# On the Ontological Foundations of Cellular Development

Patryk BUREK[a] and Nico SCHERF[b, c, 1] and Heinrich HERRE[d, 1]
[a] *Institute of Computer Science, Faculty of Mathematics, Physics and Computer Science, Marii Curie-Sklodowskiej University, Lublin, Poland,*
[b] *Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, School of Medicine, TU Dresden, Dresden, Germany,*
[c] *Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*
[d] *Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany*

**Abstract.** Time-lapse microscopy is a principal tool to unravel the mystery of how cells form and maintain organisms. The complexity of the domain of cellular dynamics demands a conceptual architecture as a solid theoretical foundation that supports the integration of knowledge obtained across experiments and theories. In this work, we outline the ontological foundation of cellular genealogies, a key concept for describing and representing of cellular development. We build the conceptual framework following the onto-axiomatic method: We first analyse the domain within the context of a top-level ontology (GFO). The resulting domain-specification provides the basis for a conceptualisation where we introduce concepts and relations. From these conceptualisations, we then construct model-structures adhering to the principles of model-theory. We finally elaborate axioms based on these model-structures. The developed framework provides the fundamental concepts underlying a Cell Tracking Ontology (CTO) that supports extraction and integration of biological knowledge from systems-level experiments across different types of observations at the single-cell level.

**Keywords.** Knowledge management, Ontology of biological reality, Theories of Developmental Biology, Microscopy, Time-lapse imaging, Cell tracking

## 1. Introduction

Cellular dynamics unfolding in space and time organise and shape multicellular life as it develops from a single fertilised egg into a complex organism. After development, cellular processes maintain the organism during its lifetime (tissue homeostasis and regeneration). To fully understand how cells build and maintain structures, we have to be able to observe cellular dynamics and cellular states from experiments [1]. One milestone was the reconstruction of the embryonic lineage tree of the nematode Caenorhabditis elegans using microscopy [2]. From these roots, modern fluorescence microscopy has turned into a powerful tool to resolve the dynamics of thousands of cells

[1] Corresponding Authors: nico.scherf@tu-dresden.de, heinrich.herre@imise.uni-leipzig.de

together with readouts of cellular states by fluorescent labels [3]. Light-sheet microscopy in particular [4] has enabled us to collect four-dimensional (4D) movies of a range of developing embryos from the fruit fly *Drosophila melanogaster* [5] to mammalian model organisms such as the mouse [6]. Complementary to recording cellular dynamics by microscopy, new genetic methods deliver single-cell atlases of gene expression in developing embryos. Those measurements yield detailed information on the genetic state of single cells across development [7] although the resolution in space and time is very coarse. Thus, a critical question in computational biology is how to integrate data from these different experimental modalities (e.g. connect time-lapse imaging with single-cell sequencing) and across experiments (e.g. imaging of the same specimen in several labs) [1]. How can we extract knowledge from such data collections? To this end, we also need to develop and refine our concepts and theories to make sense of the intricate patterns we can observe [8,9]. As state-of-the-art microscopy becomes widely available to the biology community [10], we need to establish structured and general schemes [11] concurrently to annotate and share the tracking results. We should base these annotations on a solid theoretical foundation: As pointed out in [12] we should regard the underlying terminology and formal concepts themselves as theories about the biological world. Here, we develop the conceptual architecture that supports integration and interoperability in the field of cell tracking experiments. We discuss the concept of *Cellular Genealogy* [13] (or Cell Lineage) as a fundamental notion for the development of the *Cell Tracking Ontology* [14] - an ontology designed for the integration of data obtained from cell tracking experiments.

## 2. The Ontology of Cellular Genealogies

Firstly, we define the notion of cellular genealogy and introduce essential subtypes.

### *2.1. Cell-Collective Genealogies*

We consider an individual cell as a material object; hence it has a lifetime, and since cells may divide and eventually die, the number of cells within a region under consideration (e.g. a developing organism) changes through time. Let us consider a time-segment (time-interval) I, such that during I no cell-division and no cell death occurs. Then, the cells existing during I form a collective Coll(I) that can be considered as a continuant through I [15][2]. During times when the number of cells changes, new cells may occur, and cells may disappear (i.e. die). Let us consider the life of an organism Org from fertilisation to death. Org starts as a single cell, the zygote, develops into a multicellular structure which exists for some time in a dynamic equilibrium (e.g. cells get replenished). After that time, Org dies, i.e. the structures dissolve. We divide the lifetime of Org, LifeT(Org), into a sequence of non-overlapping time-intervals I(1), ... I(n) such that the following conditions are satisfied:

(1) The intervals I(m) have a first point (they are left-closed), but no last point (right open). More precisely, they have the form [a(m), a(m + 1)) specifying the set $\{c : a(m) \leq c < a(m+1)\}$, where $0 \leq m \leq n$. Further, $\mathrm{LifeT(Org)} = \bigcup\{ I(m) : 0 \leq m \leq n\}$.

[2] We stipulate that a cell collective Coll preserves the number and the identity of the cells contained in Coll.

(2) Let Coll(I(k)) be the set of cells existing during I(k), then no cell death or division occurs during the interval I(k), k ≺ n. Further, we assume that Coll(I(k)) ≠ Coll(I(k + 1)).

These conditions imply further properties: From Coll(I(k)) to Coll(I(k + 1)) the number of existing cells changes. We consider two types: cell division and cell death. If a division of a cell $c \in$ I(k) occurs, then this process ends up with two daughter cells starting their existence at the left boundary of the interval I(k + 1). Analogously, if a cell undergoes cell death during I(k), then this ends at the left-boundary of I(k + 1). The final definition of CollGen(c(0)) then must specify which cells from Coll($k$) are related to which cells in Coll(k + 1). To this end, we introduce the relation div($x$, $y$, $z$): a cell $x$ of Coll($k$) undergoes a cell division during $I(k)$ resulting in two daughter cells y and z starting their existence at the left-boundary of $I(k$+1). We also introduce the relation id($x$, $y$) stating that $x$ belongs to Coll($k$) and $y$ belongs to Coll($k$ + 1) and both cells are identical. We further say that a cell $x$ in Coll($k$) has a successor cell $y$ in Coll(k + 1), if $y$ is either a daughter cell of x or if $y$ is identical with $x$, denoted by cell_succ(x, y). The cell collective genealogy CollGen(c(0)), is then specified by the following system CollGen(c(0)) = ({Coll($k$) | 0 ≤ k ≤ n}, div($x$,$y$,$z$), id($x$,$y$)), see Fig. 1a for an example.

