Ontological Levels in Histological Imaging

Antony GALTON [a,1], Gabriel LANDINI [b], David RANDELL [b] and Shereen FOUAD [b]
[a] *Department of Computer Science, University of Exeter, Exeter, UK*
[b] *School of Dentistry, University of Birmingham, Birmingham, UK*

Abstract. In this paper we present an ontological perspective on ongoing work in histological and histopathological imaging involving the quantitative and algorithmic analysis of digitised images of cells and tissues. We present the derivation of consistent histological models from initially captured images of prepared tissue samples as a progression through a number of ontological levels, each populated by its distinctive classes of entities related in systematic ways to entities at other levels. We see this work as contributing to ongoing efforts to provide a consistent and widely accepted suite of ontological resources such as those currently constituting the OBO Foundry, and where possible we draw links between our work and existing ontologies within that suite.

Keywords. histology, histopathology, histological image processing, digitised images, microscopy, histological stains, segmentation, mathematical morphology, discrete mereotopology, spatial logics, formal ontologies

1. Introduction: Background and related work

For the purposes of the present paper, a histological image is a digital scan of a prepared histological sample typically obtained from a living organism, through biopsy sampling, surgical excisions, smears, or other techniques, for diagnostic (histopathology) or anatomical (histology) analyses. While the histologist will typically use such images to investigate cell and tissue structure (morphology),[2] the histopathologist is interested in extracting from such images information about features that may provide indications of the presence or absence of various pathologies in the anatomical entities from which the images are derived. Highly-trained histopathologists are often able to identify and classify such features through visual inspection of the images, aided by extensive experience of working with similar images, and informed by a rich background of theoretical understanding. Given the time and expense of training such individuals, it is desirable to simulate as much of their expertise as possible by developing systems for performing at least part of the analysis automatically, thus ensuring a much higher throughput of analyses, with the potential for improved patient outcomes.

[1]Corresponding Author: Antony Galton, Department of Computer Science, University of Exeter, Exeter, UK; E-mail: apgalton@ex.ac.uk.

[2]Note, however, that while traditionally histology has focused on morphology, latterly histochemistry and immunohistochemistry have provided functional insights too, for example into the types of substances produced by cells or the genes they express.

To develop such a system it is necessary to devise automatable methods that can be used as a surrogate for the (prima facie "mysterious") exercise of judgment by a trained histopathologist. We need ways of algorithmically classifying or quantifying features of histological images in order to obtain judgments comparable to (or perhaps even better than) those provided by the experts. Such algorithms depend on our finding ways of representing the information that can be extracted from the images; and we have found that in describing such forms of representation we uncover a world of surprising ontological richness. The purpose of this paper is to provide a tour of the main features of this world.

In our work we focus attention on the quantitative and algorithmic analysis of digitised images of cells and tissue and their respective parts and structures in the pathologies of oral cancer.[3] We are particularly interested in developing imaging methods where the conformity to theoretical models of cell or tissue structure is considered necessary in order to extract reproducible and meaningful morphological and diagnostic information. Examples of this approach in our own research include assessment of the advance front of cancerous tissues as a measure of invasiveness [13,1,2], analysis of cell populations [3], and modelling of tissue architecture related to the hierarchical spatial arrangement of cells [11,12]. More recently [14], we have shown how image processing operations such as erosion and dilation from classical Mathematical Morphology (MM) [15] can be integrated within the spatial logic Discrete Mereotopology (DM) [8,9], which is used to encode topological information extracted from digitised histological images and to provide a theoretical framework on which to pin histological models. In [14,7], DM is used not just as a representation language but a spatial logic enabling symbolic automated reasoning programs to be integrated with conventional imaging algorithms where theorems provable in DM are used to justify choices made in implemented imaging algorithms.

Recently Smith *et al.* [17] have advocated the development of a Quantitative Histological Imaging Ontology (QHIO) which will provide "the resources for the creation of annotated bodies of data, which will be discoverable and shareable across institutions", and they draw attention to the "pipeline" beginning with raw pixel data, running through operations such as segmentation, registration, feature extraction, and statistical analysis, and ending up with the provision of "links to external ontologies to annotate this data".[4] In this paper we are concerned with just this pipeline in the context of histological imaging, conceptualising it in terms of a series of ontological levels characterised by distinct classes of entity. Of particular relevance here are the biomedical ontologies making up the OBO Foundry [16] and in particular the Cell Ontology, the Ontology of Biomedical Investigations, and the Information Artifact Ontology, discussed in section 3 below.

The remainder of this paper is structured as follows. In §2, we introduce the series of levels that form the main backbone of this work, then in §3 we provide some details of the ontologies associated with each level, describing some of the classes of entity characteristic of a level and the relationships between entities at the same level. In §4 we describe the relationships between the levels, with particular reference to the processes by which entities at one level are derived from those at lower levels. In §5 we provide an illustrative example, and in §6 we round off the discussion with some remarks to set our work in a broader ontological context.

[3] Quantitative histological imaging is a well-established active research area that can be be traced back many decades; for a general overview see [5].

[4] See [6] for an overview of current biological imaging tools.

