WHITE PAPER

The Annotation Exchange Format for Images (AXFI)

Overview

Image annotation and segmentation is an important analytic process in many scientific, technical, and commercial fields. Nonetheless, there are few standard formats for describing and representing image annotations, and those which do exist tend to be used in specific, relatively narrow contexts.[1] This is not a new observation; Daniel L. Rubin *et. al.*, in 2007, note that:

> Images contain implicit knowledge about anatomy and abnormal structure that is deduced by the viewer of the pixel data, but this knowledge is generally not recorded in a structured manner nor directly linked to the image. [Moreover,] the *terminology* and *syntax* for describing images and what they contain varies, with no widely-adopted standards, resulting in limited interoperability. The contents of medical images are most frequently described and stored in free-text in an unstructured manner, limiting the ability of computers to analyze and access this information. There are no standard terminologies specifically for describing medical image contents — the imaging observations, the anatomy, and the pathology. [N]o comprehensive standard appropriate to medical imaging has yet been developed. A final challenge for medical imaging is that the particular information one wants to describe and annotate in medical images depends on the *context* — different types of images can be obtained for different purposes, and the types of annotations that should be created (the "annotation requirements" for images) depends on that context. For example, in images of the abdomen of a cancer patient (the context is "cancer" and "abdominal region"), we would want annotations to describe the liver (an organ in the abdominal region), and if there is a cancer in the liver, then there should be a description of the margins of the cancer (the appearance of the cancer on the image). (http://cedarweb.vsp.ucar.edu/wiki/images/d/d9/R_19.pdf, pp. 1-2).

These challenges inspired **AIM** (the "Annotation and Image Markup" project), which "provides a solution to the ... imaging challenges [of]: No agreed upon syntax for annotation and markup; No agreed upon semantics to describe annotations; No standard format ... for annotations and markup" (https://wiki.nci.nih.gov/display/AIM/Annotation+and+Image+Markup+-+AIM). However, **AIM** itself has not been widely adopted outside the specific field of cancer research and cancer-oriented image repositories.

One obstacle to formalizing image-annotation data is that annotations have a kind of intermediate status, neither intrinsic parts of an image nor merely visual cues supporting the presentation of the image witin image-viewing software. Many applications exist which allow markup or comments to be introduced with respect to an image. From the application's point of view, these annotations are part of the application display, not part of the image — analogous to editing comments that might be added to a text document by a word processor or **PDF** viewer, which are records of user actions, not intrinsic to the document itself. Indeed, one mechanism for recording image annotations in **DICOM** (the "Digital Imaging and Communications in Medicine" format) is via "presentation state." The presentation state includes all details about how the image currently appears to **DICOM** workstation users, such as Radiologists, which could include markings they have made to indicate diagnostically

[1] Current formats include AIM (Annotation and Image Markup), CVAT XML (CVAT is the Computer Vision Annotation Tool), DICOM-SR (Digital Imaging and Communications in Medicine Structured Reporting), PASCAL VOC XML (Pattern Analysis, Statistical modelling and ComputAtional Learning Visual Object Classes), and COCO JSON (Common Objects in Context).

significant image regions or features. Insofar as image-annotations are considered to be artifacts of image-viewing software, rather than significant data structures in their own right, there is less motivation for imaging applications to support canonical annotation standards.

Nevertheless, in many scientific and technical areas image annotations *are* significant; they are intrinsic to the scientific value of a given image as an object of research or observation. Image regions, segments, and features have a semantic meaning outside the contexts of the applications that are used to view the corresponding images, which is why it is important to develop cross-application standards for describing and affixing data to image annotations.