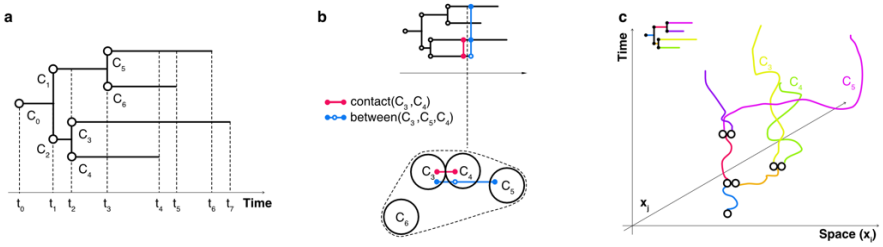


**Figure 1.** Different conceptual granularities of cellular genealogies: (a) Cell-Collective-Genealogies indicating cell division and cell death as well as the invariance intervals in time. (b) Cell-Situations-Genealogies capture spatial relations between cells in a given context (dashed outline shows convex hull). (c) Cell-Process-Genealogies describe cells as full spatio-temporal processes and their interactions.

We assume that for any organism there exists a uniquely determined cell-collective genealogy. Subsequently, we sketch a proof of this claim. Let t(0) be the starting time-point of the organism (say the zygote). A set of cells, Cells(t), is associated with any time point of the organism's life, i.e. the set of cells existing at that time-point. Now let t(1) be the latest time-point (> t(0)) such that during the time-interval [t(0), t(1)) no cells are born or die, and at time point t(1) either a cell division or a cell death occurs. t(1) is then the starting point of the next time-interval. The birth or death of a cell marks the end-point of a process, and these processes happen during [t(0), t(1)), and ending in time-point t(1) (death or birth). Assume we have carried out this construction up to time-point t(k). Then we repeat this procedure from t(k) onward, and there is a greatest time-point t(k + 1), greater than t(k), such that during [t(k), t(k + 1)) no cells are born or die, and at t(k + 1) cell birth or death is happening. A finite number of such steps t(0), ..., t(n) yields

a sequence of time intervals, [t(0), t(1)), [t(1), t(2)), ... , [t(k), t(k + 1)), ..., [t(n - 1), t(n)][3] which satisfies the desired conditions[4].

We conceptually divide the whole life-cycle of an organism (the ontogeny) in various phases (e.g. embryogenesis, growth, ageing) as it develops from a single zygote into its adult form, ages and finally dies [16]. As outlined, for every cell collective *x* there is a uniquely determined time-interval I such that no changes occur during *I*. We call this time-interval the *invariance interval* of the cell-collective; it has a left-boundary and no right-boundary. The lifetime of any cell in the collective includes the invariance interval as a temporal part. There is a successor relation between the cell-collectives. The cell-collective *y* is a successor of the collective *x* if the right boundary of the invariance interval of *x* coincides with the left-boundary of the invariance interval of *y*. Further, we group every cell collective with invariance interval I into subsets of Coll(I). During development, certain groups of cells may correspond to the development of specific structures, e.g. organs such as the heart, brain, digestive tract etc.

Understanding the structure of those sub-genealogies is an important research topic in developmental biology. The full sub-genealogy of a particular cell c, being a member of a cell-collective, contains all cells that are in the transitive closure of the successors of this cell. Any cell c generates its cell collective genealogy, denoted by CollGen(c) which is uniquely determined. *Cell lineage* indicates the development history of a tissue, organ or organism from earlier stages such as the fertilised egg.

### *2.2. Cell-Situation-Genealogies (SitGen)*

A cell situation genealogy is an extension of a cell collection genealogy: We start with the system CollGen(c(0)) and extend any collective Coll(k) of cells into an object-situation Sit(k). Sit(k) contains precisely the cells of Coll(k) as objects and is embedded into an object-situation with the time-frame *I*(*k*) and a specified space-frame which contains at least the spatial convex closure of the objects in Coll(*k*), see example in Fig. 1b). We are free to add relations between the cells or predicates of the cells in Coll(k). The set of these relations and predicates is inherently open-ended and defines the type of the corresponding situation. A signature Σ determines the situation type. Σ contains the admitted relational and predicate symbols. In the simplest case, this signature Σ is fixed for all situations of the genealogy. It may be necessary to introduce new relations and predicates during development. We consider a cell collective the simplest cell situation as its signature contains only the equality symbol =.

### *2.3. Cell-Process-Genealogies*

There are properties of cells, such as velocity or morphodynamics (changes in cell shape) along a trajectory that cannot be attributed to cells as objects and can only be captured by introducing cell processes (cf. [17]). The simplest process genealogy is defined by the transformation of the CollGen into a branching process, denoted ProcCollGen(c(0)). This process is determined by using the integration axiom of GFO and by transforming any cell of Coll(*k*) into a corresponding process. We can then define Proc(*k*) as the integration of all processes Proc(*c*) for any *c* in Coll(*k*). The processes in Proc(*k*) are usually not

[3] With the exception of the last interval all other intervals are left-closed and right-open.

[4] We assume that the time-points are represented by real numbers. A complete proof of the conditions uses the continuity of real numbers, in particular the fact that a bounded set of real numbers has a least upper bound.

isolated threads because there can be meaningful interactions between them (e.g. cells exchanging signals by direct contact or diffusible signalling molecules). Hence, there are various versions of potential process genealogies. Analogously, we may transform any cell-situation Sit($k$) into a process situation, called situoid. The investigation and classification of possible process genealogies is a research field of its own.

