Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 784–793

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

# A pattern-based approach to a cell tracking ontology.

Patryk Burek[a,*], Nico Scherf[b,d] and Heinrich Herre[c]

[a]*Institute of Computer Science, Faculty of Mathematics, Physics and Computer Science, Marii Curie-Sklodowskiej University, pl. Marii Curie-Sklodowskiej 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, Statistics and Epidemiology, University of Leipzig, Haertelstr. 16-18, 04107 Leipzig, Germany*

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

**Abstract**

Time-lapse microscopy has thoroughly transformed our understanding of biological motion and developmental dynamics from single cells to entire organisms. The increasing amount of cell tracking data demands the creation of tools to make extracted data searchable and interoperable between experiment and data types. In order to address that problem, the current paper reports on the progress in building the Cell Tracking Ontology (CTO): An ontology framework for describing, querying and integrating data from complementary experimental techniques in the domain of cell tracking experiments. CTO is based on a basic knowledge structure: the cellular genealogy serving as a backbone model to integrate specific biological ontologies into tracking data. As a first step we integrate the Phenotype and Trait Ontology (PATO) as one of the most relevant ontologies to annotate cell tracking experiments. The CTO requires both the integration of data on various levels of generality as well as the proper structuring of collected information. Therefore, in order to provide a sound foundation of the ontology, we have built on the rich body of work on top-level ontologies and established three generic ontology design patterns addressing three modeling challenges for properly representing cellular genealogies, i.e. representing entities existing in time, undergoing changes over time and their organization into more complex structures such as situations.



*Keywords:* Semantic-Based Systems; Knowledge Representation and Management; Semantic annotation of images and videos; Ontologies

* Corresponding author. *E-mail address:* patryk.burek@poczta.umcs.lublin.pl


10.1016/j.procs.2019.09.237

## 1. Introduction

In the past decades, live microscopy, i.e. imaging of the same sample over an extended period of time to create time-lapse movies, has thoroughly transformed our understanding of biological dynamics from single cells [1] to the development of entire multi-cellular organisms [2–4]. The systematic analysis of large amounts of cell tracking data such as trajectories or lineage trees [5] holds great potential but also presents a major challenge for stem cell biology [6], developmental biology [7], and theoretical biology [8]. Manual delineation and analysis of cell tracks is typically infeasible and computational methods are required for tracking, visualization, and analysis of the dynamical properties of cells. Over the past years, a large number of cell tracking methods have been published [9, 10] to tackle the problem of automated or semi-automated extraction of cell tracks. However, one major hurdle for integrating the information from different experiments into a general model of the underlying process is the lack of a standard structure for storing, annotating, and comparing cell tracking results as each tool comes with its own ad hoc format. Most tracking tools have been developed for a specialized, narrow niche of biological questions, so additional, if structured annotations of extracted cell behaviors are available at all, they are typically restricted to a few a priori fixed categories tailored to a specific question at hand. This situation limits consolidation and creation of knowledge in the field, as this requires integration of information from different experiments across laboratories or from other experimental techniques such as single cell sequencing [11]. What is needed is a more general and formal annotation scheme for cell tracking results supporting flexible and advanced analytics over the experimental results.

This poses several challenges from a modeling perspective: Cell tracking experiments focus on a wide range of scales and cellular characteristics: from assessing the motion of single cells [1], tracing the developmental histories of a few initial stem cells [12], tracking the dynamic changes of cell ensembles such as colony formation [13] or wound healing [14] to observing the formation of entire organisms at the single cell level [4]. This multitude of characteristics and levels of description lead to our major question: Is possible to query such diverse experimental data in a consistent way? The challenge is to make cell tracking data interoperable and searchable, which includes both searching for experiments themselves and searching for patterns inside each single data set. As pointed out in [15] this goes beyond the problem of data integration across different formats in the light of inherent problems such as pluralism (same name is used for different processes or the same process has different names). The underlying terminology and formal concepts are of utmost importance and could themselves be regarded as theories about the biological world.

Ontologies are an ideal tool for that task, providing a formal, explicit specification of shared conceptualization [16] and can be supplied with a rich infrastructure such as languages for ontology representation (e.g. the Resource Description Framework (RDF) [17]) and inference (Web Ontology Language (OWL) [18]) or querying (SPARQL Protocol and RDF Query Language (SPARQL) [19]). In context of biology there already exists a rich body of ontologies organized within the Open Biological and Biomedical Ontology (OBO) Foundry [20]. There are ontologies for representing experiments as such, e.g. Ontology for Biomedical Investigations (OBI) [21], ontologies for annotating cells e.g. Cell Ontology [22], and cell characteristics and behaviors, e.g. Phenotype and Trait Ontology (PATO) [23] or Cell Behavior Ontology [24]. However, as already noted in [25], according to our best knowledge there is no single ontology adequate for representing cell tracking experiments and annotating single images or entire movies obtained from these experiments. The existing ontologies support the annotation of some aspects of the domain e.g. cellular properties, or cell cycle but not the cellular genealogies which are a core knowledge structure for cell-tracking experiments in stem cell and developmental biology [6].

