Esploro Essentials: Smart Harvesting

Smart Harvesting in Action

How does Smart Harvesting work?

In this video, we will demonstrate how to run Smart Harvesting manually and explain what happens behind the scenes when you do.

Before you run Smart Harvesting, create a set of researchers.

Each researcher must have at least one of the following: an Affiliation, Research Topics, Areas of Interest, or some assets already approved in Esploro. The more information, the better, because this is the data that the algorithm will use to find author matches. The researcher must also have “Include in Smart Harvesting” enabled.

You can use Smart Harvesting to search all past records for a given researcher. This takes a lot of time and processing power. Therefore, there is a limit to the number of researchers that you can run in one batch.

We recommend that you first try to use Smart Expansion to add all a researcher’s publications to date. If Smart Expansion is not sufficient, and the researcher’s publication list is not complete, then we do recommend running retrospective Smart Harvesting to find other past publications. To do this, leave Last Smart Harvesting date blank.

If, on the other hand, Smart Expansion results in a complete publications list, then there is no need to run retrospective Smart Harvesting. You can change the last Smart Harvesting date to the current date and run Smart Harvesting on a regular basis from that date onward, to add new assets as they become available. Each time the job is run, it will search for records that are new since the last time the job was run for that researcher.

Scheduling Smart Harvesting jobs will be covered in another video

To run Smart Harvesting, open the Repository Menu. Under Smart Harvesting, select Manage Profiles. Open the CDI profile.

Scroll down to Author Matching Approval Configuration. You can choose the settings for each match rank. For example, you can select Automatic for matches ranked as Very Strong, and these author matches will be approved automatically. If you select Administrator, those matches will appear in the Author Matching Approval task list. If you choose Selected Researchers, and that researcher has been enabled to approve matches, then the match is sent to that researcher for approval.

You can choose different settings for each rank. Matched on ID means that ORCID or another identifier were found in the publication. The other ranks were matched by the Smart Harvesting algorithm with probability ranked accordingly.

Now, let’s look at the Asset Approval section. You may want to approve all assets that have been matched with at least one researcher at your institution, in which case, choose this middle option, which is the default. To manually approve all assets, choose Never automatically approve the asset. This allows you to check every asset before it becomes public. This third option allows you to automatically approve assets that meet certain conditions, either of asset type or if the asset has a DOI or PMID. Articles and conference proceedings generally have good metadata available in CDI, especially if they also have a DOI or PMID. Therefore, an asset with these conditions tends to be reliable.

Choose the set of researchers and press “Run Now.”

When you run Smart Harvesting, this is what happens behind the scenes.

1. A query is sent to CDI for every researcher included in the Smart Harvesting job, based on the researcher name and variants.
2. The records returned by CDI are first checked for an author ID, if available.
3. If not, the Author Matching algorithm is invoked. How likely is it that the publication was written by the researcher?  The matches are ranked as Very Strong, Strong, or Uncertain
4. For each match, the system goes on to check if the record is already present in your repository.  If it is – it is rejected as a duplicate.
5. If it is new, the system goes on to match co-authors, which are ranked in the same way. If no match is found, a new non-affiliated researcher is created, so that each author in the asset is matched to a researcher.

After the above is run for all researchers, the job has some additional stages (for example indexing the new assets).

Depending on your settings, some author matches may be automatically approved. Others will need approval by an administrator or the researcher. Your administrative staff can check the author matching using the Author Matching Approval Task List, which is explained in another video.

If you used the default setting for asset approval, then any record with an approved match to an affiliated researcher will be included in your repository. As we saw earlier, you can choose to manually approve some or all assets. In that case, asset approval is a separate process.

Smart Harvesting helps you automate, as much as possible, the process of adding research assets into your repository.