2. Ontological levels in histological imaging

We have found it helpful to think of what we are doing in terms of a number of different *levels*, representing the stages in deriving histological judgments from the underlying reality. Each level has its own ontology, and the ontologies belonging to different levels stand in more or less systematic relationships to each other.

2.1. Level 0: Physical reality

Level 0 is the level of physical reality, inhabited by actual biological entities, e.g., tissues, initially in situ in the living organism, but after extraction prepared in the form of thin stained slices or films, resulting in *histological preparations* or *samples*. Various different stains can be used, a widely-used combination being haemotoxylin and eosin (H&E), which selectively bind to different cell components, staining cytoplasm pink and cell nuclei dark blue or purple. Although they have been manipulated in ways that do not occur in vivo, these prepared samples still form part of the physical reality that is the target of our investigations: what is on the slides is still ontologically independent biological material, albeit in a form somewhat removed from its living state.

Biological material is three-dimensional in nature: a cell is a 3D object and even the thinnest layers of cells are 3D. The prepared samples are also 3D, but since the thickness of a section is much less than the typical diameter of a cell, one can conceptualise the section as being to all intents 2D, in the sense that all the significant variation of interest in the section is along the plane of the section, rather than orthogonal to it. This effective two-dimensionality forms the basis for the transition from level 0 to level 1.[5]

2.2. Level 1: Captured images

Level 1 contains images initially resulting from digitally scanning histological preparations at level 0. An image is a rectangular[6] array of pixels, each of which has a position, specified by coordinates (x,y), and a colour value, e.g., an RGB triple. Unlike the histological sample itself this image is strictly 2D, but its structure, as defined by the spatial distribution of its colour values, stands in systematic relationship to the sample, being derived from it in a way that has been engineered to ensure that such a relationship holds.

How the image is related to the sample depends on the resolution of the imaging process: each pixel in the image is essentially a point and its colour value represents some kind of averaging of the light transmitted by an *area* within the sample; any colour variation across that area is not differentiated within the image. Production of the initial image typically includes various "pre-processing" operations designed to eliminate image artefacts arising as systematic consequences of the imaging technology and how it is used. Such artefacts include "hot" and "dead" pixels, sensor noise, uneven illumination from vignetting, amp-glow, and many others. By "initially captured image" we understand the image obtained after any such preprocessing operations have been performed.

[5] Some imaging applications such as tomography yield three-dimensional entities at level 1: but even these can be regarded as stacks of two-dimensional "slices", and we fully expect our account of ontological levels to cover these cases too. In this case the "slices" are virtual (level 1) rather than physical (level 0).

[6] Although early image processors often used hexagonal lattices, the greater ease of computing with rectangular arrays has led to their more general adoption; here we assume that all digital images are of this kind.

Further operations performed on initially captured images can facilitate extraction of useful information from them. Colour deconvolution can be applied to RGB images to unmix contributions from stains using more than one dye, e.g., to separate eosin and haematoxylin channels from an H&E-stained tissue section. These "stain channels" are still level 1 pixel arrays but provide stepping-stones to the generation of level 2 entities. Another example is the computation of morphological domes [18] to identify dark basins in image intensity in preparation for segmenting out regions of interest at level 2.

Although we have referred to "an" image, in practice we are usually dealing with a stack of images aligned coordinate-wise ("registered"), constituting different representations of the same histological sample. Such registered stacks arise in several ways, depending on different image modalities. In light microscopy, colour images are commonly encoded as RGB images comprising three sub-images, one for each broad-band filter used. These images may be perfectly aligned when captured, e.g., using a tunable filter which takes consecutive shots with three different filter settings; or initially interleaved, e.g., using a Bayer mask, colour alignment being achieved through interpolation. The stain channels referred to above will also form a registered set. Another common method, used in confocal microscopy, is to stain sections with fluorescent molecules linked to molecules (antibodies, lectins, etc) that bind to specific targets. To visualise such bindings, lasers of different wavelengths are used to excite the fluorescent molecules, generating images of one molecular species or marker at a time. By representing the emission of excited fluorescence such multi-channel images capture different aspects of the morphology of the same sample according to molecular species distribution.

Unlike the histological preparations they depict, level 1 images are not physical objects: they are *information artefacts*. As such, they can exist in multiple copies, inhering in different physical bearers, e.g., coloured spots on a computer screen, coloured ink patches on a piece of paper, or states in a computer memory. The image itself is not any one of these physical bearers but rather the abstract pattern that they share. It is two-dimensional, not because it exists in 2D space, but because each of its constituent elements (pixels) is assigned a 'position' given by two independent numerical coordinates.

2.3. Level 2: Segmented images

Level 2 arises from the images at level 1 by segmentation, i.e., the process of carving out from the image collections of pixels which "belong together" in some way, forming regions. Although segmentation is performed algorithmically on the basis of the colour and intensity attributes of the pixels, without reference to the underlying biological reality at level 0 from which the image was derived, its purpose is to assist in the identification of areas in the image corresponding to significant entities in that underlying reality.

