composed of (continual) cells and relators connecting these cells; snapshots of cells and of temporal situations; frames identified with snapshots of temporal situations (they may be called presentic situations). All these entities are organized into cellular genealogies, forming consistent, pedigree-like structures supporting organization, querying and analysis of cell tracking data.

Keywords. Knowledge Representation and Management; Semantic annotation of images and videos; Ontologies, Live Microscopy, Cell Tracking

1. Introduction

Live microscopy is a well established, experimental technique to study the dynamics of single cells, cell colonies, tissues, organs and entire organisms [1,2,3,4]. The same sample is imaged repeatedly over an extended period of time to create a time-lapse

¹ Corresponding Author, E-mail: patryk.burek@poczta.umcs.lublin.pl. Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

movie, which is the subject of analysis in this work. Since manual analysis of such large amounts of imaging data is largely infeasible, over the past years, a lot of work has been dedicated to develop automated cell tracking methods [5, 6]. However, the lack of a standard format for storing and annotating tracking results hinders integration, and unified querying and analysis of the available data. Thus, a general and formal annotation scheme for cell tracking results is needed to support both the integration of experiment results as well as flexible and advanced analytics over the different experimental results. That requires more than a classification scheme for some specific types of entities but instead a model which supports traversing the data from a collection of cell observations into the complex knowledge structures supporting advanced queries.

Therefore, the first and most important challenge is to design a knowledge structure which supports these goals. We decided to use the most general concept of *cellular genealogies* as a foundation for the Cell Tracking Ontology (CTO) [7] which is intended for annotating cell tracking data. In the current paper we discuss ontological basis of cellular genealogies in context of the General Formal Ontology (GFO) [8] which provides the foundation for the CTO. We discuss the main ontological choices underlying the CTO i.e. Presentials vs Time-Extended Entities, Objects vs Situations. We also discuss the core notions of the CTO: cells, considered as material objects, having a lifetime and persisting through time; temporal situations, composed of (continual) cells; snapshots of cells and of temporal situations; finally cellular genealogies, which serve as the backbone model upon which the integration of specific biological ontologies into cell tracking experiments is made possible. The model organizes single cell observations into temporally extended cells and supports construction of entire developmental histories of cellular systems.

The remainder of the paper is structured as follows. Section 2 gives some introduction into the underlying framework of the GFO and into the main assumptions taken in the presented work. Sections 3 and 4 discuss the core notions of the CTO: cells, divisions and cellular genealogies as well as their ontological interpretation in terms of Presentials, Time-Extended Entities, Objects and Situations. Finally, section 5 presents conclusions, discussion and future work.

2. The Ontological Framework (GFO)

The term entity refers to anything which has a mode of existence. Entities are classified into categories and individuals. In the General Formal Ontology (GFO) the basic entities of space and time are topoids and time-intervals, also called chronoids. The ontology of space and time of GFO is inspired by ideas of Franz Brentano [10] and is presented in [11].

In the current paper we restrict the GFO-framework to the main elements relevant in the context of cell tracking experiments. We take a minimal ontological commitment and a pragmatic approach i.e. a minimal subset of GFO concepts, which is suitable for modelling the domain of cell tracking is distilled and used as a foundation of the Cell Tracking Ontology.

The main ontological categories relevant in context of the cell tracking experiments are:

- Cells, considered as material objects, having a lifetime and persisting through time.
- Temporal situations, composed of (continual) cells and relators connecting these cells.
- Snap-shots of cells and of temporal situations. Frames can be identified with snap-shots of temporal situations (they may be called presentic situations).

3. Cells as Presentials and Time-Extended Entities

Typically, the data obtained from cell tracking experiments consists of sets of observations of cells at single time points, which in the next step are organized into temporally ordered sequences. In order to represent the data adequately we introduce two distinct kinds of temporal locations: *Time Points* and *Time Periods* and establish a membership relation between them, denoted $x \in y$ meaning that a time point x is an element of a time period y . Additionally, we rely on the following relations of temporal algebra: temporal order of time temporal locations x, y , denoted $x < y$; temporal location of an entity z at some time location x , denoted $at(z,x)$.

