Towards a Cellular Genealogy Ontology

Patryk Burek[‡] & Heinrich Herre[‡*], Ingo Roeder[†], Ingmar Glauche[†], Nico Scherf[†], Markus Löffler[‡]*

[‡] Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Germany

[†] Institute for Medical Informatics and Biometry, Dresden University of Technology, Germany

ABSTRACT

Motivation: In the present paper we outline basic ideas and results about a *Cellular Genealogy Ontology* (CGO). This work is aimed at providing a framework for analyzing, specifying and annotating results of experiments and of simulations in the field of stem cell research. The real world objects of these investigations are processual cellular genealogies which are studied by using, among other methods, time lapse experiments. This framework pursues three goals: Firstly, it provides a domain independent core ontology, called Simple Process Object Ontology (SPOO). SPOO is the basis for a coherent and integrative handling of objects, processes and characteristics which are the main building blocks for any domain ontology. Secondly, this core ontology is utilized for the development of a domain ontology for cellular genealogies. Thirdly, this domain ontology CGO is intended to support and enrich tracking algorithms by providing annotations of photos, storage and structuring of analysis results, and, thus, enables the discovery of correlations between qualities, visualizations of annotations. Furthermore, CGO is intended to support semantically correct data exchange (import, export).

1 INTRODUCTION

The application of time lapse video microscopy for the analysis of cell cultures facilitates the tracing of single cells, comprising all the progeny over extended time periods up to several days. This includes the temporal analysis of cell specific parameters like morphology, expression of marker genes, cell cycle time, motility or the occurrence of cell death within the population context. All these different information can be comprised into a pedigree-like structure, referred to as *cellular genealogy* (Figure 1), in which the founder cell represents the root and the progeny is arranged in the branches. In such a framework a *cell* is perceived as a *spatially and temporally extended object*. The existence of such a cell is temporally restricted by the generating division of the paternal cell and by the terminating division that generates the descending daughters. Alternatively, a cell might undergo cell death which also precludes its existence. Automated analysis of time lapse videos from cell cultures allows the simultaneous tracking of a multitude of root cells. Automatic cell tracking procedures are based on the analysis of each individual picture taken during the time lapse experiment. Under a set of rules (characteristics e.g. size, shape, color) certain objects are identified as cell objects in every single picture taken at a particular time point *t*. This process is termed image segmentation.

2 PROCESSUAL CELLULAR GENEALOGIES

An accurate description and definition of the notion of a cell genealogy, as mentioned in section 1, leads to a number of ontological problems, which are subsequently discussed. It turns out that a cell admits different views which are described by the following observations.

(1) a cell, considered at a time-point , is completely present at this time-point and has no temporal parts,.

(2) a cell participates in a process which exhibits the change of this cell at different time-points.

(3) a cell persists through time, that means that this cell is the same at different time-points.

These conditions are, obviously, incompatible. This situation is, we believe, caused by the fact that the term “cell” denotes three different pairwise disjoint concepts which are closely related. A cell, satisfying the condition (1), is called a *presentic cell*, the corresponding predicate is denoted by PresCell(x); behind the condition (2) there is a *processual cell*, whose concept is denoted by ProcCell(x), and, finally, the condition (3) captures a type of a cell which we call *continual cell*, expressed by the predicate ContCell(x).

The relation between these three concepts is specified by several axioms, introduced in GFO (Herre 2010) , which use two ternary relations. If p is a process then timerest(p,t,q) has the following meaning: the restriction of the process p to the time-point t yields the presentic entity d . Furthermore, the relation exhibit(c,t,d) expresses the condition: the continual entity c exhibits at time-point t the presentic entity d.

We assume the following integration axioms, formulated in GFO (Herre 2010), which express fundamental interrelations between the categories Proc(x), Cont(x), and Pres(x).:

* To whom correspondence should be addressed.

$\forall x$ (Cont(x) →
$\exists y$ (Proc(y) ∧ $\forall z$ t (exhibit(x,t,z) → timerestr(y,t,z)).
$\forall x$ (Pres(x) → $\exists yt$ (Proc(y) ∧ timrestr(y,t,x))

Note, that these axioms allow to interpret a continual cell as a process. The most basic entity is the processual genealogy of the cell, denoted by ProcGen(c); this is a process which happens and unfolds into branches.. The introduced entities are related to ProcGen(c) and are displayed in Figur 1.