It is also important for image-annotation models to be broadly applicable and multi-disciplinary. While image analysis serves different goals in different contexts (segmentation of microscope images to detect cancer cells serves different ends than segmentation of camera snapshots to study traffic patterns), there is always a possibility of analytic techniques developed in one subject area to be applicable for other image-processing problems, even if the practical outcomes desired of the analyses are very different. Furthermore, certain computational domains are similar enough to image analysis to warrant inclusion in a general-purpose image-annotation framework, even if the underlying data is not "images" in the conventional sense (not, for instance, captured via photographs or microscopy). For example, **PDF** document views, Flow Cytometry data plots, and geospatial maps subject to Geographic Information Systems (**GIS**) annotations may all be considered images — by virtue of a semantic significance attributed to color and to geometric primitives as a way of characterizing phenomena observed or modeled through their data — even though such resources are not acquired by ordinary "image-producing" devices. The "Pantheon" project, characterized as a "platform dedicated to knowledge engineering for the development of image processing applications,"[2] offers one of the few attempts in bioimaging literature to rigorously define "imaging" and "image processing" in the first place. The problem of defining "images" as such, and therefore delineating the scope of image annotation, is addressed below (section ...). In brief, **AXFI** considers the realm of imaging to be more general than just graphics obtained by a direct recording of the optics of some physical scene via cameras, microscopes, or telescopes. That is to say, the image acquisition process is not necessarily one where data is generated by an instrument which produces a digital artifact by absorbing light, so that geometric and chromatic properties of the image are wholly due to the functioning of the acquisition device.

Systematically identifying the scope of "image annotion" is important for **AXFI** because doing so clarifies the sorts of domains whose semantics could reasonably be incorporated into **AXFI**, as well as the sorts of applications which would be reasonable candidates for supporting **AXFI** annotations (i.e., the capability to parse **AXFI** data and represent it vis-à-vis the relevant images). For example, if immunoflourescent Flow Cytometry (**FCM**) data plots are classified as images, then the numerical properties of the "channel" axis, with notions of "decades" and a "log/linear" distinction, become relevant to the **AXFI** vocabulary for representing spatial dimensions and magnitudes. In general, **AXFI** uses paradigms and terminology from "Conceptual Space Theory" as part of the process of formalizing geometric and dimensional notions.[3]

When defining the scope of **AXFI**, it is also important to distinguish the *data models* encapsulated by **AXFI** resources from the file formats where **AXFI** data is encoded. The choice of one file type or another to represent a data structure — **XML** versus **JSON**, for instance — does not fundamentally affect the data thereby communicated. Therefore, it is important to formalize data models in such a way that numerous different serialization languages might actually be used to share/express the data. However, in practice, format-specific standards, such as **XML** Schema Definitions, are often used as the basis for formalizing and enforcing compliance with data models. Therefore, **AXFI** cannot be complete unconcerned with the structure and requirements of files which convey **AXFI** data. This

[2] See https://hal.archives-ouvertes.fr/hal-00260065/document.
[3] See http://idwebhost-202-147.ethz.ch/Publications/RefConferences/ICSC_2009_AdamsRaubal_Camera-FINAL.pdf or https://arxiv.org/pdf/1801.03929.pdf.

is particularly true because AXFI seeks to incorporate the data models of other specifications, such as AIM and GATING-ML, which are formalized via XML specifications. Although AXFI is not primarily XML-based, in short, it intends to be (in the relevant contexts) compatible with XML languages that rely on XML schematization. Further details on how AXFI manages the relation to XML and other serialization languages are provided below (section).

A further detail that should be clarified prior to expositing a formal outline of AXFI is that of how image-annotations originate. Sometimes, of course, annotations are manually introduced on images by human users of image-viewing software. On the other hand, automated image segmentation — or similar algorithmic or AI-driven image processing without human intervention — yields partitions of images into regions, or identification of semantically important locations in an image, therefore generating annotations computationally. In short, AXFI should support both human-generated and computer-generated annotations. This becomes complicated, however, insofar as image-processing may yield analyses which overlap with annotation objectives but may not intrinsically produce annotations in the conventional sense. For example, an algorithm to count the number of cars in a highway picture may rely on statistical analysis of some quantitative image feature — such as "zero-crossings" — without in fact producing determinate image segments.