The current paper reports on the initial progress in building a tool for describing and querying cell tracking data and integrating data from complementary experimental techniques that connects to existing cell ontologies. The core component of the tool is the Cell Tracking Ontology (CTO) which provides a basic knowledge structure: the cellular genealogy which serves as the backbone model to integrate specific biological ontologies into cell tracking experiments. The cellular genealogy is a natural and generic reference frame to model the dynamic processes extracted from cell tracking experiments. It allows to organize single cell observations into temporally extended cells and also entire developmental histories of a cellular system (e.g. stem cells and its progeny in a specific niche) and opens rich opportunities for querying and analysis of experiments at the single cell level. As a first step to integrate complementary ontologies to enrich the available annotations and facilitate knowledge transfer between

different types of experiments and data, we integrate the PATO ontology as one of the most important in the context of cell tracking experiments into CTO.

## 2. Cellular Genealogies

The data obtained in cell tracking experiments typically consist of sets of observations of cells at single time points organized into time indexed sequences (a synthetic tracking experiment is shown in Fig.1). Therefore, the first challenge is to structure and organize the data for analysis. In [14] we proposed to organize the information about a cell's history into pedigree-like data structures called *cellular genealogies*, in which 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. The challenge here is to model the processes and entities that underly the cellular genealogy in a sound way. Interestingly, an analysis of the concepts underlying the interpretation of cell tracking results reveals several ambiguities in the notion of what a cell is: A cell is either considered (1) as completely present at a time-point having no temporal parts (see individual snapshots of cells at example time points in Fig. 1b) or (2) as a persisting object having dynamic qualities (see tracks in Fig. 1c) or (3) as a time extended process, which evolves through time and is never wholly present at a time point. The very structure of a cellular genealogy, as depicted in a radial tree layout in Fig. 1d), makes these distinctions explicit. Within the given genealogy the radial lines represent cells understood as objects extended in time, with cell divisions marking the boundaries of a cell's existence shown as tangential arcs. The radial dimension indicates time, i.e. the length of the radial lines represents the duration of the continual cell's existence and is a measure of the cell cycle time. However, the temporally extended cell is actually only measured and quantified at distinct individual time points as shown for two snapshots in Fig. 1b. Additionally, the entire developmental history itself, as summarized by the genealogical tree is a temporally extended process from the founding cell as the root at the center (bottom) to the outermost leaf cells. These genealogies can be enriched by annotating additional properties measured during the experiments. 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) or even only at the level of the entire genealogy (e.g. "stemness" - the property of being a stem cell, c.f. [26]).

## 3. Basic Design Patterns of the Cellular Tracking Ontology

The development of a well-defined ontology is a non-trivial enterprise. In the last two decades numerous methodologies, theoretical frameworks, and rule sets fostering this process have been introduced [27]. For the development of the CTO we have adapted two such tools, namely, top level ontologies and based on those, ontology design patterns. Top level ontologies, in contrast to domain and application ontologies, specify most generic, cross-domain and multi-purpose conceptualizations [28–31] and as such serve well for ontology development [30], ontology integration [32] and refactoring [33]. Additionally, they are well suited as a basis for highly reusable ontology patterns, called foundation design patterns [34]. The idea of ontology design pattern is inspired by the established concept of software design patterns, i.e. a generalized solution for an often occurring engineering problem [35]. OWL ontology design patterns provide generic reusable OWL snippets that can be used directly, e.g. via import into the ontology under development [28]. The application of patterns not only saves time during ontology engineering, but also assures a better quality of the developed ontology [36, 37].

The CTO ontology is designed as an integration ontology which allows integrating specific ontologies into one consistent and expressive model, which itself is minimal with respect to the number of its elements. To achieve this, CTO requires integration of data on various levels of generality and proper structuring of collected information. Thus we rely on the principle of minimal commitment [38], which is especially valid in cases where numerous ontologies are expected to be integrated.

The domain of cell tracking experiments can be decomposed into three levels:

- Level 1 consisting of cells considered as objects observed at given time points.
- Level 2 consisting of cells understood as objects extended in time (i.e. across several time points), having some dynamics involving changes of qualities and engagement in multi-cell processual relations such as cell-cell contacts or cell divisions.

- Level 3 consisting of cellular genealogies, which are time extended objects composed of objects of the first and the second level.

The representation of those levels can be reduced to three generic modeling problems i.e. (1) the representation of entities over time, (2) the representation of their quality changes over time, and (3) the representation of situations understood as collections of entities and their relations. All three topics are broadly discussed in the context of top-level ontologies. Based on the results obtained in [28–31] we have developed three patterns: (1) the Temporal Entities Pattern (TEP), (2) the Temporal Qualities Pattern (TQP) and (3) the Situations Pattern (SP) and used them as a foundation to develop CTO.