From a stack of registered images, different segmentations can be built by varying parameters or applying different algorithms. A typical technique is to convert the image into grey-scale and apply a threshold so that pixels with grey-levels at or below the threshold are classified as "foreground", the rest as "background"; one then achieves a segmentation by identifying each connected component of the foreground as a region of interest. Morphological operators such as erosion, dilation, opening, or closing are often applied first to obtain a segmentation that is judged to be more meaningful.

By the "meaning" of a segmentation we understand some such labelling as "candidate nucleus image" or "candidate cytoplasm image" applied to regions within the seg-

mentation. Candidate nucleus images are image segments which to a first approximation correspond to the parts of the image derived from actual cell nuclei in the sample. Such an interpretation links levels 0 and 2. Level 1 effectively represents an "informational bottleneck" through which information latent at level 0 is reduced to raw pixel data, from which some of that information must be reconstituted, as best we can, at level 2. This general picture is not, of course, peculiar to histological imaging but is common in some form or other to imaging processes generally, including the human visual system.

Although level 2 regions are selected for their *prima facie* histological significance, they may fail to conform to the requirements of a histological model. For example, in a valid model we normally expect to find cell nuclei and cytoplasmic components in one-to-one correspondence, with each nucleus located within the boundary of its cytoplasm, forming a single cell, with no overlap between cytoplasmic components of different cells. In the segmented images, however, such constraints may be violated in various ways. A candidate cytoplasm image might contain no candidate nucleus image, or have more than one. A candidate nucleus image might only partially overlap its candidate cytoplasm image. Such anomalies can arise for various reasons. Since the sample is only a thin slice of 3D material, it does not contain whole cells but cell slices, which may, for example, omit the nucleus. In tissue with closely-packed cells, boundaries between adjacent cells can become obscured, giving the impression of single large multi-nucleated cells. There can be isolated fragments of cytoplasm arising from imperfections in the slicing procedure, or purely optical artefacts not filtered out at the pre-processing stage.

Not all apparent anomalies at level 2 arise from faulty segmentation, however; some reflect the fact that the cells in reality (i.e., at level 0) are themselves anomalous in some way, for example a cell that is on the point of dividing can present an anomalous appearance because in the mitotic prometaphase the nuclear membrane disintegrates and the nucleus ceases to exist as a distinct compartment until the cell reaches the telophase.[7] It is important to distinguish between anomalies of these kinds.

To achieve a valid histological model it is necessary to correct anomalies arising from faulty segmentation. There are various methods for doing this, e.g., generating "virtual cells" (which we call *v-cells* [11]) using candidate nuclear segments as seeds for a watershed transform to derive regularised cell boundaries, or, alternatively, starting with regions in differently segmented registered images, labelled as nuclear and cytoplasmic, and then adjusting the images as necessary to achieve the expected correspondence between nucleus and cytoplasm. The resulting image comprises image segments (level 2) that can be directly interpreted as identifying elements of a valid histological model. It is by the addition of these identifications that we can finally pass from level 2 to level 3.

2.4. Level 3: Histological models

Once we have achieved an image segmentation which allows an interpretation consistent with histological theory we can attempt to move beyond the domain of segmented images *simpliciter* to produce a *histological model*. By this we understand a collection of level 2 image segments, each annotated by some level 0 category conformably to the ex-

[7] Other examples of such anomalies include *anucleated* cells such as erythrocytes (red blood cells) which lack nuclei altogether (so that some would say that they are not strictly cells in any case), and *multinucleated* cells such as striated muscle cells, osteoclasts, Langhans cells, and foreign-body giant cells, as well as abnormal cells such as the Reed-Sternberg cells, each with two nuclei, found in Hodgkin's lymphomas.

pectations of histology. Thus a histological model element may be represented as an ordered pair comprising a level 2 image segment and a level 0 category, such that the image segment may consistently be interpreted as depicting an instance of that category—the ensemble of all such individual interpretations in the model being consistent with histological constraints. Thus, for example, a "model cell" is an image segment that can be understood as depicting an actual cell (i.e., an instance of the level 0 category "cell")—such an understanding being possible so long as the shape and size of the image segment and its spatial relations to other such segments as well as to subsegments corresponding to model nuclei and model cytoplasms are consistent with theoretical expectations.

Although a histological model at level 3 must be theoretically correct in the sense of not violating histological constraints, it may still fail to be *factually* correct since the identification of image segments as depicting particular types of level 0 entities may be wrong—for example, a segment might be annotated as depicting a nucleus whereas in reality it arose as an imaging artefact or from some foreign body in the sample. Not all such errors will be filtered out in the production of a level 3 model. What is required, of course, is a way of engineering the process of getting from level 0 to level 3 in such a way as to minimise the probability of such mismatches occurring.

We believe that methods to integrate level 2 information into level 3 models have not been given enough attention, perhaps owing to a lack of clear understanding of the relations between the various levels. One consequence of this is the pervasive absence of level 3 models in computational histological analyses. These should be extremely useful to provide higher-level histologically relevant descriptions of image contents that can be algorithmically extracted.

3. Entities and relationships within the levels

Each of the four ontological levels we have identified constitutes a domain with its own characteristic set of entities and relationships, i.e., its own ontology. In this section we outline the main features of the ontology of each level, and in the next section we discuss relationships between the levels.

