Ex Libris
A ProQuest Company

Smart Harvesting AI

Growing Your Research Information Hub

Leveraging machine learning to streamline scholarly communications and increase the visibility of institutions' research and expertise

Universities highlight their faculty's scholarly publications and research expertise to earn peer recognition and raise their academic ranking, which in turn draws more students, talented researchers, and funding. The institution's research information hub (referred to by some as a research repository) plays a critical role in supporting this effort, as a central space for the collection and management of scholarly outputs, data, and activities.

To be most valuable, the research information hub must be as current and complete as possible, as well as keep pace with rapidly evolving technology, workflows, and user expectations. In addition to making it possible to showcase the institution's research work and expertise, a robust and up-to-date hub is critical for effective administration and actionable analytics.

The Problem of Abundance

The volume of scholarly outputs is growing at an increasing pace, with publication models and output formats evolving rapidly. Research outputs are no longer just about published articles. The Covid-19 era led to an increase in, and emphasized the value of, published preprints. It is equally important to make other types of output, research results and information easily accessible, including the datasets upon which publications are built, creative works, patents, podcasts, videos, policy changes, and media mentions.

Such changes have posed several management challenges. With data and published research found in a variety of sources, each with their own metadata and structure, rounding up all the relevant outputs for the repository is a heavily manual and time-consuming task.

With data and published research found in a variety of sources, each with their own metadata and structure, rounding up all the relevant outputs for the repository is a heavily manual and time-consuming task.

It is also vulnerable to human error, duplications and omissions, resulting in incomplete repositories and researcher profiles. This makes it hard to assess the impact and exposure of an institution's research and the full extent of the faculty's scholarly accomplishments.

A more comprehensive and effective method is needed to ensure that the research information hub reflects the true output of affiliated researchers in the institution. This poses two main challenges: matching scholars to their work; and populating the research information hub with this information at scale. Let's dive deeper to suggest a possible approach.

A rose by any other name...

The most obvious starting point for identifying a researcher's work is the researcher's full name. But names can be messy. An individual may have multiple names that are used interchangeably, such as maiden or married names, surnames and middle names, common diminutives, and the like. Their name may also appear in several different formats, such as an inconsistent use of first initials, formal suffixes, or name order. Finally, there is the challenge of two scholars – sometimes even in the same field and within same institution - with the same name, which can easily lead to misattribution.

A more advanced method for linking researchers and their work is using a persistent identifier code, such as ORCID, that uniquely identifies authors of published academic research. While the use of ORCID and other identifiers is growing, it is not yet common enough to ensure the comprehensiveness of a research information hub. In addition, such identifiers cannot be quickly or automatically linked to relevant assets and information, such as datasets and awarded research grants.

The limitations of such identification codes mean that a more comprehensive and scalable method for matching authors and their research is necessary.

Changing the Paradigm: Smart Harvesting AI

A new technology is redefining the creation of comprehensive and up-to-date research repositories. Known as Smart Harvesting, it incorporates machine learning to ensure each research output is correctly matched to the right researcher.

The solution leverages all variants of a researcher's name, learns their affiliations, research domains, years of professional activity, previously known assets, and other data, in order to correctly match authors and their work.

The solution leverages all variants of a researcher's name, learns their affiliations, research domains, years of professional activity, previously known assets, and other data, in order to correctly match authors and their work.

In examining an output record from a given source (e.g., a national or disciplinary repository, scholarly indexes), Smart Harvesting looks at the author name, title, abstract, subject matter, co-authors, journal, full text, metadata, and author affiliation. Cross-checking those data points, Smart Harvesting assesses whether each record is likely the output of a given researcher. If it is, then the solution ranks the confidence level of its determination as "Very Strong", "Strong" or "Uncertain".

