Profiles Research Networking Software (RNS) Disambiguation Engine

Documentation Version: December 21, 2011

Web Service Version: December 6, 2010

Overview

Profiles RNS uses a Disambiguation Engine to obtain Medline articles for a single person. This is an XML-based web service that can be used independently of the Profiles RNS software. To use the Disambiguation Engine, post an XML Request message to the service containing information about the person. The service will return an XML Response message with a list of matching Pubmed IDs (PMIDs).

Web Service URL

The Profiles RNS Disambiguation Engine web service is hosted at

http://profiles.catalyst.harvard.edu/services/GetPMIDs/default.asp

The data is refreshed on a weekly basis with the latest data from Medline.

The service takes about 5 seconds per person on average, provided you are the only one using the service. If another institution is calling the service at the same time, the run time will be slower. There are currently no restrictions on how frequently you can call the service, though we might have to change in this in the future if usage becomes too high.

Example XML

Below is an example Request XML message. It contains several parts. In the name tag, specify the person's first, middle, last, and suffix names. In the EmailList tag, list the person's email addresses. In the AffiliationList tag, list the person's affiliations. (See the section below on "Affiliation Strings" for information about how to choose affiliations.)

In the LocalDuplicateNames tag, indicate the number of people at your institution that have the name you provided in the Name tag. In most case this value will be 1. However, for common names at a large institution, there might be multiple people that have the same name.

The tags RequireFirstName and MatchThreshold are parameters used by the Profiles RNS Disambiguation Engine to adjust the sensitivity/specificity of the search. (See the section below on "Parameters" for more information.)

If a person has known articles, list these PMIDs in the PMIDAddList tag. To force the Profiles RNS Disambiguation Engine to exclude certain articles, lists these PMIDs in the PMIDExcludeList.

```
<FindPMIDs>
      <Name>
            <First>Griffin</First>
            <Middle>M</Middle>
            <Last>Weber</Last>
            <Suffix />
      </Name>
      <EmailList>
            <email>weber@hms.harvard.edu</email>
            <email>weber@fas.harvard.edu</email>
      </EmailList>
      <AffiliationList>
            <Affiliation>%harvard medical school%</Affiliation>
            <Affiliation>%Massachusetts General%</Affiliation>
            <Affiliation>%Brigham%Women%</Affiliation>
            <Affiliation>%@hms.harvard.edu%</Affiliation>
      </AffiliationList>
      <LocalDuplicateNames>1</LocalDuplicateNames>
      <RequireFirstName>false</RequireFirstName>
      <MatchThreshold>0.98</MatchThreshold>
      <PMIDAddList>
            <PMID>11707567</PMID>
            <PMID>12209713</PMID>
            <PMID>16359929</PMID>
      </PMIDAddList>
      <PMIDExcludeList>
            <PMID>19648504</PMID>
      </PMIDExcludeList>
</FindPMIDs>
```

Below is an example Response XML message, which lists the matching PMIDs.

```
<PMIDList>
      <PMID>11707567</PMID>
      <PMID>11788827</PMID>
      <PMID>11815958</PMID>
      <PMID>12209713</PMID>
      <PMID>12463949</PMID>
      <PMID>12659816</PMID>
      <PMID>15219292</PMID>
      <PMID>15226823</PMID>
      <PMID>16359929</PMID>
      <PMID>18541841</PMID>
      <PMID>19567788</PMID>
      <PMID>20190053</PMID>
</PMIDList>
```

Note that the Response XML only contains PMIDs. It does not return the actual citations. View the following website for information about how to retrieve the full PubMed data for a list of PMIDs.

http://www.ncbi.nlm.nih.gov/corehtml/query/static/efetchlit_help.html

Affiliation Strings

In order for the Profiles RNS Disambiguation Engine to locate publications, you must provide one or more affiliation strings. These are phrases, which can include wildcard characters ("%"), that represent the most likely ways that your researchers will list their affiliations in Medline. Strings are not case sensitive. Selecting affiliation strings is somewhat of an art. The more precise the strings, the easier it is for Profiles to find publications. However, if the strings are too narrow in scope, Profiles might miss some articles. Examples of strings that we use at Harvard include:

```
%Harvard Medical School%
%Beth Israel Deaconess Medical Center%
%BIDMC%
%@hms.harvard.edu%
%Children's Hospital%02115%
```