### *3.1. Temporal Entities Pattern*

When considering time lapse experiments, we see that only objects of the first level are directly observed and present in raw data sets. From the biologist's point of view however, the most interesting are objects of the second and the third level. Therefore, the first challenge for the design of CTO is to enable the linkage of object of all the three levels in order to support traversing from the first level objects to those of the second and third level. We use two design patterns for that purpose, based on the General Formal Ontology (GFO) [29]: the Temporal Entities Pattern and the Situations Pattern and the Situations Pattern. Fig.2 (left) depicts the Temporal Entities Pattern which is based on a simple distinction between two types of entities: the entities fully present on a single time point and those extended in time having a lifespan. The former we call Presentials (*Pres*(*x*)) and the latter Temporal Entities (*TempEnt*(*x*)). The notion of Presential is defined in GFO as follows: *"A presential is an individual which is entirely present at a time-point. The introduction of the term "presential" is motivated by the fact that presentials are individuals that may exist in the presence, where we assume that the presence has no temporal extension, hence, happens at a time-point."* [39] p. 309.

In contrast, Temporal Entities are defined as entities which are extended in time and as such not wholly present at any given single time point. The glue, which links presentials and temporal entities, is a snapshot relation (*is_snapshot_of*(*x*,*y*)), where *x* is a presential, *y* is a temporal entity and the relation indicates that *x* is a temporal snapshot of *y*. In this sense, presentials are reified temporal snapshots of entities extended in time. Temporal entities are broadly analyzed in ontology engineering and many ontologies provide principles of organizing them. A particular type of temporal entities are tangible, physical objects which often are referred as Endurants [28] or Continuants [31]. In the Temporal Entities Pattern we adapt the term Continuant (*Cont*(*x*)), to denote tangible, independent Temporal Entities. In turn, a snapshot counterpart of a Continuant is called a Presential Object (*PresObj*(*x*)). Then the following axiom holds:

$$\forall x,y\ (\mathrm{PresObj}(x) \wedge \mathrm{is\_snapshot\_of}(x,y) \rightarrow \mathrm{Cont}(y)). \quad (1)$$

A finite collection *C*(*E*) of snapshots of a temporal entity E can be ordered temporally and therefore we introduce the temporal succession relation (*has_next*(*x*,*y*)) into the pattern, meaning that the presential *y* temporally follows the presential *x*. Finally, both the first and the last snapshot in of a Temporal Entity *E* are distinguished as First Presential and Last Presential of *E* in *C*(*E*): a presential cell *E* is a first (resp. last) presential if there exists a cell *C* of which *E* is the first (resp. last) presential.

### *3.2. Temporal Qualities Pattern*

To model objects which change in time in CTO, we introduced the Temporal Qualities Pattern. The pattern is motivated by the observation that, as discussed in [40], a straightforward approach to modeling qualities in OWL cannot model the change of qualities adequately. Typically, the entities that have qualities are modeled as OWL Classes and their qualities as OWL Classes or Datatypes. Quality Ascriptions which link quality owners with their qualities are modelled as Object Properties or Datatype Properties, respectively. For instance, one can model a shape of a cell as owl:ObjectProperty named has_shape, linking an owl:Class Cell with an owl:Class Shape. Unfortunately, this approach does not allow to represent changes of qualities over time, e.g. the change of a cell's shape from round to elongated. In order to solve this limitation, we have introduced the Temporal Qualities Pattern

(Fig.2, middle). An arbitrary quality is depicted by the Quality class. An entity having a quality is called a Quality Owner and the straightforward quality assignment is introduced by the has_quality property. In situation where a quality owner is an entity located in time, i.e. Presential or Temporal Entity, the temporal location of a quality assignment is equal to the temporal location of its owner. Yet, in situations where a quality can change over time the Quality Assignment (QA) class is utilized. QA is a reified has_quality property, which itself has a temporal location. The combination of both patterns allows modeling of the following types of quality assignments:

1. Qualities observed at a single time point - by means of presential and has_quality property,
2. Static, non-changing Qualities of enduring entities - by using Temporal Entity and has_quality property,
3. Dynamic, changing Qualities of enduring entities - by using reified temporally indexed Quality Assignments.

### *3.3. Situation Pattern*

The third level of entities reconstructed from raw data of cell tracking experiments are cellular genealogies, which are complex structures comprising multiple cells linking them by relations such as cell divisions. Such complex structures are often referenced as situations [29, 41] and we adhere to this term and interpretation from GFO: "*A situation is a special configuration which can be comprehended as a whole and satisfies certain conditions of unity, which are imposed by relations and categories associated with the situation.*" [39]. In the context of cell tracking experiments we are interested in situations, which involve tangible objects such as cells. The Situations Pattern (Fig.2, right) distinguishes two types of situations depending on their temporal location: Presential Situations (*PresSit*(*x*)) and time-extended Situations (*Sit*(*x*)). Both are complex entities composed of entities participating in them. For Situations and Presential Situations the following axioms hold:

$$\forall x\ (\mathrm{PresSit}(x) \rightarrow \mathrm{Pres}(x) \wedge \exists y\ (\mathrm{PresObj}(y) \wedge \mathrm{particpates\_in}(y,x))). \tag{2}$$

$$\forall x,y\ (\mathrm{PresSit}(x) \wedge \mathrm{PresObj}(y) \wedge \mathrm{particpates\_in}(y,x) \rightarrow \exists w,v\ (\mathrm{Cont}(w) \wedge \mathrm{Sit}(v) \wedge \mathrm{particpates\_in}(w,v) \wedge \mathrm{is\_snapshot\_of}(x,v) \wedge \mathrm{is\_snapshot\_of}(y,w))). \tag{3}$$