We give a simple example to illustrate these ideas in Fig. 1 visualising a part of a cell-collective genealogy. The invariance intervals are: $[t_0, t_1)$, $[t_1, t_2)$, $[t_2, t_3)$, $[t_3, t_4)$, $[t_4, t_5)$, $[t_5, t_6)$, $[t_6, t_7)$
The cell collectives, associated with the invariance intervals, are then:
Coll(0) = ($\{c_0\}$, $[t_0, t_1)$),
Coll(1) = ($\{c_1, c_2\}$, $[t_1, t_2)$),
Coll(2) = ($\{c_1, c_3, c_4\}$, $[t_2, t_3)$),
Coll(3) = ($\{c_3, c_4, c_6, c_5\}$, $[t_3, t_4)$),
Coll(4) = ($\{c_3, c_6, c_5\}$, $[t_4, t_5)$),
Coll(5) = ($\{c_3, c_5\}$, $[t_5, t_6)$),
Coll(6) = ($\{c_3\}$, $[t_6, t_7)$).
We can extend any of these cell collectives with relations. Here we consider an extension of the cell collective Coll(3). The collective identifies only the different cells contained in it, though it does not specify anything about the relations between those cells. As an example, we introduce the following relations: contact($x$,$y$) := the cells $x$ and $y$ are in contact and between($x$, $y$, $z$) := $z$ is between $x$ and $y$. The time-frame of Sit(3) coincides with the invariance interval of Coll(3), whereas the space-frame of the situation must be specified additionally (e.g. indicated by dashed region in Fig.1b). Since the cells typically move during the invariance interval, they may change their positions relative to each other (see Fig.1c). The snapshots of such situations are presentic entities, and a situation can be accessed only through its snapshots, which are called *presentic situations*.

A presentic situation cannot adequately describe processes. Let us consider a cell c, as an object having a lifetime and thus, persisting through time: It is the same cell at every time-point of its lifetime. A cell c may move through space, and this movement (trajectory) is a process (Fig. 1c). The presentic locations of c in time is an external attributive of c. The trajectory of a cell cannot be an attributive of c, because the path itself has no presentic nature. If we consider a snapshot of the trace, its form disappears. To model these aspects, we replace any cell in the situation by its corresponding process to get a processual situation.

## 3. Formal Axiomatisation of Cellular Genealogies

In this section, we present a selection of fundamental axioms about cell-collective genealogies. We develop these axioms using the onto-axiomatic method [18] integrating Hilbert's approach in [19] with a top-level ontology as a general analytical framework. Here we use some axioms from GFO [20] and adapt them to the domain of cell development, and more generally, to the field of developmental biology. We introduce axioms at two levels, the level of GFO-axioms, which are essential to understand the domain-specific notions, and the level of cell biology. Furthermore, we introduce the idea of trans-level axioms connecting concepts across different levels of abstraction.

### *3.1. GFO-Level.*

#### *3.1.1. Objects and Processes*

A material object is a spatio-temporal individual occupying space, persisting through time that is wholly present at every time-point of its lifetime. For every material object *Obj* and a time point *t* of the object's lifetime, there is a snapshot P of Ob*j* at that time-point *t*, which we express by the expression snap (Obj, P, *t*). These snapshots differ from time-point to time-point, though there is something in common, a similarity, between them, which is captured by a universal Univ(*Obj*). Furthermore, any instance of Univ(Obj) stands for Univ(*Obj*). If we consider, for example, an individual cat *C*, and if we take into account Univ(*C*), then any instance of Univ(*C*) stands for the whole universal Univ(*C*). Also, we may imagine a prototypical cat representing this universal. The universal and a corresponding prototype express the phenomenon of persistence, whereas the instances themselves may change their properties through time.

In contrast to an object, an individual process P evolves through time and can never be wholly present at a time-point. The restrictions of P to time-points of P's temporal extension are called process-boundaries. A process boundary of a process *P* is an entity of presentic nature. However, it can never represent the entire process *P*. Processes and objects thus exhibit a fundamental duality in spatio-temporal reality. Material objects and processes are connected in a particular way, which is expressed by the integration axiom of GFO: For every material object Obj there exists a process Proc(Obj) such that the snapshots of Obj coincide with process boundaries of Proc(Obj). Proc(Obj) is a minimal process associated with Obj, because it exhibits at the process boundaries only such properties that are genuine properties of the object. Genuine properties have a presentic nature and are independent of any process. A process boundary of Proc(*Obj*) contains a snapshot of the object *Obj*. We can extend the minimal process Proc(*Obj*) by adding further properties (that depend on the process) to the process boundaries, e.g. the velocity of a moving object.

#### *3.1.2. Situations*

An object-situation (simply called *situation* in this work) is composed of objects that are connected by relators. A situation is framed by a temporal interval and a space region. The material objects contained in a situation Sit define the skeleton of Sit. There is a certain freedom to specify the time frame and the space frame of a situation. For every situation, we may consider snapshots called presentic situations (PSit). An object-situation exhibits a presentic situation at any time-point of its time frame. We generally assume that a presentic entity is a dependent entity. It is either a snapshot of an object (or an object situation), or a part of the time-boundary of a process.

#### *3.1.3 A selection of axioms*

We select some axioms as examples, [20] presents a complete system. Axioms are formulated based on signatures providing symbols, predicates and relations.

We now fix a signature $\Sigma(0)$[5] = {Chr(*x*), Obj(*x*), Proc(*x*), Pres(*x*), SReg(*x*), TimeExt(*x*), exhib(*x*, *y*, *z*), lifetime(*x*, *y*), occ(*x*, *y*), procbd(*x*, *y*, *z*), tempext(*x*, *y*), temprestr(*x*, *y*, *z*), tp(*x*, *y*), where: Chr(*x*):= *x* is chronoid; Obj(*x*) := *x* is an object (in the sense of a material

[5] $\Sigma(0)$ is a minimal signature for the material ontological region according to GFO and must be extended in various directions. We consider biology as belonging to the material stratum of reality.

