DocTAG: A Customizable Annotation Tool for Ground Truth Creation

Fabio Giachelle, Ornella Irrera, Gianmaria Silvello
fabio.giachelle@unipd.it ornella.irrera@unipd.it gianmaria.silvello@unipd.it

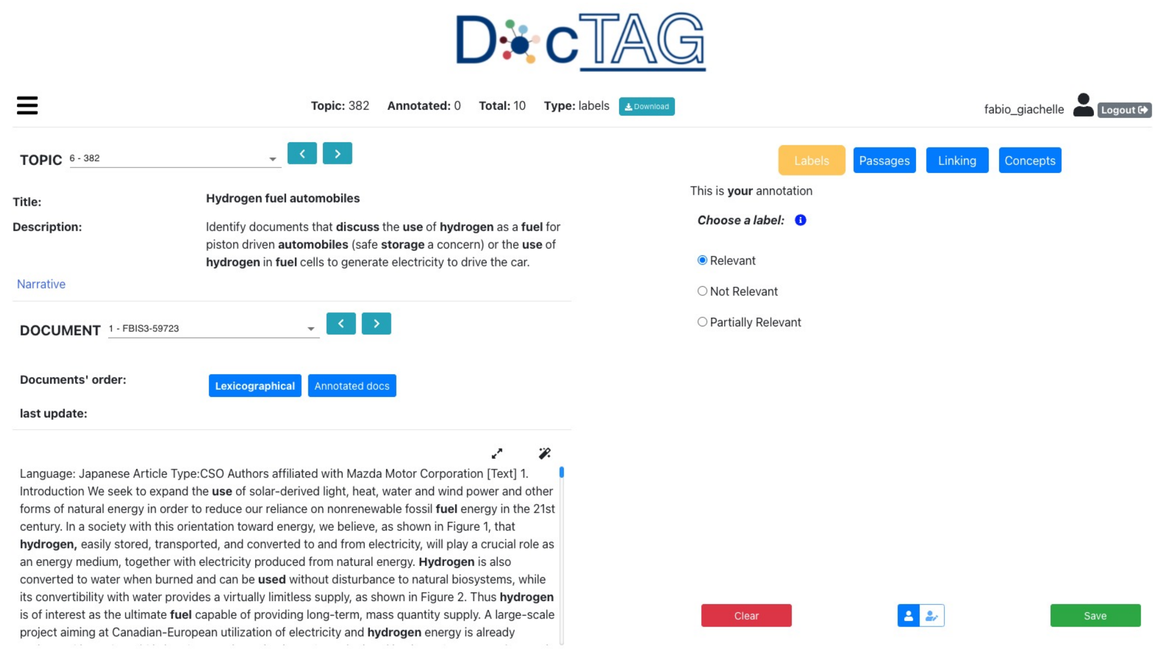
The Problem

Information Retrieval (IR) is a discipline deeply rooted on **evaluation** that in many cases **relies on annotated data as ground truth**.

Manual annotation is a **demanding and time-consuming** task, involving human intervention for topic-document assessment. To ease and possibly speed up the work of the assessors, it is desirable to have easy-to-use, collaborative and flexible **annotation tools**.

Despite their importance, in the IR domain no open-source fully customizable annotation tool has been proposed for topic-document annotation and assessment, so far.