## 4. Cell Tracking Ontology

### *4.1. Elicitation of Cellular Genealogies*

Before we demonstrate the rich potential of the CTO for querying and analyzing cellular genealogies, we address the problem that raw data usually contains only presential information and no explicit representation of cellular genealogies or even time extended cells. Below, we present a selection of axioms which allow to elicit information about time extended cells and cellular genealogies out of minimal input data given in the following form: [presential_cell_id, previous_presential_cell_id]. This allows us to enrich data with information about a position of a presential cell in a sequence of cells:

$$\forall x\ (\mathrm{PresCell}(x) \wedge \neg\exists y\ (\mathrm{PresCell}(y) \wedge \mathrm{has\_next}(y,x)) \rightarrow \mathrm{FirstPresCell}(x)). \tag{4}$$

$$\forall x\ (\mathrm{PresCell}(x) \wedge \neg\exists y(\mathrm{PresCell}(y) \wedge\ \mathrm{has\_next}(x,y))\ \rightarrow \mathrm{LastPresCell}(x)). \tag{5}$$

$$\forall x,y,z\ (\mathrm{PresCell}(x) \wedge \mathrm{PresCell}(y) \wedge \mathrm{PresCell}(z) \wedge \mathrm{has\_next}(x,y) \wedge \mathrm{has\_next}(x,z) \wedge y \neq z \rightarrow \mathrm{LastPresCell}(x) \wedge \mathrm{FirstPresCell}(y) \wedge \mathrm{FirstPresCell}(z)). \tag{6}$$

Based on First Presential Cell, Last Presential Cell and has_next relation the presential cells can be organized into a sequence, called Presential Cell Sequence (*PresCellSeq*(*x*)):

$$\mathrm{PresCellSeq}(L) = L \text{ is a List} \wedge \forall x\ (x \in L \rightarrow \mathrm{PresCell}(x)) \wedge$$

$$\begin{aligned}&\text{FirstPresCell(Head(L))} \wedge \\ &\exists!y(y \in L \wedge \text{FirstPresCell}(y)) \wedge \\ &\text{LastPresCell(Last(L))} \wedge \\ &\exists!z\,(z \in L \wedge \text{LastPresCell}(z)) \wedge \\ &\text{for every prefix w of L and v s.t. } L = w + v \text{ holds: has\_next(Last(w),First(v))}. \quad (7)\end{aligned}$$

We assume that every presential cell belongs to exactly one sequence. The sequence of presential cells provides a foundation for the reconstruction of a *Cell*(*x*) considered as a time extended entity, whose snapshots are the presential cells as the elements of the sequence:

$$\forall x\,(\text{PresCell}(x) \rightarrow \exists!L\,(\text{PresCellSeq}(L) \wedge x \in L). \quad (8)$$

$$\forall L\,(\text{PresCellSeq}(L) \rightarrow \exists!z\,(\text{Cell}(z) \wedge \forall x\,(x \in L \rightarrow \text{is\_snapshot\_of}(x,z))). \quad (9)$$

The following axiom allows to deduce qualities of temporally extended cells and time extended quality assignments from presential cells:

$$\begin{aligned}&\forall x,y,z\,(\text{has\_quality}(x,y) \wedge \text{is\_snapshot\_of}(x,z) \rightarrow \\ &\exists w\,(\text{QualityAssign}(w) \wedge \text{has\_quality\_assign}(z,w) \wedge \text{of\_quality}(w,y))). \quad (10)\end{aligned}$$

Finally, temporally extended cells can be organized into cellular genealogies as a sequence of cells linked by mother-daughter relation reconstructed from branching information present in raw data.

$$\begin{aligned}&\forall x,y,v,w\,(\text{LastPresCell}(x) \wedge \text{FirstPresCell}(y) \wedge \text{has\_next}(x,y) \wedge \\ &\text{is\_snapshot\_of}(x,v) \wedge \text{is\_snapshot}(y,w) \rightarrow \text{has\_daughter\_cell}(v,w)). \quad (11)\end{aligned}$$