object, being a continuant); Proc($x$) := $x$ is process; Pres($x$) := is a presential; SReg($x$) := $x$ is a space region; TimeExt($x$) := $x$ is temporally extended entity; exhib($x$, $y$, $z$) := the material object $x$ exhibits the entity $y$ at time-point $t$; lifetime($x$, $y$) := $x$ is the lifetime of the object $y$; occ($x$, $y$) := $x$ occupies space region $y$; procbd($x$, $y$) := $x$ is a process boundary of the process $y$; procbd($x$, $t$, $y$) := $x$ is a process and $y$ is the process boundary of $x$ at time-point $t$; tempext($x$, $y$) := $x$ is the temporal extension of the process y; temprestr($x$, $y$, $z$) := $x$ is the temporal restriction of the process $y$ to the time-interval $z$, being a temporal part of the temporal extension of $y$; tp($t$, $x$) := $t$ is time-point of the interval $x$.
We select some axioms.[6]

$$\forall x\ (\text{TimeExt}(x) \leftrightarrow \text{Obj}(x) \vee \text{Proc}(x)) \tag{1}$$

$$\forall x\ (\text{Obj}(x) \rightarrow \exists y\ (\text{occ}(x, y) \wedge \text{SReg}(y)) \tag{2}$$

$$\forall x\ (\text{Obj}(x) \rightarrow \exists y\ (\text{lifetime}(y, x)) \tag{3}$$

$$\forall x\ (\text{Lifetime}(x) \rightarrow \text{Chron}(x)) \tag{4}$$

$$\forall x\ y\ t\ (\text{Obj}(x) \wedge \text{lifetime}(y, x) \wedge \text{tp}(t, y) \rightarrow \exists z\ (\text{Pres}(z) \wedge \text{exhib}(x, z, t)) \tag{5}$$

$$\forall x\ (\text{Proc}(x) \rightarrow \exists y\ (\text{tempext}(y, x)) \tag{6}$$

$$\forall x\ y\ (\text{procbd}(x, y) \leftrightarrow \text{Proc}(y) \wedge \exists t\ s\ (\text{tp}(t, s) \wedge \text{tempext}(s, y) \wedge \text{temprestr}(x, y, t)) \tag{7}$$

We define the process boundary of a process at time point $t$ (being an element of the temporal extension of the process) by the restriction of this process to this time-point.)

We introduce the following integration law[7]. For every material object Obj there exists a process Proc(Obj) such that the snapshots of Obj coincide with the process boundaries of Proc(Obj)). This process exhibits at its boundaries only genuine properties (attributives, i.e. they have a presentic nature and are independent of any process:

$$\forall x\ (\text{Obj}(x) \rightarrow \exists y\ (\text{Proc}(y) \wedge \forall z\ t\ u\ (\text{exhib}(x, t, u) \leftrightarrow \text{procbd}(y, t, u))) \tag{8}$$

There is a difference between snapshots of objects and process-boundaries: snapshots are taken from objects, never from processes. Presentials have two sources: they can be snapshots of objects (in this case we say that an object *Obj* exhibits a presential at a time-point of its lifetime), or can be contained in boundaries of processes. There are cases when a process boundary is the same as the snapshot of an object participating in this process. In general, the process boundary contains more properties than the process associated with the object. If an object Obj participates in a process *P* then Proc(Obj) is a minimal process layer of P [21]. We say that an object Obj participates in a process *P* if any snapshot of Obj is contained within a process boundary of *P*. We introduce the following relation.
partic($x$, $y$) := $x$ is an object, $y$ is a process, and $x$ participates in $y$.

[6] Establishing a fully developed system of axioms is a research topic of its own. Here we present only a selection of particularly important axioms.

[7] The integration law is a unique condition distinguishing GFO from other current top-level ontologies.

$$\forall x\, y\, (\text{partic}(x, y) \rightarrow \text{Obj}(x) \wedge \text{Proc}(y) \wedge \forall z\, (\text{snapshot}(z, x) \rightarrow \exists u\, (\text{procbd}(u, y) \wedge \text{part\_of}(z, u)) \quad (9)$$

$$\text{snapshot}(x, y) \rightarrow \text{Obj}(y) \wedge \exists t\, (\text{exhib}(y, t, x) \wedge \text{tp}(t, \text{Lifetime}(y)) \quad (10)$$

*3.2. Cell biology level*

Cells are considered living entities in contrast to inanimate entities such as stones. However, there is no clear consensus on how to define the boundary between the animate and inanimate. Typical defining properties of life are, among others, metabolism, adaptivity and interaction with the environment, self-organisation, reproduction, heredity, and growth. These conditions define a system which must satisfy at least the following basic properties. It should have a boundary, demarcating the system from the environment, and it should have inner parts. It should further be able to sense and interact with the environment (cf. *Autopoiesis* as an attempt to define living matter using concepts from general systems theory such as self-organisation). In biology, the cell is the simplest system satisfying these assumptions. It is an open problem whether these conditions - though necessary for the definition of life – are also sufficient for determining the essence of the animate. A minority of the biologists believed that an additional life force is needed to achieve a complete picture of the world [22]. The self-organised development of a cellular genealogy, starting from a zygote, seems to be an essential feature of the animate. Hence, the ontology of biology should consider the existence of cellular genealogies as one of the basic features demarcating biology from other fields of natural science, as physics or chemistry. Thus, we include relevant concepts of cellular genealogies, such as Cell($x$), Coll($x$) cell collective, cell division, cell death, cell situation Sit($x$) and the corresponding processes in the basic notions of life. The formalization of these notions use the signature $\Sigma(1) = \{\text{Cell}(x), \text{Coll}(x), \text{CollGen}(x), \text{Sit}(x), \text{PSit}(x), \text{Dead}(x), \text{id}(x, y), \text{div}(x, y, z), \text{invar}(x, y), \text{member\_of}(x, y), \text{daughter\_of}(x, y), \text{invar}(x, y)\}$, where: Cell($x$) := $x$ is a cell; Coll($x$) := $x$ is a cell collective; and CollGen($x$) := $x$ is a cell-collective genealogy.