Basing on the above we introduce the notion of Presential - an entity existing at exactly one time point and as such immutable. The term “presential” indicates the fact that presentials exist fully in the presence, which has no temporal extension, hence, happens at a time-point.

$$Pres(x) \leftrightarrow_{df} \exists ! y (TimePoint(y) \wedge at(x,y)) \quad (1)$$

Presentials are contrasted with Time Extended Entities, which we understand as entities that exist at/over a time period.

$$TimeExtEntity(x) \leftrightarrow_{df} \exists ! y (TimePeriod(y) \wedge at(x,y)) \quad (2)$$

A snapshot relation, denoted $snapshot(x,y)$, is gluing together presentials with time extended entities in a sense that presentials are snapshots of time extended entities - for each presential there exists a time extended entity which the presential is a snapshot of. Clearly, the space of possible temporal locations of a snapshot is limited to the elements of the temporal location of the entity it is a snapshot of, i.e. if x is snapshot of y , then the time point of x is an element of time interval of y . The following axioms specify the interplay between time extended entities and presentials.

$$\forall x,y (snapshot(x,y) \rightarrow Pres(x) \wedge TimeExtEntity(y)) \quad (3)$$

$$\forall x (Pres(x) \rightarrow \exists ! y (TimeExtEntity(y) \wedge snapshot(x,y))) \quad (4)$$

$$\forall x,y (snapshot(x,y) \rightarrow \exists v,w (at(x,v) \wedge at(y,w) \wedge v \in w)) \quad (5)$$

Having the above in place we focus now on analysing the content of a single cell tracking observation. The objects identified at each observation are typically cells considered as entities present at a time of observation. The time of observation is considered as temporal point. In this sense, we may classify the elements of the observation as *Presential Cells* and denote them $PresCell(x)$. *Cells*, denoted $Cell(x)$, in

contrast to presential cells, are not temporally flat objects, but instead are the entities extended in time and as such having a life time. This is the first fundamental explicit ontological distinction made in the CTO, which, if left implicit, leads to ambiguity of the notion of cell in context of cell tracking experiments. Now, having the presential cells identified we may detect cells with the help of the following formula.

$$\forall x (PresCell(x) \rightarrow \exists y (Cell(y) \wedge snapshot(x,y))) \quad (6)$$

4. Objects and Situations: Cells, Divisions and Cellular Genealogies

The second distinction applied to the CTO is made between Objects, denoted $Obj(x)$, and Situations, denoted $Sit(x)$. The notion of object is broadly used, yet it imposes some ambiguity. In context of ontology engineering (and especially in context of the discussion between 3D and 4D approaches) the object is considered as a mutable entity that persists in time. In this sense objects are contrasted mainly to processes which usually in 3D ontologies are understood as entities that depend on objects. The relation, connecting both is a kind of participation, i.e. objects can participate in processes. In contrast, in object-oriented software engineering an object is a basic unit of code and in this sense its understanding is very broad - consequently every element of a domain, including processes, is represented as an object. However, some approaches to domain modelling with object-oriented paradigm provide a more detailed classification of objects. For instance, the Domain Driven Design [12] distinguishes between mutable objects, having a sort of conceptual identity, called Entities, and Aggregates combining one or more Entities into wholes.

The above two paradigms share some common intuitions. First, in both objects have some sort of conceptual identity and, secondly, they can be organized into bigger, more complex structures which are still perceived as wholes. Basing on this intuitions we distinguish *Objects* from *Situations*. It is out of scope of the current paper to discuss the notion of identity, which has a rich literature in context of ontology engineering, (among others [13]). In our approach, as soon as we can identify an entity (by whatever rules and means) it becomes a candidate for being classified as an object and since in the CTO we are dealing only with observable physical objects the identification is made simply by visual detection.