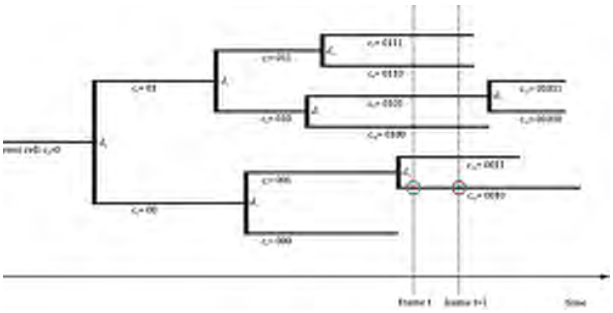

Fig. 1

Within the given five generation genealogy the thin horizontal lines represent the cells c_i whereas the divisions d_j are marked by the thick vertical bars. The horizontal dimension is time with the founding root cell c_0 indicated on the left side. Thus, the length of the horizontal lines represents the duration of the continual cell's existence and is a measure of the cell cycle time. According to the axiom IntAx the horizontal lines present a continual cell which is related to the underlying process, i.e. the processual cell. Final cells on the right side are called leaf cells. The vertical dashed lines indicate the two snapshots of ProcGen(c0) for the frames *t* and *t+1* as indicated in Figure 1. The genealogy illustrates that cell object a in frame *t* (blue circle) and b in frame *t+1* (red circle) are presentic cells which participate in the same processual cell c_{12} with the binary code 0010.
The representation , as displayed in Fig.1, is called *basic geneology of the cell* c. These basic genealogies can be annotated by additional properties.

3 SIMPLE PROCESS OBJECT ONTOLOGY (SPOO)

The SPOO exhibits a modification of a part of GFO, in particular, several new concepts are introduced, and some simplifications of GFO are proposed.

3.1 Main Ontological Choices of SPOO

In this section we outline the Simple Process Object Ontology (SPOO), which exhibits a representational level for categories and provides the most general vocabulary and structuring principles for the Cellular Genealogy Ontology (CGO). Top level of SPOO provides the most general concepts which can be used for domain modeling and, additionally, it permits to localize the concept of a process in the hierarchy of concepts.

The root category of SPOO is *Entity*, and every category of SPOO subsumes under the category *Entity* whose instances include all individuals.. The main distinction between the categories of SPOO is based on their relation to time. The ontology of time is taken in SPOO as primitive and in fact SPOO can be integrated with an arbitrary time ontology as long as it provides the notions of *time point* and *time period* or equivalent. This is a very minimal requirement on time ontology, namely that time can be organized into points and intervals.

We distinguish three categories of entities with respect to their relation to time, namely: (1) *Presentials* - those entities which are located at a time point, (2) *Temporally Extended Entities* (TEE) - entities extended through time, i.e. having some lifetime or happen in time, (3) *Abstracts* – entities, which do not have a relation to time, e.g. mathematical entities. In contradistinction to GFO, SPOO does not make a clear distinction between continuants (having a life-time), and processes (having a temporal extension)

Among abstract entities two are of particular importance, namely, *Property* and *Value*. Properties are abstract entities which we measure, observe or calculate, such as weight, color, speed or temperature. Values are, for example, volumes used in measurement, observation or calculation e.g. 10kg, green, 40$. Often they are scalars or vectors.
Both properties and values are considered as abstracts having no relation to time and independent from the entities which are characterized by them.

An individual assignment of a property and a value to an entity is called *Quality* and since this assignment characterizes an entity it is called its *Characteristic*. A host of a quality can be of an arbitrary ontological kind including a quality itself.

The second class of characteristic is *Role*, which is a notion related to the category of *Relation*, namely a role is an aspect of an entity in context of a relation. For example a cup has in context of the stands_on relation a role of a standing_object, whereas desk has the role of supporting_object.

Relations are defined as entities which glue or connect other entities via roles played by those entities in relations, e.g. stands_on relation glues a cup with a desk. Note, that in SPOO - in accordance with the GFO framework - we do not consider relations as mere sets of n-tuples of the corresponding arguments, but as concrete entities that glue their arguments together.

Relations can be contrasted to *Objects*, which are entities typically perceived and distinguished from other entities as existing to some extent independently of their surroundings. The independence of objects from their surroundings means that they can be conceived and modeled without references

to the surrounding entities. Objects typically have names without references to other entities e.g. apple, house and have characteristics by which they are conceived. This contrast objects to relations and characteristics which in turn often are named by references to their host e.g. process of *goods* transportation, color of *apple*.