It is important to remember that image processing operations do not always act directly on images themselves; sometimes algorithms are based instead on mathematical complexes derived from the image, but with their own quantitative properties. For instance, color-valued pixels may be replaced by matrices measuring the derivative of some image-feature field in eight directions around each point (an example would be "Sobel kernels" applied to the image intensity function). For a given "semantic" task — that is, an image-processing objective whose end-result is not just image-related data but some empirical observation — image segmentation, or other analyses yielding annotations, are a means to an end: one *way* to count cars is to delineate the edges of distinct cars in distinct segments, and then count the number of segments which result. However, statistical image-analysis may produce largely accurate results for such semantic tasks, given large image corpora (e.g., estimating traffic flows from highway cameras), without yielding artifacts such as human-visible segment representations. Or, in a different domain, AI-powered analysis of FCS (Flow Cytometry Standard) data could establish a largely accurate count of "events" (i.e., discrete FCS measurements of light-scattering and/or flourescent properties of cellular-scale entities) without manual "gating" (referring to the conventional practice of scientists using geometric annotations of FCS data-plots to isolate and thereby count different event-types). An AI-powered analysis of image features, or likewise of Flow Cytometry data, may yield calculations similar to those which for *human* users are achieved via image segmentation, manual gating, and similar operations which clearly yield annotation data. For AXFI, the complication arises when these AI mechanisms do not themselves yield results that would normally be considered annotations, but rather yield the desired empirical results for which the annotations would be a preliminary step — e.g., an approximate count of the number of cars in a highway photo, without a precise segmentation of the image marking the cars' respective borders.

AXFI ultimate considers these AI-related comlications to be issues resolvable at the level of software, rather than stadardized annotation models *per se*. An image annotation is, among other things, a visual (or viewable) record of some image-processing activity. If a radiologist manually clarifies a report to the effect that a given CT shows a tumor by circling the area where the tumor is visible, he or she is using the image annotation to communicate to others the thought-process which motivated the diagnostic conclusion. This is different than an AI-driven processor which would automatically demarcate an image segment outlining the tumor and use geometric properties of that segment to derive a pathological finding. In short, the data conveyed in an annotation — an image segment, rendered precisely, or rendered indirectly via a circle or polygon around the segment — may be *intrinsic* to an image-processing operation: it may be data acquired *at one stage* in an analytic workflow. However, annotations may also be *retroactive*; if a radiologist circles a tumor, he or she has completed (at least mentally) the image analysis, and is using the analysis to summarize what occured in the course of the analysis. Therefore, any image-processing task can be associated with

ex post facto annotations which summarize the process even if they are not intrinsic to it.

To continue the example of counting cars from a highway photo, an **AI**-powered observation could be retroactively *justified* by providing an image segmentation where the number of car-segments matches the **AI** count. In lieu of precise segments, it may be simpler to provide location-points for the "geometric center" of each car, or the points furthest apart in the direction of each car's front-to-back — these may be the statistical signals used for the car-counting process. Anlogously, facial recognition does not need to rely one segmenting out regions (eyes, nose, lips), but rather can be based on distances between individual points (such as the inner corners of each eye). But in any case, depending on the analytic algorithm used, it is often possible to identify some spatial/geometric feature or object that can be visualized in the image context, and that summarizes or legitimizes the analytic operation. This summarial data, then, can provide *retroactive* annotations which allow human viewers to understand and review the algorithmic process. In short, simply because image-processing tasks may not generate annotation data as part of their interal activity, it is still possible (and may be desirable) for the software operationalizing these tasks to implement annotation generators, where the resulting annotations document the operations for the scientific record and/or summarize them in **GUI** objects for the benefit of human viewers.

In general, then, **AXFI** distinguishes human-generated from computer-generated annotations, and moreover leaves open the possibility that some computer-generated annotations are *retrospective*: that instead of being internal to an imaging computation they are indirectly produced subsequent to such a computation, for purposes of documentation and validation.

In general, then, **AXFI** distinguishes human-generated from computer-generated annotations, and moreover leaves open the possibility that some computer-generated annotations are retroactive: that instead of being internal to an imaging computation they are indirectly produced subsequent to such a computation, for purposes of documentation and validation. Annotations which are not *retrospective* are called *internal*, as in, internal to a given image-processing workflow.

Related to the distinction between *internal* and *retrospective* annotations, **AXFI** recognizes a contrast between "intrinsic" and "descriptive" annotation. An example of an *intrinsic* annotation might be an image segment, where calculating the boundary of the segment is intrinsic to a specific image-processing objective, whereas an *descriptive* annotation might be an arrow *pointing toward* that segment. Here the descriptive annotation is introduced primarily for the benefit of human viewers.