Following the GFO we understand situations as special configurations, which can be comprehended as wholes and which satisfy certain conditions of unity. Situations are built upon individualized relations connecting objects - we say that objects *participate* in situations, which essentially are aggregating objects into comprehensible wholes. For every situation S there exists an object participating in S. Under this interpretation even a single object Obj can be considered as a situation, since Obj can be represented/understood as a whole, which is composed of individual properties/qualities, which are connected by the inherence relation. In the ontology of Tegtmeier [14] situations (called by him state of affairs) are the basic entities and all other entities (object, processes etc.) are special cases of state of affairs.

$$\forall x,y (participates_in(x,y) \rightarrow Obj(x) \wedge Sit(y)) \quad (7)$$

$$\forall x (Sit(x) \rightarrow \exists y (Obj(y) \wedge participates_in(y,x))) \quad (8)$$

The distinction drawn between objects and situations does not impose any temporal characteristics of those entities. In this approach both objects and situations can be located at time points or time periods. That means we may, in the first place, speak about *Presential Objects* and *Presential Situations* and define them as follows:

$$PresObj(x) \leftrightarrow_{df} Pres(x) \wedge Obj(x) \quad (9)$$

$$PresSit(x) \leftrightarrow_{df} Pres(x) \wedge Sit(x) \quad (10)$$

Since a situation is meant to be an aggregate of some objects which possesses some sort of unity, it cannot exist temporally detached from its participants and therefore it must be present at the same time on which they are present. For presentials that means that both a situation and its participant must be located at the same time point:

$$\forall x,y (PresSit(x) \wedge participates_in(y,x) \rightarrow PresObj(y) \wedge \exists z (at(x,z) \wedge at(y,z))) \quad (11)$$

The distinction between objects and situations can be applied to time extended entities as well. In this case we introduce two additional types of entities - *Continuant* and *Temporal Situation*. The former is a broadly used and discussed concept, here we understand it by following GFO: “*Continuants persist through time and have a lifetime; they correspond to ordinary objects, as cars, balls, trees etc. The lifetime of a continuant is presented by a time interval of non-zero duration*”[8].

$$Cont(x) \leftrightarrow_{df} TimeExtEntity(x) \wedge Obj(x) \quad (12)$$

In turn, a temporally extended situation (or temporal situation) is a system which can be comprehended as a whole and which consists of participating continuants (objects).

$$TempSit(x) \leftrightarrow_{df} TimeExtEntity(x) \wedge Sit(x) \quad (13)$$

Analogously as in case of presentials, a temporal situation must not be temporally detached from its participants, otherwise one could not speak about participation and a situation forming a comprehensive whole over its participants. However, in contrast to presentials, that condition is not simply reduced to having equal temporal extensions. For instance, an object can participate in many situations, yet it does not mean that its lifetime must be limited to the duration of the situations (as many of them can be much shorter). On the other hand, a situation can have a duration, which exceeds the lifetime of an object which participate (partially) into it. In this sense one could say that an object participates in a (temporal) part of situation. The ontological analysis of that interplay is the topic exceeding the scope of the current paper. Here we take a generic approach, in which we enforce some sort of co-occurrence between situation and its participant. That means that their temporal extensions somehow overlap at least at a single time point.

$$\forall x,y,v,w (TempSit(y) \wedge participates_in(x,y) \wedge at(x,v) \wedge at(y,w) \rightarrow \exists z (z \in v \wedge z \in w)) \quad (14)$$

Having the above in place, we may introduce the notion of *Cellular Division*, which is a well grounded concept in biology, understood as a process transforming a mother cell into two daughter cells. We may specify cell division as a temporal situation, in which there are exactly three participants i.e. one mother cell and two daughter cells. We identify those two different types of participants by means of specific participation relations: $participates_as_mother(x,y) \rightarrow participates_in(x,y)$, and $participates_as_daughter(x,y) \rightarrow participates_in(x,y)$. The two relations enable the introduction of the direct link between a mother and daughter cells.