In SPOO entities which are composed of at least one relation together with its players are called *Situations*. Thus e.g. a cup standing on a table is considered as a situation in which are involved two objects, namely a cup and a table, one relation: stands_on and two roles: standing_object and supporting_object. This situation should not be confused with stands_on relation which is dependent on its players but does not have them as its parts.

3.2 Matrix of SPOO entities

The matrix of SPOO entities is constructed by combining the discussed above ontological choices. In each group of entities, i.e. presentials, temporally extended entities and abstracts can be identified objects, qualities, roles, relations and situations. The summary of the SPOO entities is presented in table 1.

Table 1. SPOO Categories

Presential Entities	Time Extended Entities	Abstract Entities
Presentic Object	Object/Continuant	Abstract Object
Presentic Relation	Process	Abstract Relation
Presentic Situation	Situation	Abstract Situation
Presentic Quality	Quality	Abstract Quality
Presentic Role	Role	Abstract Role
		Property
		Value

The SPOO entity which deserves a particular attention is *Process*, which is considered as a time extended relation. Thus, in other words any binding of entities which exist through time is considered in SPOO as a process, e.g. the relation of stands_on if considered as existing through time would be a process which can have its own dynamics. This, very general treatment of processes significantly differs SPOO from other top level ontologies e.g. (DOLCE (Masolo, C. et al. 2003)) and enables representing a broad spectrum of processes covering both dynamic process such as movement of a body as well as static processes such as keeping goods in warehouse.

A number of roles are involved in processes, called *Process Participants*, are introduced i.e. *Process Executor*, *Process Resource*, *Process Operand*. A participant of a particular type is a process executor – this is a role of an entity, which is responsible for a process, hosts and realizes a process. It should be mentioned that SPOO is permits not only object but also other entities, including processes themselves, to play a role of executors. This is of particular significance in dynamic modeling in engineering and in natural science when processes often hosts and execute other processes.

An executor alone seldom realizes a process, but most likely there are also other entities contributing to the process realization. Such entities are called *Process Contributors*.

Not all participants of a process are responsible for its realization or for an active contribution to it. In SPOO such participants are called *Process Resources*. Process resources are those entities involved in a process, which are not realizing a process or contributing to its realization. A resource is involved in a process intentionally and thus it is involved in the goal of a process as well. Two types of resources can be identified: *Process Products* – resources produced by a process, and *Process Operands* – resources changed or consumed by a process.

4 REQUIREMENTS

In a biological context the cellular genealogies represent unique examples of the developmental sequence originating from the root cell as it occurs under particular assay conditions. The number of research groups doing time lapse experiments and analysis of genealogies increases (Schröder 2008, Eilken 2009, Roeder I 2006, Scherf N 2008, Glauche I 2007), however, the data exchange between them is limited, due to, among other reasons, the lack of a common data format. For the data storage, exchange as well as for the statistical analysis of cellular genealogies a precise characterization of the particular data types is required. The Ontology of Cellular Genealogies (CGO) is developed for that purpose. However, in our opinion the development of the domain ontology alone is not sufficient due to following reasons.

Firstly, the potential users of a genealogy ontology have their own interests and perspectives taken on the domain, hence, there is a high risk of refactoring and restructuring of the ontology during later stages of development or during application. However, refactoring of deployed artifacts is a difficult and expensive enterprise. This fact, well known in software engineering, is also true in the field of ontology development. The recent initiatives of refactoring the biological ontologies, e.g. (Diehl A, et al. 2009) demonstrate that biological ontologies, in order to fulfill their goals, should be well-structured and founded on a solid ground.

Secondly, the development of biological ontologies is a complicated enterprise due to many factors, including the high complexity of the domain and the dynamics of knowledge increase and evolution. This is true also for time lapse experiments. It can be expected that already at an early stage of the ontology's development users will annotate the observed genealogies with more and more information, as well as with new analysis results. Thus, a core requirement for the CGO ontology is to build a model which is easily extensible. A strategy might be the take the processual geneaology of a cell as starting point, and then use als a initial

genealogy the basic genealogy, as introduced in section 2. Any properties, introduced for the the entity ProcGen(c) must be compatible with the basic genealogy BasGen(c).

To illustrate the need of solid conceptual foundations, as well as extensibility, let us consider the most straightforward conceptual model of a cellular genealogy, namely an ontology with one concept Cell and one hasParent relation attached to Cell.