3.1. Level 0 ontology

The ontology at level 0 is populated by real-world entities. For these we may turn to the ontologies making up the OBO Foundry[8] — most importantly, for our purposes, the Cell Ontology.[9] To get to level 1 from level 0, living material may need to be sampled and the samples prepared for imaging, e.g., by mounting thin slices on slides and applying histological stains or immunohistochemical markers (see below). The result is still a part of physical reality, and therefore level 0, but in an artificially regimented state suitable for controlled digital imaging. We might think of it as a notional "Level 0.5", and as such, it is described in terms of a number of classes not present in the Cell Ontology itself.

[8] `www.obofoundry.org`

[9] `www.obofoundry.org/ontology/cl.html`. Although for the sake of potential future interoperability we follow the Cell Ontology here, it should be noted that not all the classification decisions embodied in that ontology are uncontentious. In particular, our level 0 ontology could be simplified somewhat if we followed the widespread practice of regarding blood as a tissue, contrary to the Cell Ontology.

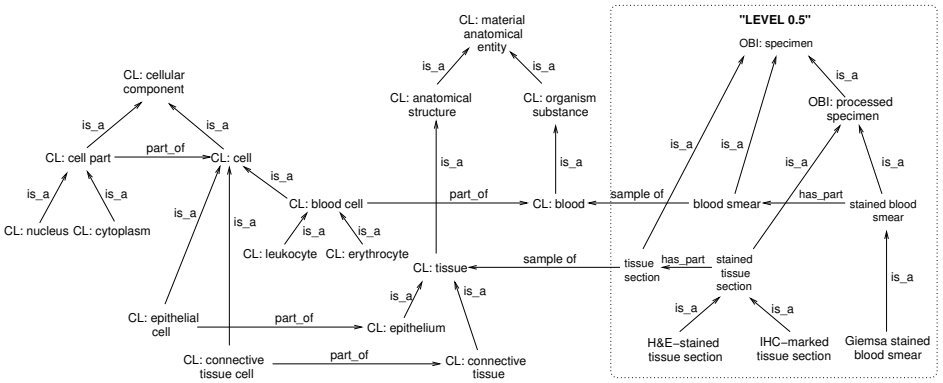


Figure 1. The ontology of Level 0. This includes a small section of the Cell Ontology and, on the right, some additional classes relating to prepared histological samples (Level 0.5). Note that some of these *is_a* links comprise a chain of such links in the Cell Ontology, with intermediate classes not represented here.

An outline of the Level 0 ontology is presented in Figure 1. This includes a number of classes from the Cell Ontology that are relevant to our work on the oral mucosa; in accordance with normal practice, these carry the annotation "CL". For simplicity's sake we only mention two tissue types, in effect presenting a very broad-brush account of oral mucosa. In reality epithelium may also contain non-epithelial cells such as melanocytes, lymphocytes, Langerhans (antigen-presenting) cells, Merkel (sensory) cells, and possibly others. In skin, there are sebaceous glands and hair follicles, which are also epithelial in origin. Some of these cell types are not detectable with H&E staining, but require special markers for their detection. We also include blood and blood cells, used for our illustrative example in §5. The classes in "Level 0.5" are shown on the right-hand side of the figure. These cover prepared histological samples, which we refer to the class *specimen* from the Ontology of Biomedical Investigations.[10] Within this class we distinguish tissue sections, which are samples of tissue, from stained tissue sections, which are not, since they contain stains in addition to the tissue itself; and likewise with blood smears and stained blood smears. In addition to tissue sections and smears, histological preparations include imprints, fine-needle aspiration biopsies, cell cultures, tissue cultures, etc.

Although we refer here mainly to stained samples this should not be understood too narrowly. An alternative to standard histological dyes is the use of immunohistochemical (IHC) markers, by means of which cells can be labelled according to whether they contain specific molecular species. Epithelial cells, for example, can be identified by means of IHC staining with anti-keratin antibodies; fibroblasts (a kind of connective tissue cell) can be identified using anti-vimentin antibodies. There are thousands of such markers, all of which would have to be included in a full account of level 0.5.

3.2. Level 1 ontology

Level 1 entities are not material objects: they are *information artefacts*. We follow the widespread understanding of these as *generically dependent* entities: they depend for their existence on being borne by some material entity, but a given information artefact

[10] `http://obi-ontology.org/page/Main_Page`

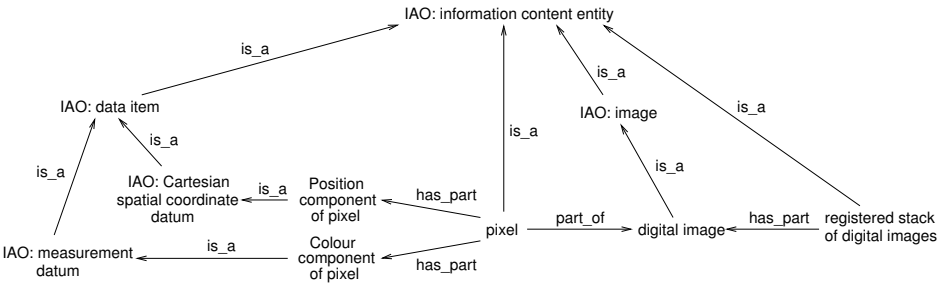


Figure 2. The ontology of Level 1. The elements here are information artefacts, so we use the Information Artifact Ontology (IAO) to classify them.