$$\forall x (CellDivision(x) \rightarrow \exists y,v,w (Cell(y) \wedge Cell(v) \wedge Cell(w) \wedge participates_as_mother(y,x) \wedge participates_as_daughter(v,x) \wedge participates_as_daughter(w,x) \wedge y \neq v \wedge y \neq w \wedge v \neq w)) \quad (15)$$

$$\forall x,y,v,w (CellDivision(x) \wedge Cell(y) \wedge Cell(v) \wedge Cell(w) \wedge participates_as_mother(y,x) \wedge participates_as_daughter(v,x) \wedge participates_as_daughter(w,x) \rightarrow parent_cell(y,v) \wedge parent_cell(y,w)) \quad (16)$$

Cell tracking data sets usually do not contain information on cell division, however this information can be reconstructed from observations of presential cells, e.g. basing on morphological characteristics of observed cells. That way the presential snapshot of a cell division, called *Presential Cell Division* considered as presential situation, can be detected. Since observations contain always information on presential objects the presential division never links both mother and daughter participants as they never co-exist at the same time point. Instead, a presential division can be detected on a frame where either a mother or daughter cells are present.

$$\forall x (PresCellDivision(x) \rightarrow \exists y (PresCell(y) \wedge participates_as_mother(y,x) \vee participates_as_daughter(y,x))) \quad (17)$$

Next, a temporal cell division can be reconstructed from presential cell division similarly as a cell has been reconstructed from presential cells:

$$\forall x (PresCellDivision(x) \rightarrow \exists y (CellDivision(y) \wedge snapshot(x,y))) \quad (18)$$

Finally, cells and their divisions can be organized into *Cellular Genealogies*, denoted $CellGen(x)$. Cellular genealogy is interpreted as a situation, in which continuant cells and division situations that link those cells participate. It always has a structure having exactly one root and an arbitrary number of leaf cells. In cellular genealogy the root of the tree represents the founder cell and its progeny is arranged in the branches of the tree. Each branching represents a cell division.

$$\forall x (CellGen(x) \rightarrow \exists ! v \exists w (root_cell(v,x) \wedge leaf_cell(w,x))) \quad (19)$$

$$\forall x,y (root_cell(x,y) \rightarrow \neg \exists z (particp(z,y) \wedge parent_cell(z,x))) \quad (20)$$

$$\forall x,y (leaf_cell(x,y) \rightarrow \neg \exists z (particp(z,y) \wedge parent_cell(x,z))) \quad (21)$$

Cellular genealogies can be enriched by annotating additional properties measured during the experiments such as e.g. number of cell generations or shape of the tree. It also supports advanced queries such as e.g. returning undifferentiated cells whose second generation descendants were differentiated and underwent apoptosis. However, some aspects of a cell, such as shape or cell state defined by gene expression profiles can be only attributed to the presential interpretation of a cell, whereas other attributes are only well-defined for a time-extended object (e.g. motion characteristics, changes of gene expressions).

5. Conclusions, Applications and Future Work

The current paper discusses the main ontological categories underlying the Cell Tracking Ontology - an ontology designed for annotating, sharing and querying cell tracking experiment results. As discussed in [7] the Cell Tracking Ontology allows for analysis of tracking results based on the underlying model of cellular genealogies, which links the raw experimental data (i.e. presential cell measurements) and builds an expressive knowledge structure out of it. To achieve this, our model itself has a minimal ontological commitment with respect to the number of its elements. The main ontological distinctions underlying the CTO are: (1) presentials vs time extended entities, (2) objects vs situations. Basing on the above four concepts all the main components of the CTO are introduced: cells, divisions and cellular genealogies. Additionally, the ambiguity of cells understood either as time extended entities or as objects observed at time points is solved by explicit separation of cells as presentials from cells as time extended entities.