The intrinsic/descriptive distinction is not always clear-cut. Consider the following two cases: in one scenario, an image-segmentation routine precisely delineates a region of interest (an outline of a red bird against blue sky, say), notating the segmentation result via a two-color-depth transform of the original image. Image-viewing software then shows the segment indirectly by encircling it, producing, in effect, a secondary annotation intended to call attention to the primary (segment) annotation. Here, clearly, the segment itself (encoding by a separate two-toned image) is intrinsic to the original analysis, whereas the secondary annotation has a purely descriptive purpose.

However, consider an alternate scenario where the original segmentation is done with less precision (this may be the case where there is a more muted color differential between foreground and background). Imperfect segmentation may still be adequate for some semantic task (e.g., identifying a bird's species). An analysis might obtain a rough segmentation by marking certain points highlighted by an edge-detector, then protruding the convex hull of the region outward so as to be sure of encompassing the whole bird-segment within a polygon, albeit allowing some background pixels into the polygon as well. This approximate segment is only an indirect representation of the region of interest (which would be the avian sub-image clearly outlined with no background included), but for analytic purposes the rough polygon might be a reasonable substitute for the finer-grained segment, analogous to how a cubic or quartic polynomial may be an adequate approximation to a more complex curve. Depending on how it is used, the approximate segment may therefore be considered merely a visual cue connoting the region of interest, or a significant region in its own

right. This contrast would, most likely, depend on whether the approximation is used itself as a basis for further analysis, or instead is mostly a presentation device. This usage-context would therefore indicate whether an "imprecise" annotation (AXFI formally uses the term *approximative*) should be classified as *intrinsic* or *descriptive*.

All told then, annotations may be *internal* or *retrospective*, and *intrinsic* or *descriptive*. AXFI also contrast *computer-generated* from *human-generated* annotations (which are related to but not the same terms as *manual* and *automated*; see below, section). These distinctions are indepenedent of one another; retrospective annotations for instance could be either intrinsic or descriptive, and either computer-generated or human-generated.

With these preliminaries concerning the scope of AXFI clarified, the following sections will (1) outline the kind of data encoded via AXFI and (2) the relation between AXFI, data types, and applications/libraries that may use AXFI. For purposes of exposition, the following specifications will discuss images mostly in two (spatial) dimensions; however, AXFI does not preclude annotations applied to image-sources with other dimensional profiles, such as 3D graphics, or movies (2D plus time), or even audio files (consistent with , which supports audio annotations because DICOM is sometimes used to share biomedical acoustical data, such as ultrasound).

Part I: Outline of the AXFI Data Model

Types of Image Annotations

The most fundamental kinds of image annotations are geometric primitives such as points, lines, and polygons. Many other forms of annotations are possible, however, mostly supplemental data which is associated with these primitives or, in some cases, with the image as a whole. In general, AXFI classifies annotations into the following groups:

Geometric Primitives These annotations delineate spatial/geometric regions in zero, one, or two dimensions (or potentially higher dimensions when working in non-2D contexts). At a minimum, AXFI data should support points, lines, polygons, polylines (considered a superkind of polygons where a polygon is a closed polyline), circles, and ellipses. Additionally, AXFI recognizes generic "closed" and "open" *curves*, which are nonlinear one-dimensional regions (or boundaries of two-dimensional regions) that are neither elliptical or circular arcs. Depending on context, AXFI can be extended to provide more detailed subkinds of curves generated by different sorts of mathematical equations, along with notations for the formulae (e.g., b-splines) that generate a particular curve.[4] More information about encoding ellipse data as well as other sorts of curves is outlined below (section).

A variation on the polygon/polygon-line alternative are polygon-lines demarcating spatial regions whose boundaries extend to one or more sides/corners of the image (e.g., a "quadrant" is defined via one central point with line segments parallal to and implictly extended to the edges). These poly-lines may not explicitly represent the overall image boundary as part of their own boundary. Conceptually, treating the image-border itself as a different *sort* of boundary than those explicitly notated may be more accurate than eliding these distinctions. This applies also to segment boundaries. For instance, the sand/ocean border in a beach scene represents a physical discontinuity between two different material substances, and so it records an observable detail of the depicted scene. On the other hand, the edge of the sand-region at the bottom of the image is not a physical boundary, but an artifact of the camery position; we assume the beach itself extends further than what the camera captures. All told, then, in addition to polygons (or curves or segment-boundaries) being *open* or *closed*, AXFI recognizes a third form of closure dubbed "incomplete," meaning that a