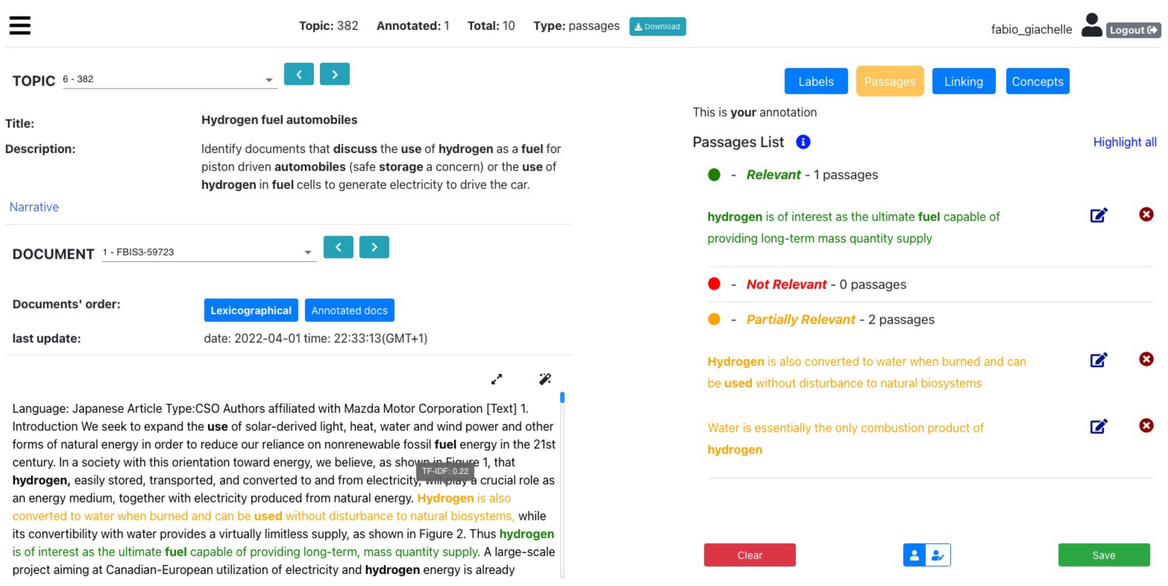
DocTAG

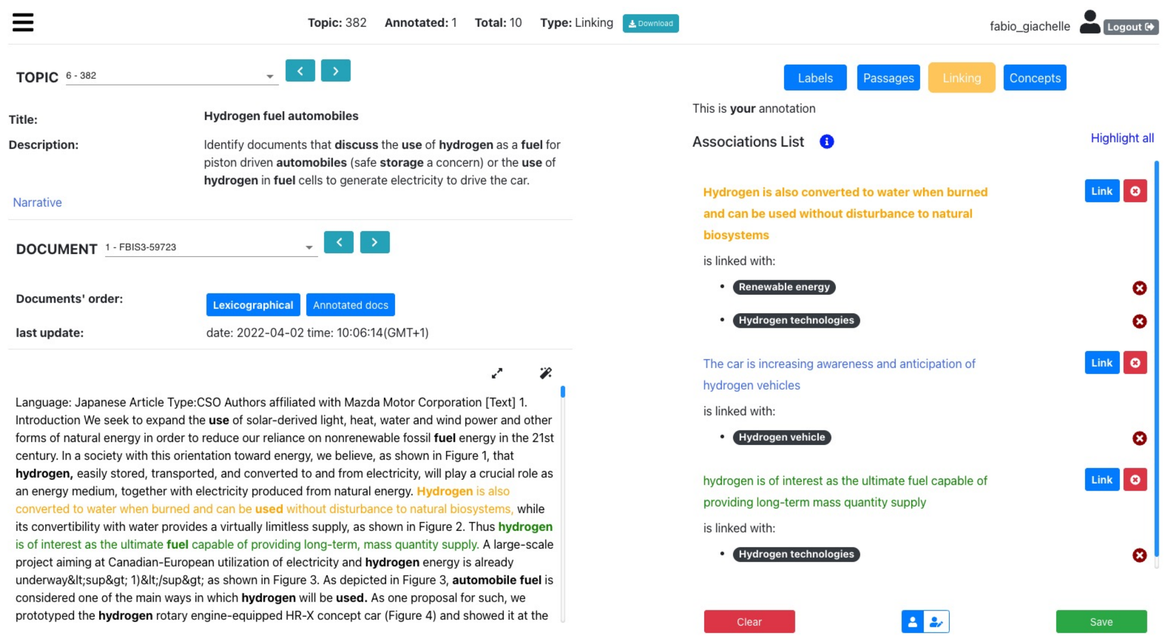


We propose DocTAG, a **customizable** web-based **annotation tool** for **ground-truth creation**. DocTAG is an **open-source** tool specifically designed to support human annotators in the IR domain. DocTAG enables the users to assess the **topic-document relevance**, by providing four different annotation modes:

- **Labels**: each topic-document pair can be associated with a single label (e.g., *Relevant*, *Not Relevant* or *Partially Relevant*).
- **Passages**: document passages can be marked with labels (one label per topic-passage pair) highlighted with different colors.
- **Linking**: each passage can be linked to user-defined/ontological concepts (one or many).
- **Concepts**: each document can be associated with several user-defined/ontological concepts.