The current paper neglects one additional but important aspect of the CTO, namely the concept of *property* applicable to all of the mentioned above components of the model. This is a crucial aspect necessary for real life applications of the CTO and should be a subject of further research. Another vector of extending the core model of the CTO is the introduction of interactions between cells that go beyond the “family relationships” captured in the genealogies such as, for instance, cell-cell contact.

In the long run, the integration of the CTO with existing tools for live cell microscopy [6, 15-18] can dramatically shorten the path from the cell tracking experiment to the analysis of its results. Currently, we are working on tools, which can automatically annotate raw data sets with CTO. That is a step towards the ultimate goal, which is to support automated generation of annotated data on cells and cellular genealogies *directly* out of raw movies without the assistance of a user.

References

- [1] Berg HC, How to track bacteria. *Rev Sci Instrum* **42** (1971), 868–871.
- [2] Huisken J, Swoger J, Del Bene F, Wittbrodt J, Stelzer EHK, Optical sectioning deep inside live embryos by selective plane illumination microscopy, *Science* **305** (2004)1007–1009.
- [3] Keller PJ, Schmidt AD, Wittbrodt J, Stelzer EHK, Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy, *Science* **322** (2008),1065–1069.

- [4] McDole K, Guignard L, Amat F, Berger A, Malandain G, Royer LA, Turaga SC, Branson K, Keller PJ, In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level, *Cell* (2018) .
- [5] Maška M, Ulman V, Svoboda D, et al. A benchmark for comparison of cell tracking algorithms, *Bioinformatics* **30** (2014) 1609–1617.
- [6] Ulman V, Maška M, Magnusson KEG, et al., An objective comparison of cell-tracking algorithm, *Nat Methods* **14** (2017), 1141–1152.
- [7] Burek, P., Scherf, N., Herre, H. A pattern-based approach to a cell tracking ontology.” To appear at *KES 2019* (2019).
- [8] Herre H, Heller B, Burek P, Hoehndorf R, Loebe F, Michalek H (2006) “General Formal Ontology (GFO): A Foundational Ontology Integrating Objects and Processes. Part I: Basic Principles (Version 1.0).” Research Group Ontologies in Medicine (Onto-Med), University of Leipzig
- [10] Brentano, F. In S. Körner and R.M. Chisholm (Eds.) “*Philosophische Untersuchungen zu Raum, Zeit und Kontinuum* Hamburg: Felix Meiner-Verlag (in German), 1976.
- [11] Baumann R.; Loebe F.; Herre H. Axiomatic Theories of the Ontology of Time in GFO, *Applied Ontology* **9** (2014), 171-201.
- [12] Evans, E. *Domain-Driven Design: Tackling Complexity in the Heart of Software* , Addison-Wesley, 2004.
- [13] Guarino, N., Welty C., Ontological Analysis of Taxonomic Relationships. In Leander, A., Storey, V. (eds.): Proc. ER-2000. Springer-Verlag LNCS (2000).
- [14] Tegtmeier, E. “Facts and Connectors.” In M. Reicher (Ed.), *States of affairs* (2009) 71-82.
- [15] Eliceiri KW, Berthold MR, Goldberg IG, et al “Biological imaging software tools, *Nat Methods* **9** (2012) 697–710
- [16] Wolff C, Tinevez J-Y, Pietzsch T, et al., Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *Elife* (2018).
- [17] Pietzsch T, Saalfeld S, Preibisch S, Tomancak P., BigDataViewer: visualization and processing for large image data sets, *Nat Methods* **12** (2015) 481–483.
- [18] Meijering E, Carpenter AE, Peng H, Hamprecht FA, Olivo-Marin J-C, Imagining the future of bioimage analysis, *Nat Biotechnol* **34** (2016), 1250–1255.