Examples of strings that we do not use at Harvard because they would be too broad are:

```
%Beth Israel%
%Department of Medicine%
```

Although the affiliation strings help the service find publications, it does not limit the search to just those affiliations. The affiliation strings are used to identify "seed" publications. These are publications that are most likely correct matches. The Profiles RNS Disambiguation Engine then searches all of Medline, using information about the seed publications, such as their titles, MeSH terms, coauthors, and journals, to find additional articles.

Parameters

The Profiles RNS Disambiguation Engine uses a probability model to estimate the likelihood that person wrote a particular publication. Only publications whose match likelihood is above a probability threshold will be returned. By default, the Profiles RNS Disambiguation Engine uses a 98% probability threshold, meaning it will only return publications that are very likely correct matches. You have the option of lowering this threshold by specifying a different value in the MatchThreshold tag in the Request XML. This will reduce the chances that correct publications are missed, but it will increase the chances that incorrect publications are returned. In general, select a low threshold if your goal is to retrieve the "most complete" publication lists, and use a high threshold if your goal is to create the "cleanest" lists. The Profiles RNS software uses the high default threshold when obtaining publications for investigator profiles because it is easy for people to add missing publications, but the website loses much of its value if the search results return incorrect results.

In the Request XML message, if the RequireFirstName tag's value is set to true, then the Profiles RNS Disambiguation Engine will only select seed publications where the author's entire first name (not just the initial) is used. This will improve the specificity of the search (i.e. reduce the number of false matches), though it might also eliminate some true matches. It is good to set this parameter to true when you have two people at your institution with the same last name and same first initial. Another use case is when searching for publications for young investigators (e.g., post-docs), because they will have few publications before 2002, the year when Medline began including author first names. By requiring a first name match for these

people, it should have little effect on correct publication matches, but it has the potential to eliminate older publications that are likely to be incorrect matches.

Notes about Disambiguation

The Profiles RNS Disambiguation Engine will have the most difficulty with common names (e.g., J Smith), names with multiple parts (e.g., a hyphenated last name), names with foreign characters, and people who only recently joined your organization. We are continually working to improve the service to address these issues.

Sometimes it is not obvious to users why the Profiles RNS Disambiguation Engine fails to return any matching publications. We receive this question very frequently. In most cases, this turns out to be because the user thinks his/her name is unique, but there are similar names in Medline that make disambiguation difficult. For example, if there are author names “G. F. Weber” and “G. M. Weber” in Medline, then the person “Griffin Weber” could be either of those two authors, even though it is unlikely that both of those authors’ full first names are “Griffin”.

How to Cite Us

Profiles RNS is a grant-funded project. Our continued support depends on us being able to report on who is using our software. Please let us know how you are using the Profiles RNS Disambiguation Engine by contacting us at profiles@hms.harvard.edu. If you are using this service to populate a website, we ask that you include the following acknowledgment on an easily accessible page of your site:

> This service is made possible by the Profiles Research Networking Software developed under the supervision of Griffin M Weber, MD, PhD, with support from Grant Number 1 UL1 RR025758-01 to Harvard Catalyst: The Harvard Clinical and Translational Science Center from the National Center for Research Resources and support from Harvard University and its affiliated academic healthcare centers.

More Information

For more information about Profiles RNS, please visit

http://profiles.catalyst.harvard.edu

The Harvard development team can be reached at profiles@hms.harvard.edu. We will try to reply promptly, though we cannot guarantee that we will be able to answer all questions.

Commercial support options are available through Recombinant Data Corp. Harvard has no financial relationship with Recombinant, but we recommend them as an Authorized Service Provider for Profiles RNS. For more information, contact Recombinant at results@recomdata.com or call (617) 243-3700.