The distinguished cells in a genealogy called Root Cell and Leaf Cell are specified as follows:

$$\forall x\,(\text{Cell}(x) \wedge \neg\exists y\,(\text{Cell}(y) \wedge \text{has\_daughter\_cell}(y,x)) \rightarrow \text{RootCell}(x). \quad (12)$$

$$\forall x\,(\text{Cell}(x) \wedge \neg\exists y\,(\text{Cell}(y) \wedge \text{has\_daughter\_cell}(x,y)) \rightarrow \text{LeafCell}(x). \quad (13)$$

Basing on the two above notions and the has_daughter_cell relation a sequence of Cells, denoted *CellSeq*(*x*) can be introduced. This, in turn, serves as a foundation for deduction of a cellular genealogy.

$$\begin{aligned}&\text{CellSeq(L)} = \text{L is a List} \wedge \\ &\quad \forall x\,(x \in L \rightarrow \text{Cell}(x)) \wedge \\ &\quad \text{RootCell(Head(L))} \wedge \\ &\quad \exists!y(y \in L \wedge \text{RootCell}(y)) \wedge \\ &\quad \text{LeafCell(Last(L))} \wedge \\ &\quad \exists!z(z \in L \wedge \text{LeafCell}(z)) \wedge \\ &\quad \text{for every prefix of L w and v s.t. } L = w + v \text{ holds has\_daughter\_cell(Last(w),First(v))}. \quad (14)\end{aligned}$$

$$\forall x\,(\text{Cell}(x) \rightarrow \exists!L\,(\text{CellSeq}(L) \wedge x \in L). \quad (15)$$

$$\forall L\,(\text{CellSeq}(L) \rightarrow \exists!z\,(\text{CellGen}(z) \wedge \forall x(x \in L \rightarrow \text{participates\_in}(x,z))). \quad (16)$$

The application of the above set of axioms builds the enriched knowledge base directly from raw input of only presential objects and allows querying across all three levels from presential cells, temporally extended cells to cellular genealogies.

### *4.2. Ontology Structure and Query Capabilities*

The Cell Tracking Ontology is an application ontology targeted to support analysis of cell tracking experiments. Since the biomedical domain has numerous specialized ontologies, we decided to provide a simple model which supports analysis of cellular genealogies based on raw presential data and leverages the expressiveness of existing biomedical ontologies. As a first step we have integrated PATO ontology providing a rich vocabulary to describe qualities of cells. The sound backbone of the CTO ontology is provided by the patterns discussed above, which adapted to the domain of cell tracking yields a simple, but expressive model consisting of: (1) presential cells; (2) presential cellular situations such as cell contact, cell death, cell division; (3) cells considered as time extended entities; (4) cellular situations such as cell contact, cell death, cell division; (4) cellular genealogies considered as situations and (6) cellular quality assignments, which support representation of quality changes over time.

The combination of the discussed patterns and the PATO gives the CTO expressiveness to handle queries on entities present in raw data i.e. presential cells annotated with PATO concepts. For instance, the SPARQL listing below returns the position of all differentiated cells where PATO:0002099 identifies a differentiated cell and PATO:0000140 a location of a cell:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX hs: <http://www.onto-med.de/ontologies/HappeningsSituations/0.9#>
PREFIX tq: <http://www.onto-med.de/ontologies/TemporalQualities/0.9#>
PREFIX te: <http://www.onto-med.de/ontologies/TemporalEntities/0.9#>
PREFIX cto:<http://www.onto-med.de/ontologies/cto/0.9#>
PREFIX PATO: <http://purl.obolibrary.org/obo/pato.owl#>

SELECT ?location ?presential_cell
WHERE
{
    ?presential_cell a cto:PresentialCell .
                ?presential_cell tq:has_quality  ?location .
    ?location a PATO:0000140 .
    ?presential_cell tq:has_quality PATO:0002099 .
}
```

CTO allows also more sophisticated queries on cellular genealogies: e.g. the listing below returns states/potency of all daughter cells of undifferentiated cells together with the number of occurrences of each daughter cell grouped by state/potency. PATO::0001397 is a potency class grouping all potencies and PATO:0002100 is undifferentiated potency.

```
SELECT ?potency (COUNT (?potency) ?countPotency)
WHERE
{
    ?daughter_cell a cto:Cell .
    ?mother_cell a cto:Cell .
    ?daughter_cell tq:has_quality ?potency .
    ?potency a PATO::0001397 .
    ?mother_cell cto:has_daughter_cell ?daughter_cell .
    ?mother_cell tq:has_quality PATO:0002100 .
}
GROUP BY ?potency .
```

Finally, it is possible to query for changing properties of cells or genealogies: the query below returns undifferentiated, growing cells (i.e. cell size changes from normal or decreased to increased) which differentiate (i.e. divide into differentiated cells). The PATO concepts PATO:0002100, PATO:0002099, PATO:0045050,

PATO:0000587, PATO:0000586 identify undifferentiated and differentiated potency, normal size, decreased size and increased size respectively.