A cell-collective has members, member_of($x$, $y$) := the cell $x$ is member of the collective $y$. Its invariance interval determines the lifetime of a collective: inv($x$, $y$) := $x$ is the invariance interval of the collective $y$. We distinguish two kinds of Time-Entities: Time Points and Time Intervals, where a time point is an element of a time interval. We use two types of time-intervals, those which are closed (they have a first point and a last point), and such which are left-closed and right open (i.e. they have a first point, but no last point). Notable examples are cell-division, cell-death and the various structural and morphological properties of cells. Subsequently, we present a selection of axioms.[8] These axioms can be easily transformed in pure first-order formulas, as exemplified by axiom 14.[9]

$$\forall x\, (\text{Cell}(x) \rightarrow \text{Obj}(x)) \quad (11)$$

[8] A more complete axiomatization of cellular genealogies is work in progress.

[9] *let x = [a,b), point(u,x) := u is a point of the interval x*
*x has a first time-point := $\exists v$ (point(v,x) $\wedge \forall w$ (point(w,x) $\rightarrow v \leq w$))*
*x has no last time-point := not ($\exists v$ (point(v,x) $\wedge \forall u$ (point(u,x) $\rightarrow u \leq v$)).*

$$\forall x\, y\, z\, (\mathrm{div}(x, y, z) \rightarrow \mathrm{Cell}(x) \wedge \mathrm{Cell}(y) \wedge \mathrm{Cell}(z) \wedge y \neq z \wedge x \neq y \wedge x \neq z) \quad (12)$$

$$\mathrm{Invar}(x) \leftrightarrow \exists y\, (\mathrm{Coll}(y) \wedge \mathrm{invar}(x, y)) \quad (13)$$

$$\forall x\, (\mathrm{Invar}(x) \rightarrow x \text{ has first time-point} \wedge x \text{ has no last time-point}) \quad (14)$$

$$\forall x\, (\mathrm{Coll}(x) \rightarrow \exists y\, (\mathrm{invar}(y, x)) \quad (15)$$

$$\forall x\, y\, z\, (\mathrm{Coll}(x) \wedge \mathrm{member\_of}(y, x) \wedge \mathrm{lifetime}(y, z) \wedge \mathrm{invar}(u, x) \rightarrow \mathrm{part\_of}(u, z) \quad (16)$$

A cell situation, Sit, contains a cell collective forming the skeleton of the situation, and various relations between cells, called the situation's signature. As an example of a signature consider $\Sigma = (\mathrm{contact}(x, y), \mathrm{between}(x, y, z), \mathrm{equidistance}(x, y, u, v))$.

$$\forall x\, (\mathrm{Sit}(x) \rightarrow \exists y\, (\mathrm{Coll}(y) \wedge \forall z\, (\mathrm{member\_of}(z, y) \leftrightarrow \mathrm{obj\_in}(z, x))) \quad (17)$$

For every situation S there exist a cell-collective C such that the members of C are exactly the objects in S. Here we assume that the situations are spanned by the cells of a collective.

$$\forall x\, (\mathrm{Sit}(x) \rightarrow \exists y\, (\mathrm{Coll}(y) \wedge \forall z\, (\mathrm{member\_of}(z, x) \leftrightarrow \mathrm{obj\_in}(z, y)) \quad (18)$$

Since the cells of a situation can move during the situation's time-frame the relations between them may depend on time, e.g. two cells c, d are in contact at time point *t*, and separated at another time-point *t'*. Hence, the relations (e.g. contact(*x*,*y*)) must be extended by a time-argument such as contact(*x*, *y*, *t*). The time-frame of situation *S* is the invariance interval of the collective contained in *S*. coll_succ (*x, y*) denotes a successor relation such that *x* and *y* are cell collectives, and *y* is the successor of *x*.

There exists exactly one cell-collective without a predecessor and exactly one cell collective without a successor. A cell-collective genealogy CollGen is a temporally extended structure consisting of a sequence of invariance intervals and the cell-collectives associated with these intervals: CollGen = ($\mathrm{Coll}_1$, ..., $\mathrm{Coll}_m$, $\mathrm{inv}_1$, ..., $\mathrm{inv}_m$, coll_succ(*x*, *y*), id(*x*, *y*), div(*x*, *y*, *z*), Dead(*x*)). A sequence of intervals specifies such a cell genealogy Int(CGen) =($\mathrm{inv}_1$, ..., $\mathrm{inv}_m$), by the collectives $\mathrm{Coll}_m$, with added coll-successor relation, and (at least) two relations id(*x*, *y*), and div(*x*, *y*, *z*) between the cells of a collective, and the cells of the successor collective. By adding relations to the cell-collectives of a genealogy, we define the notion of a cell-situation genealogy, denoted by SitGen. A situation genealogy is said to be stable if the signature is the same for any situation of the genealogy. A many-sorted model-structure of a cell-collective genealogy can be specified as follows: CollGen = ((L,$\mathrm{Inv}_1$,...,$\mathrm{Inv}_n$, <), Cell, $\mathrm{Coll}_1$, ..., $\mathrm{Coll}_n$, lifetime(*x*, *y*), cell_succ(*x*, *y*), id(*x*, *y*), div(*x*, *y*, *z*)).

Here, Cell(*x*) is a predicate, the extension[10] of which contains all cells occurring during the full temporal extension of the genealogy. The extension of Coll(*i*) are subsets of the extension of the predicate Cell(*x*). (*L*, <) is a dense linear ordering, presenting the set of time-points, and $\mathrm{Inv}_i$ are left-closed and right-open intervals of (*L*, <).

[10] The extension of a predicate P(*x*) is the set of all entities satisfying this predicate. This notion can be explicated based on a model-structure established according to the methods of logic and model theory [23].

By adding further relations, presented formally by a signature $\Sigma = (r(1), ..., r(n))$, we get a model-structure for a situation genealogy SitGen: SitGen = (CollGen, int(Σ)). int(Σ) is the interpretation of the relational symbols of Σ in the corresponding cell-collectives $Coll_i$[11].