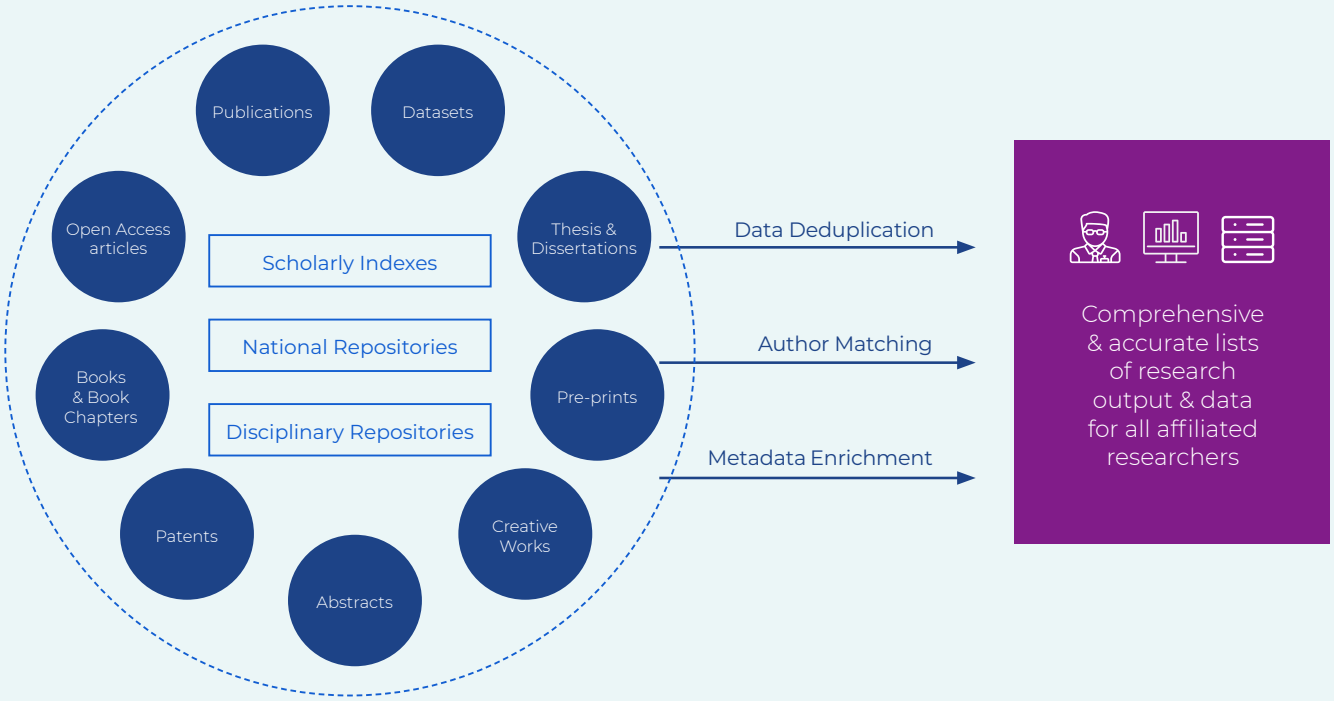
Manually deposited outputs and data already included in the hub can also be enriched using Smart Harvesting. Missing metadata and links can be added, author and co-author misidentifications can be corrected, and duplicated records can be removed.

The first time Smart Harvesting is run for a given researcher it will naturally collect records for all possible outputs. After the initial run, the solution can be scheduled to periodically check for new outputs on an ongoing basis and to continuously enrich existing outputs with additional data.

The first time Smart Harvesting is run for a given researcher it will naturally collect records for all possible outputs. After the initial run, the solution can be scheduled to periodically check for new outputs on an ongoing basis and to continuously enrich existing outputs with additional data.

The more information it collects from these forays, the better able it is to accurately and consistently identify each researcher's work.

The institution can always set conditions for automatic or manual approval of the candidate records to be added, such as the strength of the match ranking. With its intelligent automation, the Smart Harvesting technology can also eliminate the need for manual intervention to remove null or indeterminate results.



ExLibris
A ProQuest Company

Finding relationships with machine learning

The effectiveness of Smart Harvesting in correctly matching researchers and their work is due to unique machine learning algorithms.

The initial stage of training the intelligent solution involves defining three features, or measurable properties, of the data it is assessing: researcher names; general information from the outputs; and semantic content. The Smart Harvesting technology is then exposed to a very large set of output records and researcher profiles, from which it begins to learn about patterns of information in scholarly publications and for each researcher.

The initial stage of training the intelligent solution involves defining three features, or measurable properties, of the data it is assessing: researcher names; general information from the outputs; and semantic content.