[4] In formal statements, this and other AXFI documentation will use the term "kind," as well as "superkind" and "subkind," to indicate groups of values/entities identified via a classification. Vocabulary based on "kind" is preferred to "type" or "class" because of the distinct meanings these latter terms have in computational contexts which are also discussed in reference to AXFI.

spatial region is geometrically closed by the edge of the image but that this closure has no observational or semantic significance. **AXFI** allows a notation that a spatial region is "closed by fiat" when the closure results from the intrusion of the global image boundary. A polygonal line may be closed by fiat when it does not explicitly include lines along the global boundary, but forms a closed polygon when such lines are included in practice; the result is called a *fiat polygon* (similarly an annotation can be classified as a *fiat curve* or *fiat segment*.

Segments and Regions A *segment* is considered to be, canonically, an integral subimage with a semantically precise separation of "foreground" from "background". There may be vagueness or approximation in how the segment is precisely individuated from its surroundings, but these ambiguities are considered to be practical limitations due to limited computer power, limiting pixel resolution, and so forth. A *region* or *region of interest* is similar to a segment, but defined more loosely; regions can have vague descriptors, and spatially disconneted parts of an image can be treated as part of the same region. In a photograph of a flock of birds, say, there may be multiple segments, each outlining one single bird. However, the flock as a whole may be outlined by one *region*, which could (without being deemed approximative) include some of the background sky. A *segment* can be seen as subkind of *region* with stricter granular and topological requirements.

A *partition* of an image is a segmentation which exhaustively classifies every point into one or another segment. A *selective* segmentation, unlike a complete partition, only isolates certain segments or regions of interests. **AXFI** adopts these terms as parameters that can characterize segmentation processes, and by extension the resulting segments/regions.

Locations In **AXFI**, *locations* are considered to be designations of areas within an image (of varying dimensions) which are significant by virtue of their directional, morphological, or topological relations to the rest of the image, rather than by virtue of their intrinsic shape. Conceptually, a *location* is in many cases similar to a *point*, but **AXFI** does not require locations to be zero-dimensional. A location may be designated by a small disk, or even a region/segment. The distinguishing feature of locations is that their spatial shape or extent are not semantically significant; instead, what is important about locations is their position in the image and how this position relates to surrounding image content. For instance, a location might be the point/position where two roads intersect, or it could be the leftmost point on the segment-boundary of a car's fender or a bird's wings, or the geometric center of a car's body or a bird's torso.

AXFI distinguishes location-annotations from secondary annotations used to identify locations (insofar as these may be visually distinct). For instance, a position may be modeled as a single point, but visually conveyed by an arrow, or by a circle hovering above the relevant point.

Focal Points The concept of *focal points* integrates locations and geometric primitives for certain analytic tasks. In general, focal points are important for positional rather than morphological regions, similar to locations. However, focal points may function more like segments or geometric objects when used as part of an analytic objective. For instance, points embodying the center of a car's body or a bird's torso could be used to count the number of birds or cars appearing in a photograph. In this case the concept of "geometric center" has no meaningful morphological properties, so it is analogous to a location. On the other hand, the center may be treated as a proxy for, or a most significant component of, a segment or region; in this sense the point is *part of* a segment. This mereological aspect makes focal points act conceptually more like geometric primitives than like locations. In short, the classification *focal point* is available for spatial objects which behave somewhere between locations and regions/points, and particularly when they are used in some proxying or indicative relation to other regions (e.g. for counting).

[Secondary Images] In some contexts, such as segmentation, image-transforms are used to convey image-processing operations, in contrast to geometric-style annotations that can be defined via vertex coordinates. A crisp segment can be defined via a two-toned image with the same dimensions as the annotated image, but with only two logical colors (colors are called "logical" insofar as the relevant detail is how the colors compare to one another, irregardless of the optical colors used to

render them[5]). A "fuzzy" segment can likewise be defined via a greyscale image (anything which is pure-background becoming either pure white, or pure transparent, depending on context). Secondary images can be used to denote annotations which are too granular to be summarized by any mathematical expression. In these cases, the actual annotation data should identify the secondary image (e.g., via a file path) and explain how it relates to the primary (or "ground") image, wheres the secondary file itself fills in the annotation details.

