FAQS for OCLC Data Labeling Project

The data labeling tool is now closed. Thank you for your support and participation in this initiative.

Why is OCLC engaging member libraries to help identify duplicates in WorldCat?

Cleaning up duplicate records in WorldCat is one of the most impactful measures we can take to improve the quality of WorldCat and improve the experience for libraries and their users—it is a key component of our cooperative effort to maintain accurate and useful data. While OCLC experts and members of the library community are best suited to do this work, we know that scaling that work to the entirety of WorldCat using humans is simply not a feasible or sustainable solution. So, we’ve developed a machine learning model to identify duplicate records in WorldCat and subsequently take the necessary actions to merge those records. That’s where you come in. We need humans to validate and enhance our model’s understanding of duplicate records to scale the work, which ultimately improves the quality of WorldCat for the entire cooperative and library community.

What are duplicate records?

Duplicates records are records that are considered “functionally equivalent” for an item. True duplicates make library processes less efficient and impede discovery. But records that look equivalent at first glance may actually represent items that differ in important ways.

How does machine learning work?

The labeled data that is collected from this project will be used to confirm and enhance our machine learning model’s understanding of duplicate records in WorldCat. All machine learning models require a lot of “right” answers at the beginning. The data we’re collecting in this project helps ensure the model has the right answers.

Some, but not all, of these right answers (called Training Data) are given to the model so it can learn what it should and should not be looking for. In this case, it will learn what a duplicate record looks like and what it doesn’t look like. After that, we take the remaining “right” answers (called Testing Data) and score how well the model does at getting the answer we want. If the results, or accuracy, reach a certain quality threshold, we can apply the model to data where we don’t already know the answer. Just like with humans, the more we train the model the better it does on the test and, eventually, real life.
What happens to records identified as duplicates?

Following this project, the machine learning model will be presented with pairs of potential duplicates from WorldCat. It will then analyze the records and predict the likelihood of the records describing the same item. Records identified as duplicates will be sent to the WorldCat quality team within OCLC where they will be merged according to OCLC rules and procedures.

Who can I contact with questions about the data labeling project?

You can contact us at datalabeling@oclc.org for questions or assistance.

What information about me or my library is gathered during this project?

The only user information gathered during this project is counts associated with user tokens. We do not keep usernames, login credentials, or any other personally identifiable information.

What is the Leaderboard?

The Leaderboard displays the top 25 users based on the highest number of responses submitted (potential duplicates labeled). Users can access the Leaderboard by clicking on the “Leaderboard” button in the upper right corner of the screen—the rankings are updated each time you click the “Leaderboard” button. You, the current user, are displayed on the top row and highlighted in purple. The number displayed after “User” is your ranking (e.g., “User 7” means that you are currently in 7th place). The number on bottom right shows the current total of all labeled data up to 60,000 records.