*3.2.1. Description of the relations*

We introduce the following relations: lt($x$) := lifetime of the cell $x$ and is defined by the following condition $lt(x) = y \leftrightarrow$ lifetime($x, y$); daughter($x, y$) := $\exists z$ (div($x, y, z$); $Inv_i(x)$ := $x$ is an element of the i-th invariance interval; $Init_i(x)$ := $x$ is the initial time-point of the interval $Invar_i$; Init($x$) := $\vee\{ Init_i(x) \mid i \leq n\}$; init($x, y$) := $x$ is a cell and $y$ is the initial time point of the cell's lifetime; $Coll_i(x)$ := $x$ is an element of the $i$-th cell collective. The definition of succ($x, y$) uses the following formulas: $\varphi_i(x, y)$ := $Coll_i(x) \wedge Coll_{i+1}(y) \wedge$ (id($x, y$) $\vee$ (daughter($x, y$)); then, cell_succ($x, y$) := $\vee\{\varphi_i(x, y) \mid 0 \leq i \leq n -1\}$.

*3.2.2. Selection of axioms:*

($L$, <) is a dense linear ordering. $Inv_i$, $i$ = 1,...,$n$, are intervals, such that the following conditions are satisfied:

$$\forall x\ (L(x) \leftrightarrow Inv_1(x) \vee \ldots \vee Inv_n(x)) \tag{19}$$

$$\Phi_i = \exists x\, y\ (\forall u\ (\ Inv_i(u) \leftrightarrow x \leq u \prec y), i = 1,..., n \tag{20}$$

$$\bigwedge \{ \sim \exists x\ (Inv_i(x) \wedge\ Inv_j)(x) \mid i \neq j \} \tag{21}$$

$$\forall x\, y\ (Inv_i(x)\ \wedge Inv_{i+1}(y) \rightarrow x \prec y), i = 1, 2, ..., n - 1 \tag{22}$$

$$\forall x\ (Coll_i(x) \rightarrow Cell(x)) \tag{23}$$

$$\forall x\ (Coll_i(x) \rightarrow Inv_i \subseteq lt(x)) \tag{24}$$

$$\forall x\, y\ (\text{cell\_succ}(x, y) \rightarrow \text{daughter}(x,y) \vee \text{id}(x, y)) \tag{25}$$

$$Coll_i(x) \wedge \text{div}(x, y, z) \rightarrow \exists u\ (\ \text{Init}(u) \wedge \text{init}(y, u) \wedge \text{init}(z, u)) \tag{26}$$

A cell situation genealogy SitGen is based on a cell-collective genealogy CollGen extended by adding relations to any of the cell-collectives.

## 4. The experimental framework and its Formalisation

*4.1. Basic conditions - Ontology of Frame-Sequences*

In this section, we investigate and analyse cell tracking experiments based on the principle of time-lapse microscopy. In reality (*in vivo* as well as *in vitro*) cells are moving

[11] If $r(x, y, z)$ is a ternary relation symbol in Σ, then an interpretation of $r$, denoted by int($r$), in $Coll_i$ = {a(1), ...., a($n$)} is a subset of $Coll_i \times Coll_i \times Coll_i$ (e.g. the relation between($a$, $b$, $c$)).

and changing continuously in time and space. Hence, the time-points are densely ordered: after a time-point, there is no direct successor. In the considered experiments, discrete snapshots of the continuous dynamics are taken. These snapshots provide incomplete information about an individual situation genealogy of the independent reality. Let a given situation genealogy SitGen be specified by the structure SitGen = ((L, $Inv_1$, ..., $Inv_n$, <), $Sit_1$,..., $Sit_n$, lt($x$,$y$), cell_succ($x$, $y$), id($x$, $y$), div($x$, $y$, $z$), int(Σ)). The time-points at which the snapshots are taken are from a finite subset S ⊆ L of the linear ordering (L, <), hence (S,<) is a finite linear ordering which can be ordered by natural numbers. A snapshot at time point t yields a presentic situation PSit(t), which is called the frame at t, denoted by Fr($t$). Any experiment Exp of this type results in a finite sequence Seq(SitGen)) = (Fr(t(1)),..., Fr(t(n)) of frames, called components of the sequence. This sequence, related to an experiment Exp, is denoted by Seq(Exp). We say that a time-lapse experiment Exp is adequate for the situation genealogy SitGen if for any situation Sit in SitGen there exists a snapshot of Sit in Seq(Exp).[12] These sequences are the entities to be investigated. Any of the pictures Fr(k) reflects a snapshot of a situation from SitGen. For the sake of simplicity, we identify the frame Fr(i), being a picture, with the snapshot of the reflected situation. In the following, we fix a sequence Seq(Exp) as a result of a certain experiment.

Every frame is a snapshot of a situation. Hence a frame is a presentic situation (PSit). Further, any presentic situation in FrSeq contains presentic cells, also called presentials. Presentials possess various properties and can relate to other entities. Some properties are inherent to the objects, e.g. the form (based on metrics) or the number of proteins of a certain type. Others are external to the cells, such as the distance between two cells, and the position of the cell in space. Further important relations between two presentic cells are: contact($x$, $y$) := the cells I and y are in contact; or relative spatial positions between the cells $x$ and $y$, for example, the cell $y$ is right to the cell $x$, $y$ is left to $x$, above, below or spatial relations with respect to an (often anatomical) frame of reference (e.g. dorsal, ventral, distal etc…). Also, spatial relations with more than two arguments are possible, e.g. between($x$, $y$, $z$) the cell $y$ is localised between the cell $x$ and the cell $z$. A further example of a relation with four arguments is equidist($x$, $y$, $z$, $u$) meaning that the distance between $x$ and $y$ is the same as that between $z$ and $u$.

Although this set of spatial relations may seem quite limited, from a biology point of view, it is in itself already useful to describe a large class of symmetries that are established or broken [24] during development (e.g. mirror symmetry in bilateral animals). Further, from a theoretical perspective, the relations of betweenness and equidistance are even sufficient to establish the whole elementary planar Euclidean geometry [25]. We emphasise that in a single frame, only presentic properties can be identified, that are independent of any process. Aspects such as the circularity of a cell path or the morphodynamics of a cell (its particular pattern of shape changes) cannot be detected in any one individual frame and are thus not presentic, but processual properties. The derivation of a processual property of a cell or a cell-collective can only be achieved by an analysis of a sequence of frames. In essence, a sequence of frames can be transformed into a video providing processual properties. Since the famous works of the photographer Eadweard Muybridge to study motion from a sequence of static pictures, the method of changing time scales via slow-motion or time-lapse/fast motion provided

[12] This condition implies that the temporal distances between the snapshots are sufficiently small to acquire the relevant information about a cell-division.

many insights in the processual properties of nature, and in particular into the properties of embryogenesis [8,9].