This model is sufficient to represent the cellular genealogy illustrated on figure 1. Moreover, it permits for some extensions e.g. assignment to the cell concept of properties such as e.g. life time or type. However, implementing such a model, although sufficient at the early development stage, would shortly bring us to trouble when it turns out that some of the properties change during the cell life time and, more importantly, some properties do not concern cells, but processes in which they are involved, e.g. cell division or cell death.

5 CELLULAR GENEALOGY ONTOLOGY (CGO)

The Cellular Genealogy Ontology (CGO) is a part of a conceptual framework underlying the annotation schema developed for the purpose of structuring and annotating experiment and simulation results, obtained in frames of the research on the cellular genealogies which is based on the SPOO ontology.

The development of CGO is work in progress, though, in its current state it permits already the description of cellular genealogies and their components in context of research activities which produce them. The main notions of CGO are briefly discussed in the current section.

A *Cellular Genealogy* is considered as a product (called *Cell Tracking Product*) of a *Cell Tracking Process*. Two types of Cell Tracking Processes are considered, namely *Cell Tracking Experiment* and *Cell Tracking Simulation*. The former corresponds to time lapse experiments, the later to the simulations of in silico cell cultures. This distinction permits to handle with the same representation schema both the genealogies simulated and observed and yet distinguish them. Currently, a cell tracking process and cell tracking product are merely placeholders which at the later phases of the ontology development will be extended for proper handling of data describing the social context of research process such as e.g. scientist and lab data. For this purpose integration of CGO with existing schemas such as e.g. Dublin Core is planned (Dublin Core Metadata Intiative 2010).

A Cell Tracking Process results in a sequence of *Frames*. Each Frame is a presential situation describing a cell colony at a particular moment of time. In case of Cell Tracking Experiment a Frame depicts a photo of a colony taken. Each Frame consist of zero to many *Presential Cells*. A presential cell is a presential object being a part of presential situation. Presential cells can be depicted by qualities such as e.g. *Position* or *Shape*. The SPOO mechanism of characteristics permits a user to define additional qualities of presential cells. A presential cell is characterized not only by presential qualities but also by roles it plays participating in presential relations such as e.g. *Cell-Cell Contact*.

Presential Cells belonging to different Presential Frames can be related by abstract n-ary relation of *Abstract Link* which comes in three types: *Succession*, *Division* and *Fusion*.

Succession is an abstract link between exactly two presential cells located at different Frames indicating that both cells are considered to represent the same time extended cell. In context of succession the older presential cell is called *Predecessor* and the later - *Successor*. Division is a relation which links a single presential cell located at the former frame and called a *Parent Presential Cell* with two cells in the later Frame called *Daughter Presential Cells*. Finally, fusion is a relation gluing two presential cells at a former frame with one at a later frame. All of the above relations come with their own characteristics such as e.g. *probability*, *confidence of human expert*.

Out of presentic cells and their abstract interrelations can be constructed cellular genealogies. A Cellular Genealogy is considered as a SPOO situation, that is a complex time extended entity in which other entities participate. Two types of entities participate in Genealogical trees, namely, *Cells* and *Cellular Processes*.

Cells are represented in CGO as objects, i.e. entities existing in time and to some extent independent of their background and of the processes in which they participate. Cells are constructed out of a chain of presential cells linked with succession relation. Each cell can be characterized with a number of characteristics such as qualities (e.g. *morphology*, *shape*, *lineage assignment*) and roles played by cells in relations. Qualities of cells can be either calculated out of qualities of corresponding presential cells e.g. *velocity* or can be genuine time extended qualities. Cell qualities can be time-parameterized by the temporal location which permits to document a quality value changes overtime.

Among the qualities of cells is *Cell Generation* which organize them in a genealogy. Cell Generation is itself characterized by such qualities as *Division Probabilities* (*Asymmetric*, *Symmetric*, *Undifferentiated Symmetric*) and *Cell Death Index*.

