|  |  |  |
| --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION**TELECOMMUNICATIONSTANDARDIZATION SECTOR**STUDY PERIOD 2017-2020 | FG-AI4H-I-047 |
| **ITU-T Focus Group on AI for Health** |
| **Original: English** |
| **WG(s):** | DASH | E-meeting, 7-8 May 2020 |
| **DOCUMENT** |
| **Source:** | Editors |
| **Title:** | DEL05.x: Annotation tool - Draft requirements study |
| **Purpose:** | Discussion |
| **Contact:** | Henry HoffmannAda Health GmbHGermany | Tel: +49 177 6612889Email: henry.hoffmann@ada.com |
| **Contact:** | Marc LecoultreML|LAB.AISwitzerland | Email: ml@mllab.ai |

|  |  |
| --- | --- |
| **Abstract:** | This document contains an initial draft of the deliverable for the annotation tool. It is based on the previous work carried out by TG-Symptoms and continues the discussions started at the H meeting in Brazil.  |

Table of Contents

[1 Introduction 3](#_Toc39736915)

[1.1 Benchmarking Test Data 3](#_Toc39736916)

[1.2 Challenges with Test Data 3](#_Toc39736917)

[1.3 Annotation Tools 3](#_Toc39736918)

[1.4 Challenges with existing Annotation-Tools 4](#_Toc39736919)

[1.5 A Joint Annotation Tool 4](#_Toc39736920)

[2 Analysis of Existing Annotation Tools 4](#_Toc39736921)

[3 Requirement 4](#_Toc39736922)

[3.1 Topic Group Annotation Requirements Survey 4](#_Toc39736923)

[3.2 High Level Requirements 5](#_Toc39736924)

[3.2.1 End to End annotation process 5](#_Toc39736925)

[3.2.2 The Global Scope 5](#_Toc39736926)

[3.2.3 Meta-data & Bias 5](#_Toc39736927)

[3.2.4 Quality 5](#_Toc39736928)

[3.2.5 Safety and Security 5](#_Toc39736929)

[3.2.6 Transparency 6](#_Toc39736930)

[3.2.7 Compliance by Design 6](#_Toc39736931)

[3.2.8 Interoperability / Data Format / Encoding 6](#_Toc39736932)

[3.2.9 Baseline 6](#_Toc39736933)

[3.2.10 Modality Support 7](#_Toc39736934)

[4 Implementation Options 7](#_Toc39736935)

# Introduction

The AI4H focus group is organized in working groups and topic groups. Working groups are taking care of horizontal aspects like regulatory considerations, ethics or the handling of data and AIs. The vertical topic groups are working on creating the actual benchmarking for a “topic” that might be a set of diseases, a medical specialty or a modality like X-ray. Common for all topic groups is that they aim for establishing a standardized benchmarking for their topic. This always includes a test dataset labelled/annotated with the output an AI is expected to respond with if given a test set entry.

## Benchmarking Test Data

The test data contains the inputs given to the AI and the expected outputs. Depending on the task the expected output is as simple as a single number or a class. For other tasks it also contains aspects as acceptable outputs, or particular wrong outputs that are e.g. clearly ruled out by the input and therefore must be avoided.

Depending on the topic, the datasets consist of real patient data for instance from a hospital, synthetic data generated for testing certain aspects of the AIs like robustness to noise, multi-morbidity or the response to meaningless nonsense. For some topics the test data also needs to contain test cases defined by experts for instance to cover the most typical presentation of a disease or the known corner cases.

## Challenges with Test Data

In principle, real patient-data is available in a multitude of databases. However, the encoding of this data is quite heterogeneous. Data formats, media type, level of detail, encoding, completeness or even smaller details like if a finding was not found is explicitly given or just omitted differ widely.

Independent of the source, the data entries also need to be annotated with the expected AI output which in machine learning is often referred to as the “label”. Even here, aspects like the quality, encoding, completeness vary. To determine which AI output would be correct, acceptable or wrong is not trivial. Studies show that doctors often do not agree on a correct label (inter-observer variability). For some tasks it was shown that doctors do not even agree with their own label if asked a few hours later (intra-observer variability). There are also systematic differences in how doctors decide on certain aspects and therefore label cases from a different perspective.

The process of annotating datasets is complicated by the fact that e.g. in a diagnosis task the correct disease confirmed a week later by imaging is not necessarily the correct AI output based on less information. For instance, it might not be correct to expect an AI to respond with “brain cancer” if given only “headache” as a symptom even if a later MRI unknown to the AI would confirm this. Especially the opinions of doctors about the “correct answer” if not all information is available differs since there is no easily obtainable ground truth.

Another challenge with every test data from a real-world context is that it is always biased in some way. For the interpretation of the results this bias needs to be known.