### *4.2 Formal Axiomatisation of frame sequences*

#### *4.2.1 Some predicates and informal description of axioms.*

FSeq($x$) denotes a frame-sequence $x$, and its components are called frames. Every frame is a snapshot of a situation, denoted by PSit. We introduce a linear ordering between the components of a frame-sequence, hence such a sequence can be presented by the structure FSeq = ({F(1), …, F($n$)}, <), where F(1) < … < F($n$). Let Seq be a frame sequence; we say that a component G of Seq is a successor of the component F of Seq, if F < G and there is no component between F and G. In this case we introduce the relation seq_succ(F,G), G is a sequence-successor of F. We assume that in any frame there occur cells, that these cells are presentials, and any such presentic cell is a snapshot of a uniquely determined cell (with lifetime> 0).

#### *4.2.2 Selection of formal axioms*

We first introduce a signature Σ(2) on which the axioms are based: Fr(x) := x is a frame, FSeq(x) := x is a frame sequence, PCell(x):= x is a presentic cell, PSit(x) := x is a presentic situation, comp(x,y) := x is a component of the frame-sequence y, < is the linear ordering between the components of frame sequence, ipart(x,y) := x is an image-part of the frame y.

$$\forall x\, y\, (\mathrm{FSeq}(x) \wedge \mathrm{comp}(y, x) \rightarrow \exists z\, (\mathrm{Sit}(z) \wedge \mathrm{snapshot}\, (y, z)) \tag{27}$$

$$\forall x\, y\, (\mathrm{comp}(y, x) \rightarrow \mathrm{Fr(x)} \wedge \mathrm{FSeq(x)}) \tag{28}$$

$$\forall x\, y\, (\mathrm{FSeq}(x) \wedge \mathrm{comp}(y, x) \rightarrow \exists z\, (\mathrm{PCell}(z) \wedge \mathrm{ipart}(z, y))^{13} \tag{29}$$

$$\forall x\, (\mathrm{FSeq}(x) \rightarrow \exists y\, z\, (\mathrm{comp}(y, x) \wedge \mathrm{comp}(z, x) \wedge \sim \exists u\, (\mathrm{comp(u,}\, x) \wedge u < y)\ ) \wedge \sim \exists v\, (\mathrm{comp}(v, x) \wedge y < v)) \tag{30}$$

$$\forall x\, (\mathrm{FSeq}(x) \rightarrow \exists y\, (\mathrm{SitGen}(y) \wedge \forall u\, (\mathrm{comp}(u, x) \rightarrow \exists v\, (\mathrm{sit\_of}\, (v, y) \wedge \mathrm{snapshot}(u, v)). \tag{31}$$

Axiom (31) establishes a link between the experiment and the independent reality of situational genealogies. Such an axiom should be postulated for any type of experiment as each experiment is directed at objects to be studied.

We have established a relation between cellular genealogies and sequences of frames from time-lapse experiments. The final reconstruction of genealogies is then an information artefact that captures relevant knowledge about the real-word genealogies.

[13] There is an ambiguity between *part of* a frame and *image-part* of a frame. For sake of simplification we do not distinguish between the image of an entity and the entity itself. We could simply say that a presentic cell is a part of a frame. Though, a frame can possess image parts as artefact to which no real entity corresponds.

## 5. Conclusion and Future Research

In this work, we outlined an ontological foundation of cellular genealogies concerning a fundamental theory and a formal representation of a type of experiments and its results. The full framework will provide three levels of abstraction. This paper addresses the first two levels: the theory and the experiment level. At the *theory level*, we analysed cellular genealogies as independent real-world entities using the onto-axiomatic method. We proposed a partial formal axiomatisation of knowledge assumed to be true for every cellular genealogy. At the *experiment level*, we formally described time-lapse experiments and developed an axiomatic foundation of this domain. Any experimental framework should be considered as a mediator between a theory and the real-world entities to be studied. Experiments provide data about a domain of interest; they play an indispensable role for supporting or disproving a theory, and thus for further development and revisions of theories. Our development of the overarching conceptual framework follows the onto-axiomatic method adhering to the principles of model-theory [23,26] as introduced in biology by [27].

As a conceptual next step, we will extend the genealogy types outlined here, to model the self-organising processes in biology as complex, interacting systems (embryonic development being a prime example). Building on existing work on collective phenomena by [15], we will consider groups of cells and their mutual interactions. We could model groups of interacting cells[14] as object-situations, e.g. a cell-group can be a specific tissue (or a group of precursor cells). We may further introduce material boundaries and dynamics of these cell-groups to build a cell-group genealogy. A critical problem will be to find appropriate levels of granularity. Here, we will build on ideas from complex systems research.

Finally, we outline directions for future research, as we feel that the presented framework paves the way for new questions and might even open new fields:

*Development of suitable representational levels.* We presented theories on a general level in this paper. However, the instance-level is still needed, if we want to study individual cellular genealogies. We are currently investigating various representational levels as continuation of the present paper.

*Extending the ontological foundation of cellular genealogies.* To elaborate on the presented framework, we will analyse existing knowledge in developmental biology and successively transform it into formalised axioms based on the onto-axiomatic method.

*Elaborating Genealogy-Theories for particular model species.* An ideal first step would be the development of a complete genealogy-theory for the model organism *Caenorhabditis elegans* as much is known about its genetics and development [2,28].

*Extending the outlined theory to other levels of granularity.* Our current genealogy-theory refers to the single-cell level as a 'middle-out' starting point as already proposed by [29][15]. We will consider two canonical extensions of granularity levels: We explicitly model the state of cells at the molecular level using [30,31] and we model cell-groups as tissue-level entities.

[14] Such as the parasegments forming during patterning of Drosophila embryos [16].

[15] Sidney Brenner is being credited with saying 'I believe very strongly that the fundamental unit, the correct level of abstraction, is the cell and not the genome' by [29].