Cells participate in two types of processes: *Cell Division Processes* and *Cell Death Processes*. Cell division process is a process operating on/consuming one cell called *Parent Cell* and producing two *Daughter Cells*. Both the parent cell and the daughter cell are roles of a cell in context of a cell division process. The process of cell division is constructed out of abstract relation of division. It is worth mentioning that in CGO the abstract division relation between presential cells is distinguished from the process of division. The first is mere representation of the fact that instead of one cell at a frame two cells were observed, whereas the second represent a biological process of cell division which can be

further characterized. For example, a division process can be characterized by at most two different division classes with respect to the chosen *view on cell fate identification.* In CGO are introduced two views - *Prospective* and *Retrospective View*. The chosen view has also an influence on the cell lineage assignment, discussed above.
The notion of *Cell Death Process* indicates a process of cell death which operates on one cell.
Root Cell of a cellular genealogy is the cell from which the observation of a cellular development starts and which is a trunk of a genealogy. Technically, the root cell is a role of a cell in context of a genealogy. Out of a root cell new cells are developed by means of the *Cell Division Processes.*
In addition to the discussed above structure of a genealogies and their participants, a number of qualities depict a genealogy, namely *total number of leaf cells, total number of divisions, range of branch lengths, symmetry indices, generalized cell death index.*

6 CONCLUSION

In the current paper is discussed a work in progress on a framework for modeling and representing data on cellular genealogies. The framework consist of two ontologies – a top level ontology of SPOO which is a part of GFO tailored for conceptual modeling and the domain ontology of CGO. The former provides a general structuring principles and handling of cross-domain general notions such as object, process and characteristic. Its primary goal is to provide a design patterns for constructing well-structured and easily extensible domain ontologies. The later is a domain ontology providing the vocabulary for describing results of time lapse experiments and simulations.
Currently, the framework is utilized for first applications such as:

- An object-oriented domain model and database schema for cell tracking software developed within the frames of EuroSyStem Project (EuroSyStem Project 2010).
- An export/import format for software tools developed for analysis of time lapse experiments.

The ontology currently is tailored mainly to satisfy the requirements for the purposes of the DynaMo Research Group (DynMo). However, in the second step additional standardization effort and cooperative work with other groups working on cellular genealogies is necessary.

In addition, number of issues concerning the structure of ontologies require further work. In particular these are patterns of abstract links between presentials cells and integration with current biological ontologies e.g. Cell Type Ontology (Bard J, Rhee SY, Ashburner M 2005).

ACKNOWLEDGEMENTS

Research for this paper has been realized within EuroSyStem Project (EuroSyStem Project 2010).

REFERENCES

Bard J, Rhee SY, Ashburner M (2005). An ontology for cell types. Genome Biol. 2005;6:R21.

Diehl, Alexander, Augustine, Alison, Blake, Judith, Cowell, Lindsay, Gold, Elizabeth, Gondré-Lewis, Timothy, Masci, Anna Maria, Meehan, Terrence, Morel, Penelope, Nijnik, Anastasia, Peters, Bjoern, Pulendran, Bali, Scheuermann, Richard, Yao, Q. Alison, Zand, Martin, and Mungall, Christopher. (2009) Hematopoietic Cell Types: Prototype for a Revised Cell Ontology. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2009.3635.1>

Dublin Core Metadata Intiative (2010). Available <http://dublincore.org/>, cited 2010.

DynaMo http://tu-dresden.de/die_tu_dresden/fakultaeten/medizinische_fakultaet/inst/imb/forschung/MM_MS

Eilken, H. M., S.I. Nishikawa, and T. Schroeder. Continous single-cell imaging of blood generation from haemogenic endothelium. *Nature* 457, 896 (2009)

EuroSyStem Project (2010). Available <http://www.eurosystemproject.eu/>, cited 2010.

Glauche, I., M. Cross, M. Loeffler, I. Roeder. (2007) Lineage specification of hematopoietic stem cells: Mathematical modeling and biological implications. *Stem Cells* 25, 1791 (2007)

Glauche, I: *Dissertation* Glauche, I. Theoretical studies on the lineage specification of hematopoietic stem cells, Dissertation, University of Leipzig, 2010

Herre, H. General Formal Ontology- A Foundational Ontology for Conceptual Modeling. In: *Theory and Application of Ontology*, Ed R. Poli et al. Springer (2010)

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. (2003) WonderWeb Deliverable D18. Ontology Library (final). Version 1.0. Laboratory For Applied Ontology, ISTC-CNR, Trento, Italy.

Roeder, I. m R. Lorenz. Asymmetry of stem cell fate and the poö tential impact oft he niche observations, simulations, and interpretations. *Stem Cell Rev* 2, 171 (2006)

Scherf, N., I. Roeder, I. Glauche. in: Proceedings of the Fifth International Workshop on Computastional Systems Biology. *WCSB* 2008 pp. 161-164 (2008)

Schroeder, T. Imaging stem-cell-driven regeneration in mammals. *Nature* 453, 345 (2008)