```
SELECT ?cell
WHERE
{
    ?cell a cto:Cell .
    ?daughter_cell a cto:Cell .
    ?cell tq:has_quality ?potency .
    ?potency a PATO::0001397 .
    ?cell cto:has_daughter_cell ?daughter_cell .
    ?daughter_cell tq:has_quality PATO:0002099 .
    ?cell tq:has_quality_assignment ?qa1 .
    {?qa1 tq:of_quality PATO:0045050 } UNION  {?qa1 tq:of_quality PATO:0000587} .
    ?cell tq:has_quality_assignment ?qa2 .
    ?qa2 tq:of_quality PATO:0000586 .
    ?qa1 te:has_next ?qa2 .
}
```

## 5. Conclusions and Future Work

The current paper reports on the development of the Cell Tracking Ontology dedicated to describing and querying cell tracking results and integrating data from complementary experimental techniques. The CTO ontology facilitates integration of ontologies focusing on different aspects of a cell, such as being observed at a single time point, being a temporally extended object undergoing changes and finally being part of a bigger developmental process. All those aspects are organized into one consistent and expressive model. To achieve this, our model itself has a minimal ontological commitment with respect to the number of its elements. In the present paper we have demonstrated the rich query and analysis potential of the CTO by integrating as the first step a single bio-ontology: PATO. In the next steps we will integrate the CTO with other biological ontologies extending the expressiveness and querying potential over the biological domain, and also with the OWL-Time [42] increasing the expressiveness with respect to temporal information and reasoning. That will open further possibilities, such as inference of potential models of cell organization over time.

The CTO requires both the integration of data on various levels of generality and the proper structuring of collected information. Here, we have used the theoretical model of cellular genealogies - a pedigree-like structures supporting organization, querying and analysis of cell tracking data as the structuring principle. To provide a sound foundation of the ontology and using the rich body of work on top-level ontologies, we have established three generic ontology design patterns: (1) the Temporal Entities Pattern suited for the representation of entities over time, (2) the Temporal Qualities Pattern suited for the representation of quality changes over time and (3) the Situations Pattern suited for the representation of situations understood as collections of entities and their relations.

To make our approach as general as possible and alleviate the problem of inconsistent and non-standard data formats in cell tracking tools, we have also addressed the aspect that raw data does not necessarily need to explicitly contain all the information on cellular genealogies. We have demonstrated a selection of axioms, which support structuring of information on cells considered as temporally extended entities as well as cellular genealogies basing on raw data only. We are currently implementing these as a computational tool to elicit that information from raw results. The ultimate goal is to support automated generation of annotated data on cells and cellular genealogies out of raw movies by integrating automated cell tracking tools [10] and abstract representations/concepts learned by deep neural networks [43]. One important challenge here will be to quantify and represent the intrinsic uncertainty in automated (and manual) cell tracking. To systematically map the influence of tracking errors on downstream data analysis and knowledge extraction we will concentrate on simulated cell tracking experiments, where the mechanistic model behind the observed dynamics is completely known. It will be interesting to investigate, if the underlying mechanisms (e.g. differentiation rules) of the computational cell model can be recovered from simulated tracking results annotated with our scheme. That in turn will support systematic analysis of how sensitive the knowledge extraction is to the level of uncertainty and errors in automated cell tracking.