## Annotation Tools

To address some of these aspects, test data creation and annotation tools have been developed. They help to make sure that test data is encoded and annotated in a unified way. They also sometimes implement a dedicated workflow for the peer review of test data and labels. For some tasks a simple spreadsheet with some macros or predefined drop-down fields – possibly accompanied with an annotation guideline – is sufficient. Other tasks require functions for annotating more complex AI outputs like for instance an 3D surface around a tumour in a volumetric data set. These tools then also enforce constraints specific for a modality like that the surface in the volumetric examples is closed, free of self-intersection and not empty. Some tools also help with the task itself e.g. by pre-selecting things that then only need to be refined.

Many of the features are shared by the existing tools. This includes the handling of meta data including the source, the method how it was acquired and when, devices used and people involved in the process. Annotation tools using similar input data also have similar features like zoom, pan, rotate, contrast, brightness, drawing of outline, boxes, circles etc. for 2D image-based tasks.

Even though the tools have many features in common, the work in the focus group also showed that many AI tasks require some annotation details so specific that they are not needed by other AI tasks.

## Challenges with existing Annotation-Tools

The design goal of most annotation tools is to support the creation of data for the training and testing of AI systems by their developers. The tools are therefore often highly specialized for one AI task, require a steep learning curve to be used and are usually not designed for scalability or flexibility. Other companies using the tools primarily for the curation of training data in a domain where high quality data is scarce developed more sophisticated tools since they consider them a key part of their IP and competitive advantage, however these tools are not publicly available.

As primarily company-internal tools the existing annotation tools have virtually no protection against any attacks and attempts to compromise the benchmarking since for the AI developers. Out-of-the-box the tools are therefore not always an option for annotating the benchmarking data needed inside the focus group.

## A Joint Annotation Tool

All topic groups inside the focus group work on establishing a standardized benchmarking for their topic and need annotated benchmarking test data. Most of them will require some kind of annotation tool. One example is the topic group on AI-based symptom assessment (TG-Symptom) that started working on a tool for creating and annotating case vignettes for their benchmarking. Another example of an annotation tool presented to the focus group is Insight GT a web-based ground truth annotation tool for medical 2D images developed by Hulilabs.

Given the complexity of building and maintaining an annotation tool secure and scalable enough for the worldwide collection for representative test data, the focus group started during meeting G to discuss the potential of a modular annotation tool that could be shared by the whole focus group.

It was also pointed out that for some tasks there might be the potential to collect the data for several topic groups at the same time. For instance, in TB patients it could make sense to collect the X-Ray, symptoms and coughing audio signals at the same time rather than collecting different data sets repeatedly from the same patients.

During meeting H the first thoughts on the possibility of a joint tool where outlined in document FG-AI4H-H-038-R01.

# Analysis of Existing Annotation Tools

This section gives an overview of existing annotation tools and their features. This allows both, the identification of features that are similar across the different tools or certain modalities, and also the identification of features that are specific.

TODO review the list of tools collected

# Requirement

## Topic Group Annotation Requirements Survey

In addition to the features identified by analysis of the available annotation tools, a survey was designed for collecting the topic group’s requirements for annotating benchmarking data. The survey was extended by several questions focusing on the annotations tool requirements. The insights from this survey will be summarized in this section. The questionnaire can be found here: <https://forms.gle/51iuHG5SrP6E8Hfr7>

TODO

## High Level Requirements

This section summarizes the requirements for an annotation tool that could be shared by the different topic groups inside the focus group.

### End to End annotation process

The general goal is to facilitate the complete process of the continuous creation of the datasets needed for benchmarking the AI systems in the different topic groups. It needs to cover all steps from the planning and organization of the creation of a dataset, the execution and monitoring of the collection including the review and conflict resolution process to the final handover of the dataset to the system executing the actual benchmarking.

### The Global Scope

The AI4H focus group can only be successful if the benchmarking results allow stakeholder all over the world to make decisions on which AI systems are a viable option in a given context. This requires the benchmarking data to be representative on a global scale and to contain cases from every country, every region, every ethnicity, age, sex and all other relevant dimensions. The annotation tool needs to support the worldwide, parallel collection and annotation of benchmarking data. It is a platform connecting hundreds of clinics, doctors and fields experts, the topic groups that want to perform a next benchmark run and the entities then executing it.

### Meta-data & Bias

Even if the annotation tool allows to collect the test data in the most representative way, the reality is that every data set from every region is biased in some way and that even combining all of them will not completely remove the bias. More important – every single context of use has a bias for many dimensions e.g. the local populations ethnicity, age distribution or social situation. Rather than showing leader boards the benchmarking systems allow to drill down to the performance estimates of the different solutions in a given – heavily biased context. An example might be a decision on some screening system for pregnant women in southern Africa. To obtain the benchmarking data for this context the benchmarking results need be filterable for the different parameter dimensions specifying this context.

One of the core features of the annotation tool is to provide the means for the definition of such meta-data dimensions, their scales/states and to then allow to specify them for each data point.