*Modelling cellular genealogies in disease.* To support computational approaches in systems medicine, we should elaborate a specific theory for abnormal genealogical patterns as can be found in certain cancers, such as leukaemia [32] and related diseases.

## References


[1] J.B. Wallingford, The 200-year effort to see the embryo, *Science*. **365** (2019) 758–759.
[2] J.E. Sulston, E. Schierenberg, J.G. White, and J.N. Thomson, The embryonic cell lineage of the nematode Caenorhabditis elegans, *Dev. Biol.* **100** (1983) 64–119.
[3] S.G. Megason, and S.E. Fraser, Imaging in systems biology, *Cell*. **130** (2007) 784–795.
[4] J. Huisken, J. Swoger, F. Del Bene, J. Wittbrodt, and E.H.K. Stelzer, Optical sectioning deep inside live embryos by selective plane illumination microscopy, *Science*. **305** (2004) 1007–1009.
[5] L.A. Royer, W.C. Lemon, R.K. Chhetri, Y. Wan, M. Coleman, E.W. Myers, and P.J. Keller, Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms, *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3708.
[6] K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L.A. Royer, S.C. Turaga, K. Branson, and P.J. Keller, In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level, *Cell*. **0** (2018). doi:10.1016/j.cell.2018.09.031.
[7] R.M. Harland, A new view of embryo development and regeneration, *Science*. **360** (2018) 967–968.
[8] J. Wellmann, Model and movement: studying cell movement in early morphogenesis, 1900 to the present, *Hist. Philos. Life Sci.* **40** (2018) 59.
[9] J. Wellmann, Die Form des Werdens: eine Kulturgeschichte der Embryologie; 1760-1830, Wallstein, 2010.
[10] R.M. Power, and J. Huisken, Putting advanced microscopy in the hands of biologists, *Nat. Methods*. (2019). doi:10.1038/s41592-019-0618-1.
[11] A.N. Gonzalez-Beltran, P. Masuzzo, C. Ampe, G.-J. Bakker, S. Besson, R.H. Eibl, P. Friedl, M. Gunzer, M. Kittisopikul, S.E. Le Dévédec, S. Leo, J. Moore, Y. Paran, J. Prilusky, P. Rocca-Serra, P. Roudot, M. Schuster, G. Sergeant, S. Strömblad, J.R. Swedlow, M. van Erp, M. Van Troys, A. Zaritsky, S.-A. Sansone, and L. Martens, Community Standards for Open Cell Migration Data, *BioRxiv*. (2019) 803064. doi:10.1101/803064.
[12] S. Leonelli, The challenges of big data biology, *Elife*. **8** (2019). doi:10.7554/eLife.47381.
[13] I. Glauche, R. Lorenz, D. Hasenclever, and I. Roeder, A novel view on stem cell development: analysing the shape of cellular genealogies, *Cell Prolif.* **42** (2009) 248–263.
[14] P. Burek, N. Scherf, and H. Herre, A pattern-based approach to a cell tracking ontology, *Procedia Comput. Sci.* **159** (2019) 784–793.
[15] Z. Wood, and A. Galton, A taxonomy of collective phenomena, *Appl. Ontol.* **4** (2009) 267–292.
[16] L. Wolpert, and C. Tickle, Principles of Development, OUP Oxford, 2011.
[17] P. Burek, N. Scherf, and H. Herre, Ontology patterns for the representation of quality changes of cells in time, *J. Biomed. Semantics*. **10** (2019) 16.
[18] R. Baumann, F. Loebe, and H. Herre, Axiomatic theories of the ontology of time in GFO, *Appl. Ontol.* **9** (2014) 171–215.
[19] D. Hilbert, Axiomatisches Denken, in: D. Hilbert (Ed.), Dritter Band: Analysis · Grundlagen Der Mathematik · Physik Verschiedenes: Nebst Einer Lebensgeschichte, Springer Berlin Heidelberg, Berlin, Heidelberg, 1935: pp. 146–156.
[20] H. Herre, General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling, in: R. Poli, M. Healy, and A. Kameas (Eds.), Theory and Applications of Ontology: Computer Applications, Springer Netherlands, Dordrecht, 2010: pp. 297–345.
[21] H. Herre, B. Heller, P. Burek, R. Hoehndorf, F. Loebe, and H. Michalek, 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, 2006.
[22] H. Driesch, Philosophie des organischen; Gifford-vorlesungen gehalten an der Universität Aberdeen in den jahren 1907-1908, (1921). https://www.worldcat.org/title/philosophie-des-organischen-gifford-vorlesungen-gehalten-an-der-universitat-aberdeen-in-den-jahren-1907-1908/oclc/3408806.
[23] W. Hodges, School of Mathematical Sciences Wilfrid Hodges, and H. Wilfrid, Model Theory, Cambridge University Press, 1993.
[24] A.C. Neville, Animal asymmetry The Institute of Biology's Studies in Biology, *London, UK: Edward Arnold*. (1976).

[25] A. Tarski, What is elementary geometry?, in: Studies in Logic and the Foundations of Mathematics, Elsevier, 1959: pp. 16–29.
[26] C.C. Chang, and H.J. Keisler, Model Theory, Elsevier, 1990.
[27] J.H. Woodger, and W.F. Floyd, The Axiomatic Method in Biology, By J.H. Woodger. With Appendices by Alfred Tarski and W.F. Floyd, Cambridge University Press, 1937.
[28] S. Brenner, Nature's gift to science (Nobel lecture), *Chembiochem*. **4** (2003) 683–687.
[29] D. Noble, The music of life: biology beyond the genome, Oxford: Oxford University Press, 2006.
[30] J. Bard, S.Y. Rhee, and M. Ashburner, An ontology for cell types, *Genome Biol.* **6** (2005) R21.
[31] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.* **25** (2000) 25–29.
[32] C. Bahr, L. von Paleske, V.V. Uslu, S. Remeseiro, N. Takayama, S.W. Ng, A. Murison, K. Langenfeld, M. Petretich, R. Scognamiglio, P. Zeisberger, A.S. Benk, I. Amit, P.W. Zandstra, M. Lupien, J.E. Dick, A. Trumpp, and F. Spitz, A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies, *Nature*. **553** (2018) 515–520.