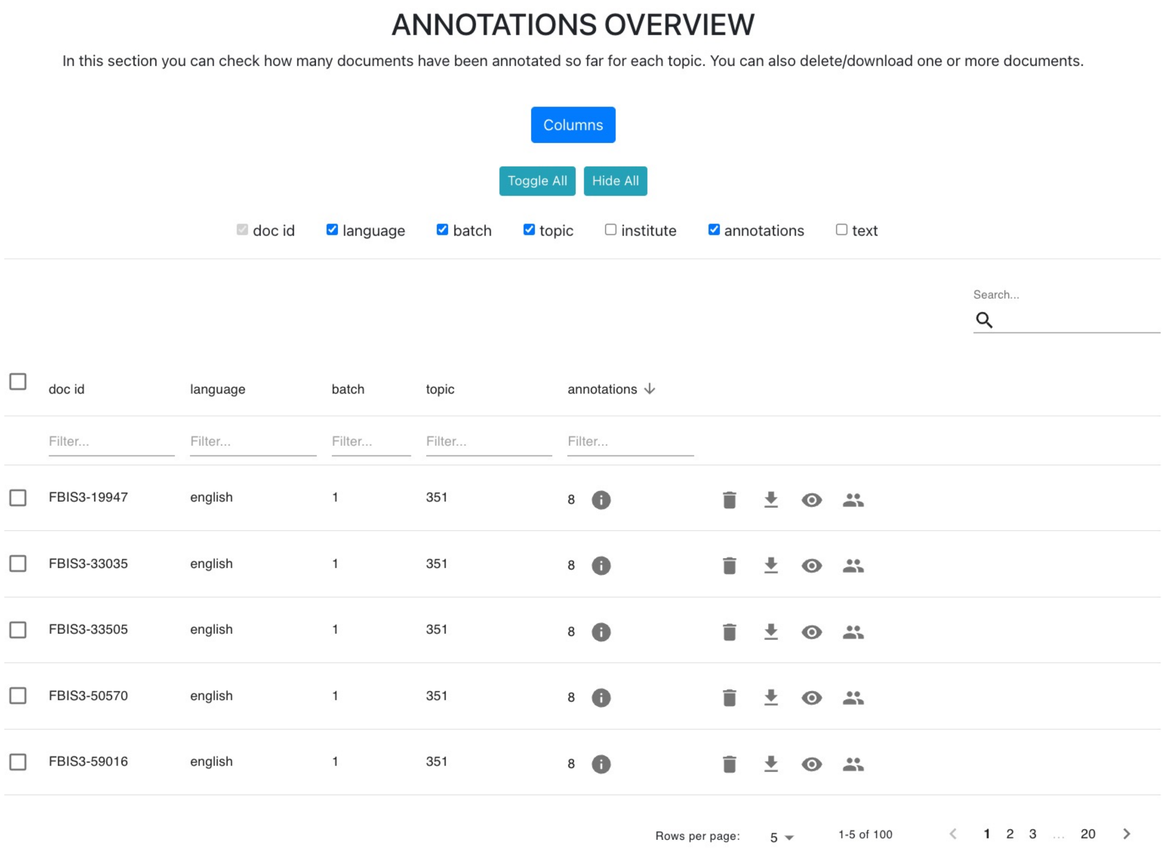
Labels and concepts are **document-level**, whereas passages and linking are **passage-level** annotation modes.



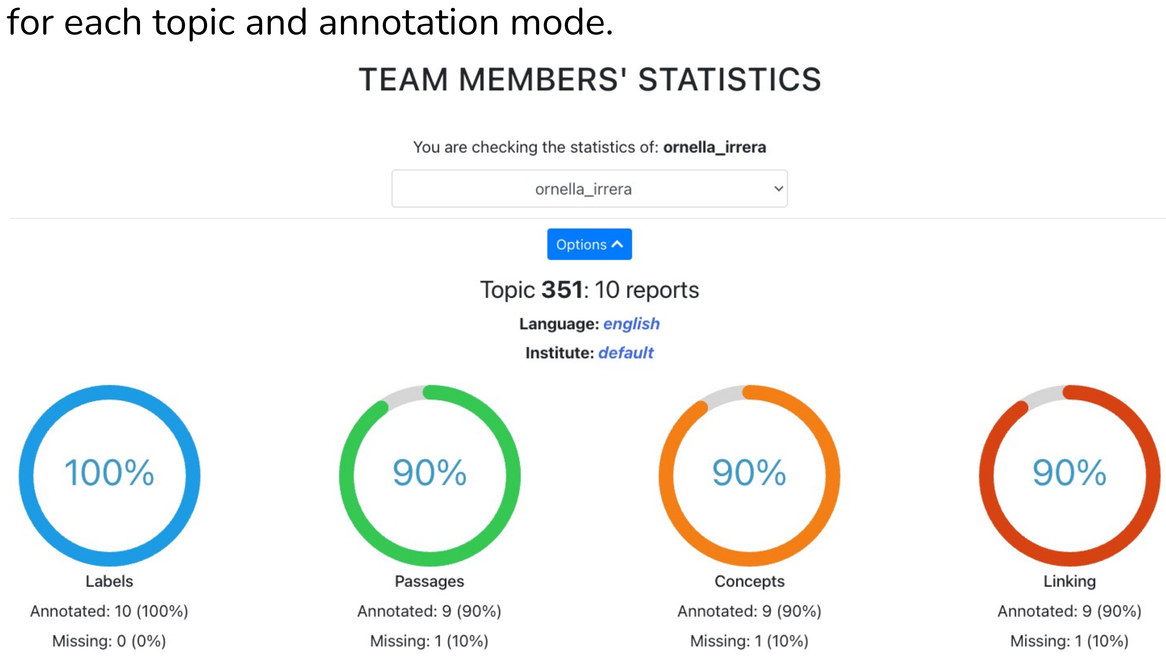


Annotation process control and statistics

Checking the annotation process advancements and coordinating the annotators could be burdensome without the aid of automatic solutions. To this aim, DocTAG provides a **collaborative setting** for monitoring the annotation work in terms of number of annotated documents for each topic and the authors of the annotations. DocTAG provides an overview of the annotation work carried out, indicating for each document the number of annotations done. DocTAG provides also **inter-annotator agreement facilities**, such as the automatic generation of the **majority vote ground-truth** among a set of given annotators.



DocTAG allows the users to check other team members' annotations and copy them in their own profiles for further edits.
In addition, DocTAG provides **annotation statistics** for each team member, such as the number/percentage of annotated documents for each topic and annotation mode.



https://www.unipd.it https://www.examode.eu http://iiia.dei.unipd.it https://ecir2022.org