### Quality

One of the biggest challenges if it comes to benchmarking AI systems is to design tools and processes for guaranteeing the necessary high quality of the data. All benchmarking results are only as good as the test data and since stakeholders might base their decisions on this data, it must be free of any doubts if it comes to the quality. Given that individual doctors are not good enough for creating and annotating cases, the annotation tool therefore needs to support a sophisticated peer reviewing process. Since the number of reviewers and the mode of conflict resolution vary across the different AI tasks, the tool needs to be flexible in this respect.

### Safety and Security

Beside quality the most important requirement for any annotation tool is trustworthiness. If there are any doubts that the results have been manipulated to favour any AI manufacturer or that a participant in the benchmarking could have had access to the benchmarking data before the benchmarking the results are of no value for decision makers. The benchmarking tool needs to protect itself and the data against any unauthorized access. This includes a detailed rights and role management system, encryption, the separation of case data and labels, a complete version history and a design that makes it impossible to manipulate data or history without leaving traces. It also includes the mitigation of any risks connected to the software and the systems running it.

The creation and annotation of benchmarking data sets large enough for representative benchmarking results involves a considerable amount of manual work. The annotation system needs to make sure that the data collected is protected against any kind of data loss in case of technical failure and/or attacks.

### Transparency

Another dimension of trustworthiness is transparency. Given that potentially many topic groups, the participants in their corresponding benchmarks and the decision makers need to rely on the software, it is crucial that everyone has access to the source code to check for bugs and to make sure that it was not manipulated.

### Compliance by Design

For generating representative benchmarking results it is often necessary to also use real case data from real patients. This kind of data is protected by GDPR and similar regulation protecting personal data in general and health related data in particular. The annotation tool needs to deal with this data in a compliant way. It is important to recognise that some AI tasks require data that cannot be synthesized or anonymized effectively and that therefore documenting consent, data donation or a legitimate public interest need to be considered and included in the software design.

### Interoperability / Data Format / Encoding

Most developers of AI systems have some approach for creating training and testing data. In this case it is only important that the data is similar to the data used in the context where the AI is intended to be used i.e. using the same encoding and format or having an effective mapping technology in place. The benchmarking pursued by the focus group in contrast focuses on testing all available systems for a given task in a way where the results can be compared. All AI systems need to be able to understand the input data even if they are designed for slightly different input encoding. For this the topic groups need to agree on an ontology. As in the example of TG-Symptom this is for instance most likely a SnomedCT subset selectively extended by missing concepts. Other topic groups might be able to use existing ontologies without any changes. The annotation tool should support annotating data based on an explicitly defined ontology, preferably using mainstream standard ontology everyone can agree on. This will also allow the easier integration of existing clinical data.

### Baseline

The main purpose of the benchmarking is to provide the decision makers with the means to decide if an AI is a viable solution in a given context. Understanding if a solution is “viable” also requires assessing how it relates to the performance of possible alternatives. In most contexts this is defined by the performance of the available health workers. For several diverse reasons their performance is not available in most contexts so that it is difficult to compare it to AI systems.

The annotation systems should therefore also allow to measure the performance of health workers on the benchmarking data using the same global network of clinics used to create and peer review the cases to collect the baseline. As the benchmarking data itself it needs to be assessed locally in the clinics in the different regions so that representative comparison for a given context is possible.

### Modality Support

Important parts for the annotation and especially the annotation of the metadata is shared by many AI tasks. Beside this the tool needs to support the annotation tasks specific for the modalities. The annotation tool therefore needs to have a modular architecture that allows to easily plugin components for annotating data for the relevant modalities. While the base functionality of these components might be shared between e.g. several topic groups dealing with 2D image classification tasks, the topic groups most likely need to add further details only relevant for them. The collection of such modality annotator plugins might include for instance:

● 2D image area segmentation + labeling

● 2D image boundary segmentation + labeling

● 2D image entity counting + labeling

● 3D image area segmentation + labeling

● 3D image boundary segmentation + labeling

● 3D image entity counting + labeling

● 2D-series image area segmentation + labeling

● 2D-series image boundary segmentation + labeling

● 2D-series image entity counting + labeling

● 3D-series image area segmentation + labeling

● 3D-series image boundary segmentation + labeling

● 3D-series image entity counting + labeling

● Audio data segmentation + labeling

● Low frequency N-channel time series labelling (ECG, EEG)

● Semantic case vignette editors enforcing ontologies for symptoms, conditions, and factors

All the annotator plugins would support all the necessary tools for navigation (zoom, pan, scale, etc.) and enhancement (brightness, color, contrast, amplitude, etc.) that allow for user friendly, efficient and correct annotation. They might also include semi-automatic tools e.g. providing editable automatic pre-annotations.

# Implementation Options

The implementation of a focus group overarching annotation tools is a complex nontrivial software development task. Under normal circumstances this would require a company with a strict software development process following e.g. ISO quality standards. However, given the limited resources that individual contributors can invest into the implementation and given that no one outside a topic group really understands the specific details of a topic group the development can only be joint effort coordinated between the topic groups. Since the task is complex it might be necessary to have a “driver” similar to this role inside the topic groups. For this case Hulilabs who has already worked in annotation tools in the past offered to take this role.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_