is not dependent on any one such entity—rather, it can have numerous "copies", each borne by a different material object [10,4]. A digital image, for example, can be borne by a computer disc, a screen display, or a print-out. An appropriate ontology for handling these entities is the Information Artifact Ontology.[11]

The elements at level 1 are linked to those of level 0 in two ways. A digital image is *derived from* a prepared sample by means of an imaging process that is designed so that the resulting image can be regarded as *depicting* that sample. How faithfully it does so depends on the details of the imaging process, which is engineered to maximise the salient information about the sample in the image. Extracting that information is, of course, what motivates the transition from level 1 to levels 2 and 3.

We include in the ontology here the category *registered stack of digital images*. This is a set of images all of the same histological sample, but with different stains; they are registered in the sense that their pixel grids are aligned, i.e., the pixels at a given position in the grid in all the images of the stack are derived from the same portion of the sample.

3.3. Level 2 ontology

After segmentation, we still have an image, but it is now conceived as made up not of pixels but of regions, each obtained by grouping together certain pixels of the level 1 image. This is a *segmented image*. In the ontology, these regions are classified as *image segments* (see Figure 3). They are not regions in the sense of being actual regions of space; rather, they are still information artefacts capable of multiple realisation in different physical forms. Informally, it is natural to call them regions, and there is no harm in doing so as long as a more correct terminology is adopted when formulating the ontology in a precise fashion. Image segments may be subclassified by type—as noted above, the purpose of segmentation is to identify parts of the image corresponding to histologically relevant entities, and they are therefore, perhaps tendentiously, labelled with reference to those entities as candidate nucleus images, candidate cell images, etc.

3.4. Level 3 ontology

Level 3 represents the point at which, ideally, we have a model that is fully conformable to our understanding of histological reality. Each element of the model is derived by as-

[11] `http://bioportal.bioontology.org/ontologies/IAO`

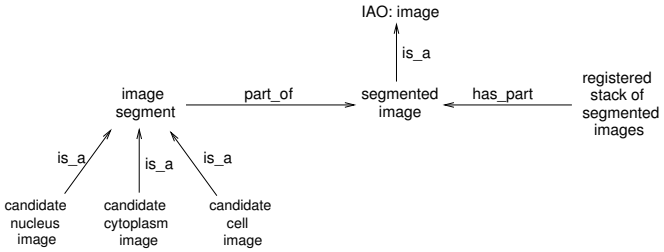


Figure 3. The ontology of Level 2. Image segments correspond to what we informally refer to as regions.

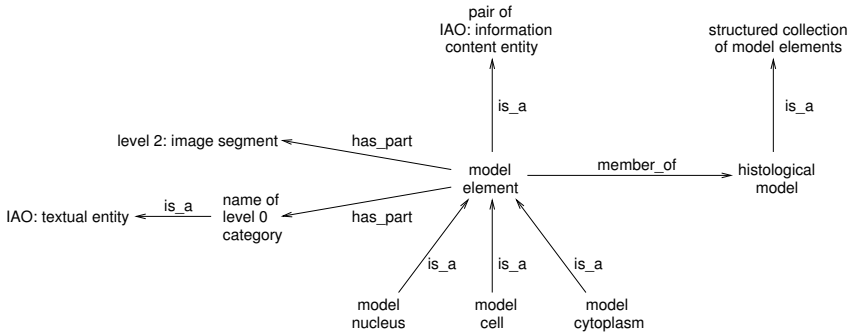


Figure 4. The ontology of Level 3. Model elements together constitute a coherent histological model.

sociating an image segment from level 2 with a histological category from level 0; the idea is that the image segment is interpreted in the model as depicting an element of the associated category (e.g., cell, nucleus, epithelium). We represent these elements as composites comprising these two components. Thus a model element is specified as an ordered pair in which the first component is an image segment and the second component is the name of a histological category.[12] As such, the entities at level 3 are still information artefacts; they are information artefacts defined in such a way as to enable us to reason about the underlying reality by reasoning about the model that they constitute.

An important feature of level 3 (see Figure 4) is that the model elements together constitute a coherent histological model. By this we mean not just a collection of model elements but a *structured collection*, in which the spatial relations between the image segment components of different elements are in conformity with the expected spatial relations between instances of their histological category components. This conformity is secured by the fact that in order to be included as components in a level 3 model, image segments at level 2 have to pass appropriate tests, being manipulated in various ways if necessary to ensure this. For example, in a model nucleus, the categorical component will be the level 0 category-name "CL:*nucleus*", and the image segment component will normally be a proper part of the image segment component of some model cell.

[12]Note that the category "name of histological category" cannot be modelled as, e.g., IAO: *written name*, as the latter is restricted to entities denoting particulars, not universals. The Information Artifact Ontology does not appear to include any category for category-names that is more specific than *textual entity*.