In the long run, the integration of the CTO with existing tools for live cell microscopy [10, 44-47] can dramatically shorten the path from the cell tracking experiment to the analysis of its results. We intend not only to annotate the content of the tracking result but also the experiment settings us such, which will support search and query of data sets. For instance, integrating the CTO with the Ontology of Biomedical Investigations might allow us to leverage complementary information from other experimental sources to support cross-experiment analysis, consolidation of knowledge and meta-analysis. This would be an important step to leverage the large amounts of heterogeneous data from single cell analysis to inform our understanding of cell biology.

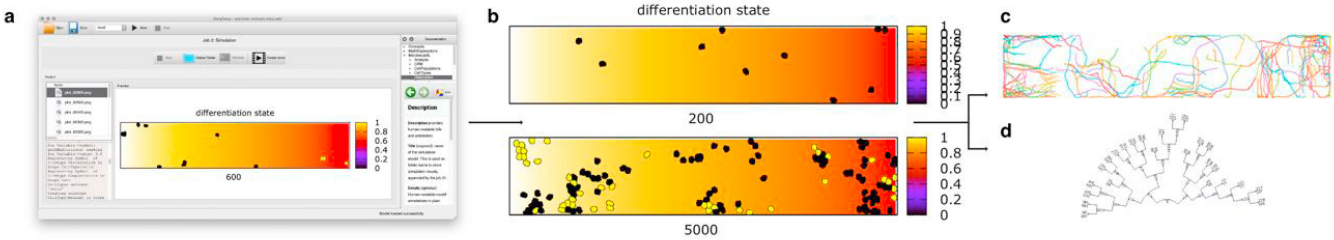


Fig. 1. (a) Simulation of single cell dynamics [48] creates virtual time lapse experiments. (b) Two example time points of a simulated experiments (color code shows differentiation state of cells (black - stem cells, yellow- differentiated cells) and external molecular signal gradient (white - low, red - high). Extracted information from time lapse movies: (c) cell tracks and (d) a sample cellular genealogy.

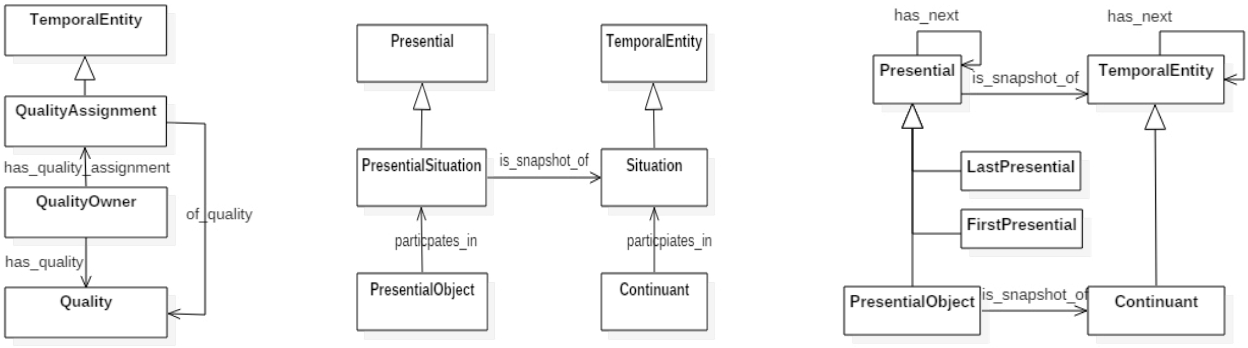


Fig. 2. UML diagrams presenting the three design patterns underlying the Cell Tracking Ontology. (Left) Diagram of the Temporal Entities Pattern, (Centre) - the Temporal Qualities Pattern (Right) the Situations Pattern.

## References

[1] Berg HC (1971) "How to track bacteria." *Rev Sci Instrum* 42:868–871
[2] Huisken J, Swoger J, Del Bene F, Wittbrodt J, Stelzer EHK (2004) "Optical sectioning deep inside live embryos by selective plane illumination microscopy." *Science* **305**:1007–1009
[3] Keller PJ, Schmidt AD, Wittbrodt J, Stelzer EHK (2008) "Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy." *Science* **322**:1065–1069
[4] McDole K, Guignard L, Amat F, Berger A, Malandain G, Royer LA, Turaga SC, Branson K, Keller PJ (2018) "In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level." Cell. doi: 10.1016/j.cell.2018.09.031
[5] Glauche I, Lorenz R, Hasenclever D, Roeder I (2009) "A novel view on stem cell development: analysing the shape of cellular genealogies." *Cell Prolif* **42**:248–263
[6] Schroeder T (2008) "Imaging stem-cell-driven regeneration in mammals." *Nature* **453**:345–351
[7] Scherf N, Huisken J (2015) "The smart and gentle microscope." *Nat Biotechnol* **33**:815–818
[8] Myers G (2012) "Why bioimage informatics matters." *Nat Methods* **9**:659–660
[9] Maška M, Ulman V, Svoboda D, et al (2014) "A benchmark for comparison of cell tracking algorithms." *Bioinformatics* **30**:1609–1617
[10] Ulman V, Maška M, Magnusson KEG, et al (2017) "An objective comparison of cell-tracking algorithms." *Nat Methods* **14**:1141–1152

[11] Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF (2018) "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis." Science. doi: 10.1126/science.aar3131
[12] Hoppe PS, Schwarzfischer M, Loeffler D, et al (2016) "Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios." *Nature* **535**:299–302
[13] Scherf N, Herberg M, Thierbach K, Zerjatke T, Kalkan T, Humphreys P, Smith A, Glauche I, Roeder I (2012) "Imaging, quantification and visualization of spatio-temporal patterning in mESC colonies under different culture conditions." *Bioinformatics* **28**:i556–i561
[14] Sunyer R, Conte V, Escribano J, et al (2016) "Collective cell durotaxis emerges from long-range intercellular force transmission." *Science* **353**:1157–1161
[15] Leonelli S (2019) "The challenges of big data biology." Elife. doi: 10.7554/eLife.47381
[16] Guarino N, Oberle D, Staab S (2009) "What Is an Ontology?" In: Staab S, Studer R (eds) *Handbook on Ontologies*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–17
[17] Cyganiak R, Wood D, Lanthaler M, Klyne G, Carroll JJ, McBride B (2014) "RDF 1.1 concepts and abstract syntax." W3C recommendation 25:
[18] W3C OWL Working Group (2012) "OWL 2 Web Ontology Language Document Overview (Second Edition)." In: W3C Recommendation. http://www.w3.org/TR/owl2-overview/.
[19] Prud'hommeaux E, Seaborne A, Others (2017) "SPARQL query language for RDF. W3C Recommendation (2008)."
[20] Smith B, Ashburner M, Rosse C, et al (2007) "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." *Nat Biotechnol* **25**:1251–1255