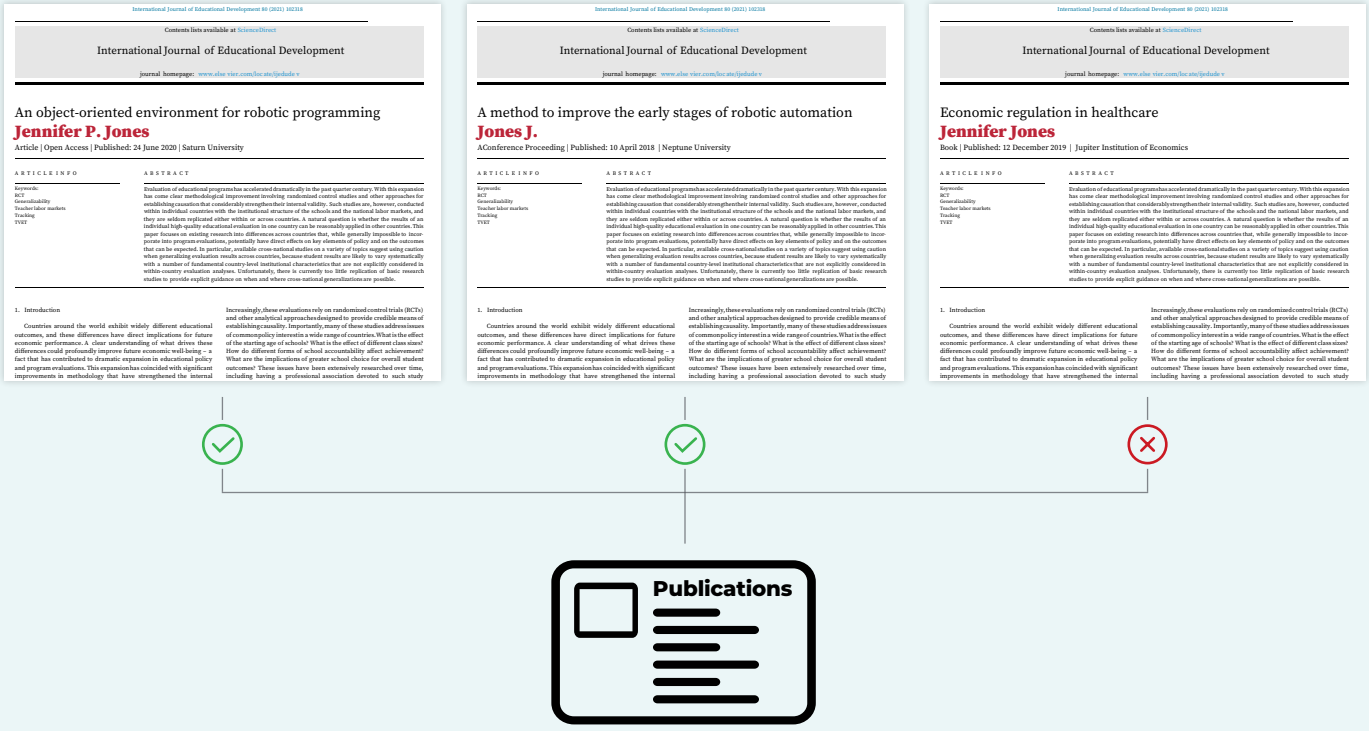
The "name features" include comparing identified names for similarity, the use of initials or nicknames, name frequency, and name variants. "General features" include author affiliation, country, academic discipline, years of activity, and more. The "semantic features" are the presence, absence or combination of key textual elements in abstracts, titles, journal names, and the like.

In assessing the "semantic features", Smart Harvesting uses a natural language processing tool called "word embedding". This tool examines a large corpus of texts and presents a multidimensional map of the relative proximity of selected words, indicating how close they are likely to be to one another.

With its machine learning underway, Smart Harvesting can quickly assess the name, general information and semantic features of a novel research output record in connection with a given researcher. The result is a numerical or qualitative ranking indicating the likelihood that the output is the product of the researcher.

The power of Smart Harvesting AI is its ability to accurately discover the relationship between an output and a researcher and connect related scholarly information. This means that the solution can link research outputs, activities, and entities to enable institutions to create a comprehensive view of research projects and their outcomes.