4. Relations between the levels

In describing ontological relations between the levels an overriding consideration is their *processual* nature: each level is derived from its immediately lower level by some set of processes which determine, at least in part, its ontological character. These processes are occurrents rather than continuants; we mention some of them here, but do not include them in the formal ontology: this is future work.

4.1. Relations between level 0 and level 0.5

The relevant processes here are the methods for preparing samples for imaging, including extraction of samples from living tissue, preparation and mounting of sections from samples, and staining or marking of samples. There are many ways of extracting biological material from an organism, including for example both biopsy (sample is extracted by cutting) and smears (extracted by scraping). Different methods may preserve different features of the original—in smears, for example, one does not see the tissue architecture relations or layers that are seen in sections, instead there are just single cells or small clumps that do not necessarily show the same relations as exist in a biopsy or in vivo.

After the histological sample has been extracted and mounted, it will then undergo the final preparations for imaging, including for example application of a histological dye or IHC marker. The resulting prepared sample can be characterised as belonging to an appropriately designated type such as, for example, *H&E-stained_tissue_section*, which is a subclass of *stained_tissue_section* identified in terms of its method of preparation.

4.2. Relations between level 0.5 and level 1

The processes defining the relations here are the imaging processes. In detail, a huge amount could be said about this, but the overall picture is that we end up with an image—as noted, an information artefact—that can be said to *depict* the histological sample from which it was derived. Ontologically, the depiction relation can have as many subtypes as there are different imaging techniques. Of paramount practical importance is the general issue of *image quality*, broadly understood as a measure of the extent to which an image may be regarded as a faithful depiction of the level 0 entities it is supposed to depict. This is a function of the imaging process, which in turn depends on not only the available technologies but also the precise experimental set-up used, including e.g., focus, illumination, resolution, image compression, RGB Bayer mask artefacts, chromatic aberration, and colour quantisation, all of which can impact on the quality of the image. This in turn determines what information it is possible to transmit to levels 2 and 3.

Although an image depicts the sample by virtue of the values of its constituent pixels and the processes by which they were assigned, the individual pixels themselves do not depict anything. Pixels in an image are roughly analogous to letters in a text: the letters on their own carry no meaning, but the meaning of the text as a whole depends on the letters it contains. The analogy is not exact, however, since the colour value of a pixel is derived from the part of the sample located in the position corresponding to its position value, whereas the association of meanings with particular combinations of letters in language is largely arbitrary and conventional. The different images in a registered stack depict the same sample in different ways, and again, they do this by virtue of comprising pixels whose values were obtained by different means from the original sample.

4.3. Relations between level 1 and level 2

An image segment at level 2 is derived from a collection of pixels at level 1, which are picked out by the particular image-processing operations applied. These may include thresholding, smoothing, or non-linear increase of contrast, any of which could be applied either locally to smaller parts of the image or globally. This may reduce the information in the image although paradoxically that may be exactly what is needed for extracting the information we are interested in, like getting a clearer view of the wood by suppressing some of the information about the trees. Different procedures can be applied to the same or different images in a stack, to produce a new stack whose members are not necessarily in one-to-one correspondence with the original.

Level 2 entities (image segments) have properties that do not directly derive from any analogous properties of level 1 entities (pixels). Level 2 is the first level where it is meaningful to refer to individual features of regions such as shape, size and topology; computationally, this involves different data structures. With candidate nucleus images at level 2, we can already look for features of nuclei (size, shape, etc) that conform to the expectations of histological theory; the identification of such features plays a crucial role in promoting the candidates to labelled nuclei at level 3.

It is worth noting here that while conceptually there is a unidirectional progression from level 1 to level 2, the actual algorithms employed for accomplishing this transition can involve iterated back-and-forth translations between these levels.

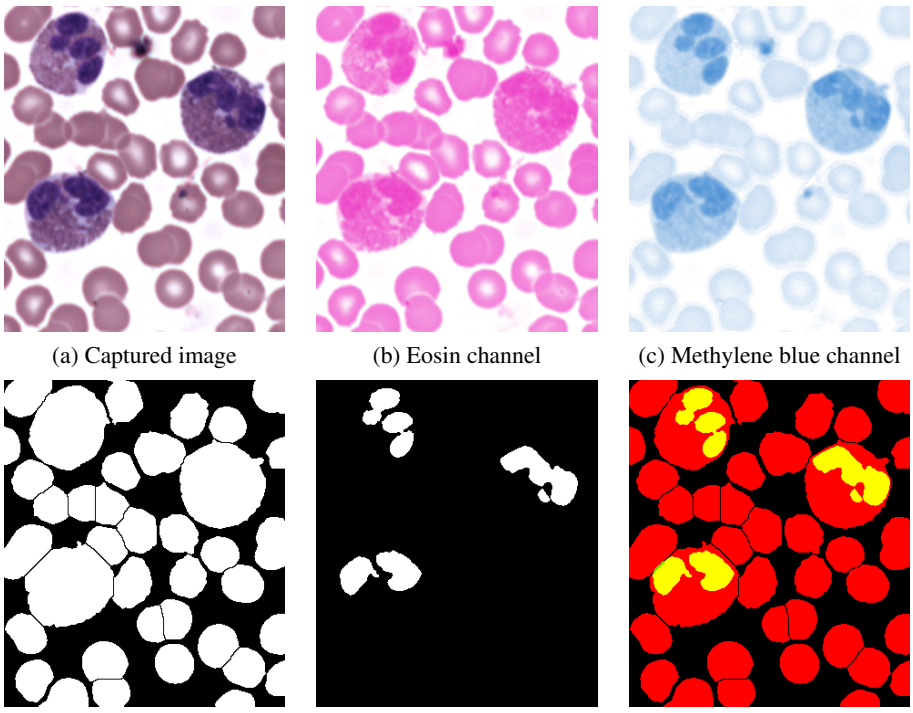
4.4. Relations between level 2 and level 3