[21] Bandrowski A, Brinkman R, Brochhausen M, et al (2016) "The Ontology for Biomedical Investigations." *PLoS One* **11**:e0154556
[22] Diehl AD, Meehan TF, Bradford YM, et al (2016) "The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability." *J Biomed Semantics* **7**:44
[23] Gkoutos G (2019) "Phenotype and Trait Ontology." http://purl.obolibrary.org/obo/pato.owl.
[24] Sluka JP, Shirinifard A, Swat M, Cosmanescu A, Heiland RW, Glazier JA (2014) "The cell behavior ontology: describing the intrinsic biological behaviors of real and model cells seen as active agents." *Bioinformatics* **30**:2367–2374
[25] Scherf N, Kunze M, Thierbach K, Zerjatke T, Burek P, Herre H, Glauche I, Roeder I (2013) "Assisting the Machine Paradigms for Human-Machine Interaction in Single Cell Tracking." In: Meinzer H-P, Deserno TM, Handels H, Tolxdorff T (eds) *Bildverarbeitung für die Medizin 2013*. Springer Berlin Heidelberg, pp 116–121
[26] Herre H (2019) "Towards a New Foundational Ontology of Properties, Attributives and Data." In: *Ontology Makes Sense - Essays in honor of Nicola Guarino*. pp 194–210
[27] Staab S, Studer R (2009) Handbook on Ontologies. Springer Berlin Heidelberg
[28] Presutti V, Gangemi A (2008) "Content Ontology Design Patterns as Practical Building Blocks for Web Ontologies." In: *Conceptual Modeling - ER 2008*. Springer Berlin Heidelberg, pp 128–141
[29] 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
[30] Arp R, Smith B, Spear AD (2015) Building Ontologies with Basic Formal Ontology. MIT Press
[31] Sowa JF (1999) Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA
[32] Schulz S, Boeker M, Stenzhorn H (2008) "How Granularity Issues Concern Biomedical Ontology Integration." In: *eHealth Beyond the Horizon - Get IT There, Proceedings of MIE2008, The XXIst International Congress of the European Federation for Medical Informatics, Göteborg, Sweden, May 25-28, 2008*. pp 863–868
[33] Burek P, Loebe F, Herre H (2015) "A UML profile for functional modeling applied to the Molecular Function Ontology." In: *Proceedings ICBO*. pp 12–16
[34] Falbo RA, Guizzardi G, Gangemi A, Presutti V (2013) "Ontology patterns: clarifying concepts and terminology." In: *Proceedings of the 4th International Conference on Ontology and Semantic Web Patterns - Volume 1188*. CEUR-WS.org, pp 14–26
[35] Johnson R, Gamma E, Vlissides J, Helm R (1995) Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley
[36] Gangemi A, Presutti V (2009) "Ontology Design Patterns." *Handbook on Ontologies* 221–243
[37] Blomqvist E, Gangemi A, Presutti V (2009) "Experiments on Pattern-based Ontology Design." In: *Proceedings of the Fifth International Conference on Knowledge Capture*. ACM, New York, NY, USA, pp 41–48
[38] Gruber TR (1993) "A translation approach to portable ontology specifications." *Knowledge Acquisition* **5**:199–220
[39] Herre H (2010) "The Ontology of Mereological Systems: A Logical Approach." In: Poli R, Seibt J (eds) *Theory and Applications of Ontology: Philosophical Perspectives*. Springer Netherlands, Dordrecht, pp 57–82
[40] Burek P, Scherf N, Herre H (2014) "OWL Patterns for Modeling the Change over Time exemplified by the Cell Tracking Ontology." ONTOLOGIES AND DATA IN LIFE SCIENCES (ODLS 2014)
[41] Barwise J (1987) "Situations and small worlds." In: *The Situation in Logic*. pp 79–92
[42] Cox S, Little C (2017) "Time ontology in OWL." In: W3C Recommendation. https://www.w3.org/TR/owl-time/.
[43] Kan A (2017) "Machine learning applications in cell image analysis." Immunology and Cell Biology
[44] Eliceiri KW, Berthold MR, Goldberg IG, et al (2012) "Biological imaging software tools." *Nat Methods* **9**:697–710
[45] Wolff C, Tinevez J-Y, Pietzsch T, et al (2018) "Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb." Elife. doi: 10.7554/eLife.34410
[46] Pietzsch T, Saalfeld S, Preibisch S, Tomancak P (2015) "BigDataViewer: visualization and processing for large image data sets." *Nat Methods* **12**:481–483
[47] Meijering E, Carpenter AE, Peng H, Hamprecht FA, Olivo-Marin J-C (2016) "Imagining the future of bioimage analysis." *Nat Biotechnol* **34**:1250–1255
[48] Starruß J, de Back W, Brusch L, Deutsch A (2014) "Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology." *Bioinformatics* **30**:1331–1332