FAQS for OCLC WorldCat Duplicate Data Labeling

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 an AI machine learning model to identify duplicate records in WorldCat. That's where you come in. We need you 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.

How do I participate to help identify duplicates in WorldCat?

To participate in data labeling and validate our AI machine learning model's understanding of duplicate records in WorldCat, follow these participation instructions. At the end of March 2025, we will begin analysis of the collected data.

I don't know my WorldShare login credentials. Who can I contact for assistance?

If you don't know your WorldShare login credentials, the best course of action is to contact the WorldShare admin librarian at your institution, who will be able to look up your account information. If you are unable to contact that person, you may contact OCLC at orders@oclc.org. In the email, be sure to provide your institution's symbol, your name, and your email address.

What are duplicate records?

Duplicates records are records that are considered "functionally equivalent" manifestations—records that describe the same resource in the same language of cataloging. True duplicates make library processes less efficient and impede discovery. But records that look equivalent at first glance may actually represent manifestations that differ in important ways. For additional guidance on determining whether records are "functionally equivalent", please see Bibliographic Formats and Standards (BFAS) Chapter 4.

How does machine learning work?

The labeled data that is collected from this effort is used to continually enhance our AI 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’ve collected and continue to collect helps ensure the model has the right answers for all the different materials reflected in WorldCat.

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?

The AI 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 manifestation. 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?

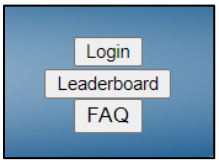
You can contact us at datalabeling@oclc.org for questions or assistance.

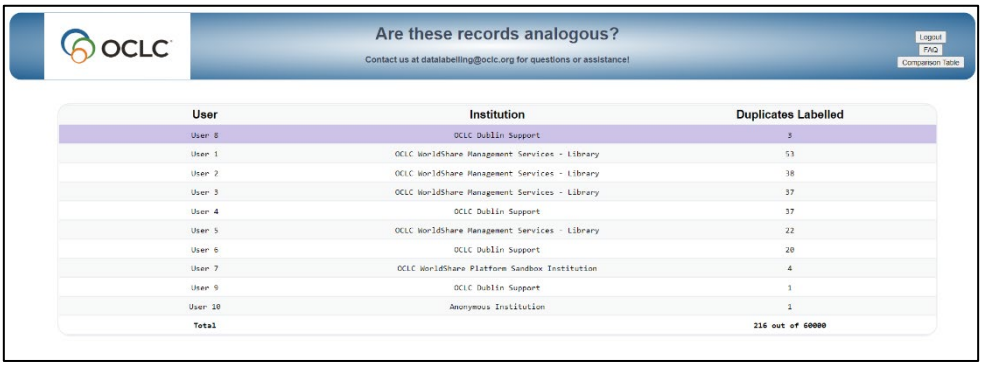
What information about me or my library is gathered during this project?

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.





User	Institution	Duplicates Labelled
User 8	OCLC Dublin Support	3
User 1	OCLC WorldShare Management Services - Library	53
User 2	OCLC WorldShare Management Services - Library	38
User 3	OCLC WorldShare Management Services - Library	37
User 4	OCLC Dublin Support	37
User 5	OCLC WorldShare Management Services - Library	22
User 6	OCLC Dublin Support	20
User 7	OCLC WorldShare Platform Sandbox Institution	4
User 9	OCLC Dublin Support	1
User 10	Anonymous Institution	1
Total		216 out of 60000

OCLC