At level 2 the relations between, say, nuclear and cytoplasmic image segments may not conform to histological constraints, and hence these segments cannot immediately provide models of nuclei and cytoplasm respectively; the transition between levels 2 and 3 is accomplished by testing features and relations of candidate segments, adjusting them where necessary, so that, after annotation with level 0 category-names, they can provide histologically consistent models. Such models can correctly capture salient properties of level 0 reality and support reasoning concerning metric and topological relations between cells, layering or clustering of cells, and other features indicative of specific pathologies. This transition is in many ways the most challenging of the whole process, and forms a major technical focus of our research.

5. Illustrative example

In Figure 5 we show how the ontological levels relate to image computation operations when querying an image. Figure 5 depicts key stages in the segmentation of a Giemsa stained (Eosin and Methylene Blue) human blood smear. The image shows eosinophil leukocytes (with lobulated nuclei), erythrocytes (blood cells with no nuclei), and small platelets. The task is to identify the two main cell types. This is achieved by a series of operations and tests to check whether the interpreted image segments conform to a set of constraints arising from our assumed background histological model.

First the captured image (a) is separated into different stain channels (b and c) using colour deconvolution. These images are still purely at level 1. Then we take the



(a) Captured image (b) Eosin channel (c) Methylene blue channel

(d) Cytoplasm regions (e) Nuclear regions (f) Model cell candidates

Figure 5. Stages in the identification of model cells in a blood smear. In (f), cytoplasm is shown in red, nuclei in green, and overlaps between them in yellow (dark grey, light grey, and white in monochrome printed versions of the paper).

eosin stain image (b), fill the grayscale holes and threshold the result using an automated (Huang) method. A watershed operation on the binary image separates potential merged cell regions and any holes are filled, yielding level 2 segments.

Regions smaller than a pre-determined size are deleted, leaving candidates for annotation as either erythrocytes or leukocytes (the white segments in (d)). Next the methylene blue image (c) is thresholded using the minimum method, holes are filled, and regions smaller than a specified size are deleted (e), leaving candidates for annotation as nuclei. To test for model leukocytes, we compute the spatial relations between cell and nuclear regions using Discrete Mereotopology [14] (e.g., whether a candidate nucleus forms a part of or partially overlaps some host cell) and the cardinality of the mappings between nuclei and cells (e.g., eosinophil leukocytes have cell nuclei but blood cells do not). Cell/nucleus pairs that pass these tests can be annotated as leukocytes, giving model cells at level 3. The remaining cells, not paired with nuclei, are annotated as erythrocytes.

If all the tests succeed, the segmentation represents a model, meaning that the segments can be interpreted as depicting histological entities classified in level 0's ontology and named accordingly. In image (f) (a composite of images (d) and (e)) we see a case where the attempt to generate a valid model has met with only qualified success. The smaller red segments containing no yellow inclusions can be unproblematically modelled as erythrocytes; the three larger segments are candidate leukocyte images but some

anomalies remain. First, each of the two left-hand segments appears to show a cell with two nuclei; this appearance is deceptive, arising from the characteristic lobulated form of the eosinophil leukocyte nucleus, which often shows apparently disconnected components in images. Second, close inspection of the lower left cell shows that its nucleus partially overlaps the cell body, the external part showing up as a thin green line filling part of the gap between the leukocyte and a neighbouring erythrocyte. Both these anomalies need to be corrected or otherwise allowed for in order for the image to support a valid histological model. For an example of such correction, see §5.3.3 of [14].

It should be emphasised here that it is the explicit use of these model-based tests of conformity that enables one to identify anomalies in the segmentation process which otherwise may go unnoticed using 'blind' pixel-based histological imaging segmentation routines. With large histological image datasets now comprising whole-slide images and tissue micro-arrays (TMAs) there is increasing need for automated high-throughput methods; this is where integration of formal ontologies for histology and histopathology with practical aspects of histological image processing has a clear role to play.

6. Concluding remarks

Our purpose in this paper has not been to propose a fully-fledged ontology of histological imaging, but rather to advocate the adoption of an ontological point of view to gain a clear understanding of the stages by which we can arrive at a histologically consistent understanding of microscopic images acquired from tissue samples. As a side effect of this, however, we have proposed a number of categories which we feel could make a valuable contribution to ongoing work in the development of such ontologies. In particular, we highlight here our identification of a clearly-defined sequence of ontological levels to which the various entities handled during the imaging and analysis process can be assigned, as well as the proposal to analyse model histological elements as image segments (level 2) annotated with labels referring to classes of biological entity (level 0).