Secondary images may be derived from ground images merely to present annotations, but they may also be intermediate analysands which are themselves annotated. In the latter case, annotations on secondary images are usually meaningful also as annotations on ground images, so the interrelations between both images should be notated in the annotation data.

Proscriptive Annotations AXFI allows for the designation of certain annotations as "proscriptive" when they do not formally delineate a geometric object, but indirectly express such an object's shape or how it may be derived. A canonical example is the use of color — perhaps color-enhanced secondary images — to designate segments or regions of interest. For instance, one common segmentation technique uses color simplification; smoothing out color patches can facilitate image partitioning by reinforcing the boundaries between different regions. With sufficient color manipulation, image-segments can potentially be described chromatically; a region may for instance be identified as "the area all of whose pixels display a red channel above" some threshold. AXFI data should therefore allow such indirect designations of segments (and spatial/geometric regions in general) to be encoded as annotations.

The concept of proscriptive annotations is not only chromatic; one can imagine other scenarios where morphological features, such as symmetries, may be employed to similar effect. In describing a chess board, for instance, a set of image-regions (each square on the board) can be derived via translational transforms of an initial two-segment kernel (one black square and one white). This representation is possible because of translational symmetries in the underlying image. Such geometric transforms and symmetries can sometimes be used to "generate" meaningful image segments, so they should be notated in AXFI data when appropriate.

Semantic Labels Some annotations supply data about other annotations (what AIM calls "Annotation on Annotation" as opposed to "Annotation on Image"). In general, we can supply a label, description, or semantic classification indicating what an image segment or region is "about", i.e., what it "depicts" (a bird, a flock of birds, a car, a traffic jam, and so forth). Such meta-annotation may be seen as a semantic postlude to an imaging process, but it may also be seen as providing further detail about the process itself. For instance, suppose segmentation isolates a bird from the sky: the epilogue to this operation is an ability (for a human user or a computer algorithm) to classify the segment as "bird." However, we can also say that the given fact of the segment capturing the optics of a bird (with its given anatomical outline and coloration) is what allowed the segmentation to be possible in the first place. Therefore, the label "bird" serves both to utilize the segmentation for further analytic/classificatory purposes and also, potentially, to characterize the inner workings of effective feasibility of achieving the desired processing objective. It should be kept in mind, in short, that label annotations are not exclusively intended to be used for practical labeling in the contexts of tasks such as subject-marking images in corpora; labeling could also be used to notate how semantic properties of the image make segmentation (and other processing objectives) more or less feasible, or computationally intensive/error-prone.

The semantics of labels themselves is outside the scope of AXFI. AXFI makes no effort to proscribe what sorts of terms are useful as labels ("bird," "car," "ocean," "tumor," etc.), or whether labels are simple strings/words or have internal structure of their own. AXFI does, however, make the following minimal stipulations about labels. First, AXFI distinguishes "mass," "count," and "plural" label targets — e.g., *ocean*, *one bird*, *two birds*. These distinctions are directly relevant

[5] In QT, for instance, the **QColorConstants::Color0** and **QColorConstants::Color1** color values are considered to be special "colors" (i.e., **QColor** instances) which define the foreground and background of a two-toned image; they are not formally assigned to a fixed visual color, like black or white.

to how segments are bounded and counted; for instance, labeling a body of water as *a lake* (e.g. on a satellite image where the lake's full shore is visible) versus *ocean* (continuing to the horizon) implies different desiderata for how the labeled region spatially extends in relation to the surrounding image. Secondly, with respect to semantic labels, **AXFI** recommends that applications and libraries enable labels to be typed values — instances of application-specific data types — as well as simple character strings (like "bird.") This is consistent with **AXFI**'s general approach to application integration and type systems, to be discussed further in Part II.

Data Transformations This category

Image Metadata This category of annotation concerns any information "about" an image which is apart from the specific image content — including computational/encoding details such as color depth, pixel dimensions, and image file format, as well as acquisiton details such as the make and settings of the camera whose photograph yields the current image.