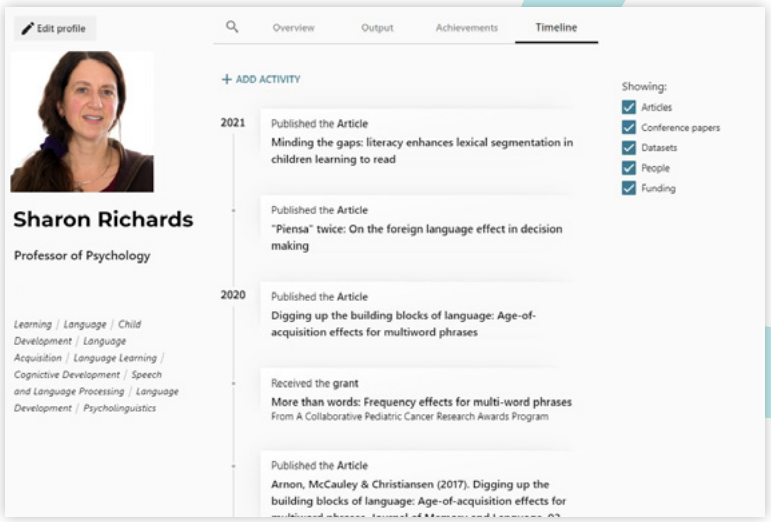
The power of Smart Harvesting AI is its ability to accurately discover the relationship between an output and a researcher and connect related scholarly information.



ExLibris
A ProQuest Company

The Smart Harvesting Impact

As Smart Harvesting automates the updating of researcher profiles and the research information hub with more comprehensive data, it reduces the time and effort librarians and researchers must spend manually maintaining them. This relieves significant pressure felt by researchers who are expected to keep their scholarly records up to date. Research Office staff and Faculty Affairs leaders also clearly benefit when they have access to accurate information on faculty activities, publications and citations.



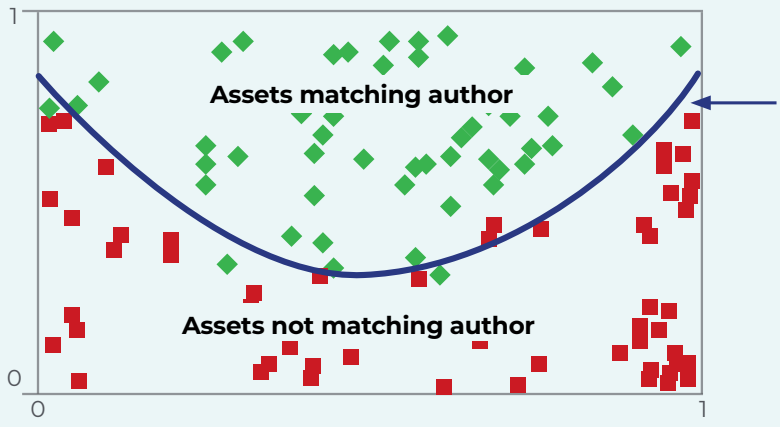
The Smart Harvesting technology can enhance the exposure of an institution's academic research through discovery and networked portals. More comprehensive data and the robust analytics possible with a more accurate repository make it easier to track metrics and set benchmarks. This, in turn, makes it possible to demonstrate the value of research activities to institute stakeholders.

Smart Harvesting in Action

Ex Libris Esploro provides research institutions with Smart Harvesting capabilities for creating an accurate representation of research outputs and data of affiliated researchers. The solution enables universities to comprehensively identify, aggregate and showcase multiple types of scholarly information, activities, and entities in one place, and link everything using automated processes and integrated workflows.

The main source of information on the relevant research assets is the Ex Libris Central Discovery Index (CDI), which includes over four billion records from more than 5,000 content providers and data sources. These include articles, books, chapters, digital resources, datasets, open access content, dissertations, patents, audiovisual content, conference proceedings, and other research outputs. Records in the CDI cover all disciplines, including STEM, Humanities, Social Sciences, and the Arts.

Experience from Esploro customers demonstrates that institutions significantly increase the coverage and completeness of their research portal. This in turn helps universities secure more funding, facilitate collaborations, attract talented researchers, and boost the institution's reputation.



ExLibris
A ProQuest Company

About Ex Libris

Ex Libris, a ProQuest company, is a leading global provider of cloud-based solutions for higher education. Offering SaaS solutions for the management and discovery of the fullspectrum of library and scholarly materials, as well as mobile campus solutions driving student engagement and success, Ex Libris serves thousands of customers in 90 countries. Visit www.exlibrisgroup.com

ExLibris
A ProQuest Company