There are many potential benefits from using ontologies in histological imaging. Amongst other benefits, ontologies can

- help to establish a common understanding of histological images, by making annotation of complex histological components more efficient and interactive;
- facilitate the integration of histological images from various sources, thereby enabling reuse and sharing of knowledge;
- encourage communication and cooperation between histologists;
- enable histological imaging retrieval as well as the query process;
- assist in the development of informatic tools to improve automated diagnosis reasoning systems to support clinical decisions;
- demarcate histological imaging from general biomedical imaging;
- improve histological image analysis as well as image classification;
- establish histological knowledge-based validation and verification systems.

As emphasised in [17], the full extent of these benefits cannot be realised unless the ontology is integrated with a widely-recognised suite of ontologies such as those of the OBO foundry; for this reason we have, where possible, indicated which categories within our system of levels can be identified with categories taken from ontologies within that

suite. Further development of the ideas presented in this paper should be pursued in conjunction with a more thoroughgoing integration with existing established ontologies.

Acknowledgments

This research is supported by EPSRC through funding under grant EP/M023869/1 "Novel context-based segmentation algorithms for intelligent microscopy".

References

[1] R. Abu Eid and G. Landini. Quantification of the global and local complexity of the epithelial–connective tissue interface of normal, dysplastic, and neoplastic oral mucosae using digital imaging. *Pathology Research and Practice*, 199(7):475–482, 2003.

[2] R. Abu Eid and G. Landini. The morphometry of pseudoepitheliomatous hyperplasia: An objective comparison to normal and dysplastic oral mucosae. *Analytical and Quantitative Cytology and Histology*, 27(4):232–240, 2005.

[3] R. Abu Eid, F. Sawair, G. Landini, and T. Saku. Age and the architecture of oral mucosa. *Age*, 34(3):651–658, 2012.

[4] R. Arp, B. Smith, and A. D. Spear. *Building Ontologies with Basic Formal Ontology*. The MIT Press, Cambridge MA and London UK, 2015.

[5] P. H. Bartels and R. Montironi. Quantitative histopathology: The evolution of a scientific field. *Analytical and Quantitative Cytology and Histology*, 30(1):1–4, 2009.

[6] K. W. Eliceiri, M. R. Berthold, I. G. Goldberg, L. Ibáñez, B. S. Manjunath, M. Martone, R. F. Murphy, H. Peng, A. L. Plant, B. Roysam, N. Stuurman, J. R. Swedlow, P. Tomancak, and A. E. Carpenter. Biological imaging software tools. *Nature Methods*, 9(7):697–706, 2012.

[7] R. Flight, G. Landini, I. Styles, R. Shelton, M. Milward, and P. Cooper P. Automated optimisation of cell segmentation parameters in phase contrast microscopy using discrete mereotopology. In T. Lambrou and X. Ye, editors, *Proceedings of the 19th Conference on Medical Image Understanding and Analysis, Lincoln UK, Jul 15-17, 2015*, pages 126–131. British Machine Vision Association, 2015.

[8] A. Galton. The mereotopology of discrete space. In C. Freksa and D. M. Mark, editors, *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Science*, pages 251–66, New York, 1999. Springer-Verlag. Proceedings of International Conference COSIT'99.

[9] A. Galton. Discrete mereotopology. In C. Calosi and P. Graziani, editors, *Mereology and the Sciences: Parts and Wholes in the Contemporary Scientific Context*, pages 293–321. Springer, 2014.

[10] A. Galton. On generically dependent entities. *Applied Ontology*, 9(2):129–153, 2014.

[11] G. Landini and I. E. Othman. Architectural analysis of oral cancer, dysplastic and normal epithelia. *Cytometry A*, 61A:45–55, 2004.

[12] G. Landini, D. A. Randell, T. Breckon, and J. W. Han. Morphologic characterization of cell neighborhoods in neoplastic and preneoplastic epithelium. *Analytical and Quantitative Cytology & Histology*, 32:30–38, 2010.

[13] G. Landini and J. W. Rippin. How important is tumor shape? Quantification of the epithelial–connective tissue interface in oral lesions using local connected fractal dimension analysis. *The Journal of Pathology*, 179:210–217, 1996.

[14] D. A. Randell, G. Landini, and A. Galton. Discrete mereotopology for spatial reasoning in automated histological image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):568–581, 2013.

[15] J. Serra. *Image Analysis and Mathematical Morphology*, volume 1. Academic Press, 1982.

[16] B. Smith *et al.* The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25:1251–1255, 2007.

[17] B. Smith *et al.* Biomedical imaging ontologies: A survey and proposal for future work. *Journal of Pathology Informatics*, 6, 2015.

[18] L. Vincent. Morphological greyscale reconstruction in image analysis: Applications and efficient algorithms. *IEEE Transactions on Image Processing*, 2(2):176–201, 1993.