Comparison of possible corpus source websites regarding the possibilities of automated data acquisition

webofscience.com

The basic structure of the profile URLs is as follows:

```
https://www.webofscience.com/wos/author/record/1177820
```

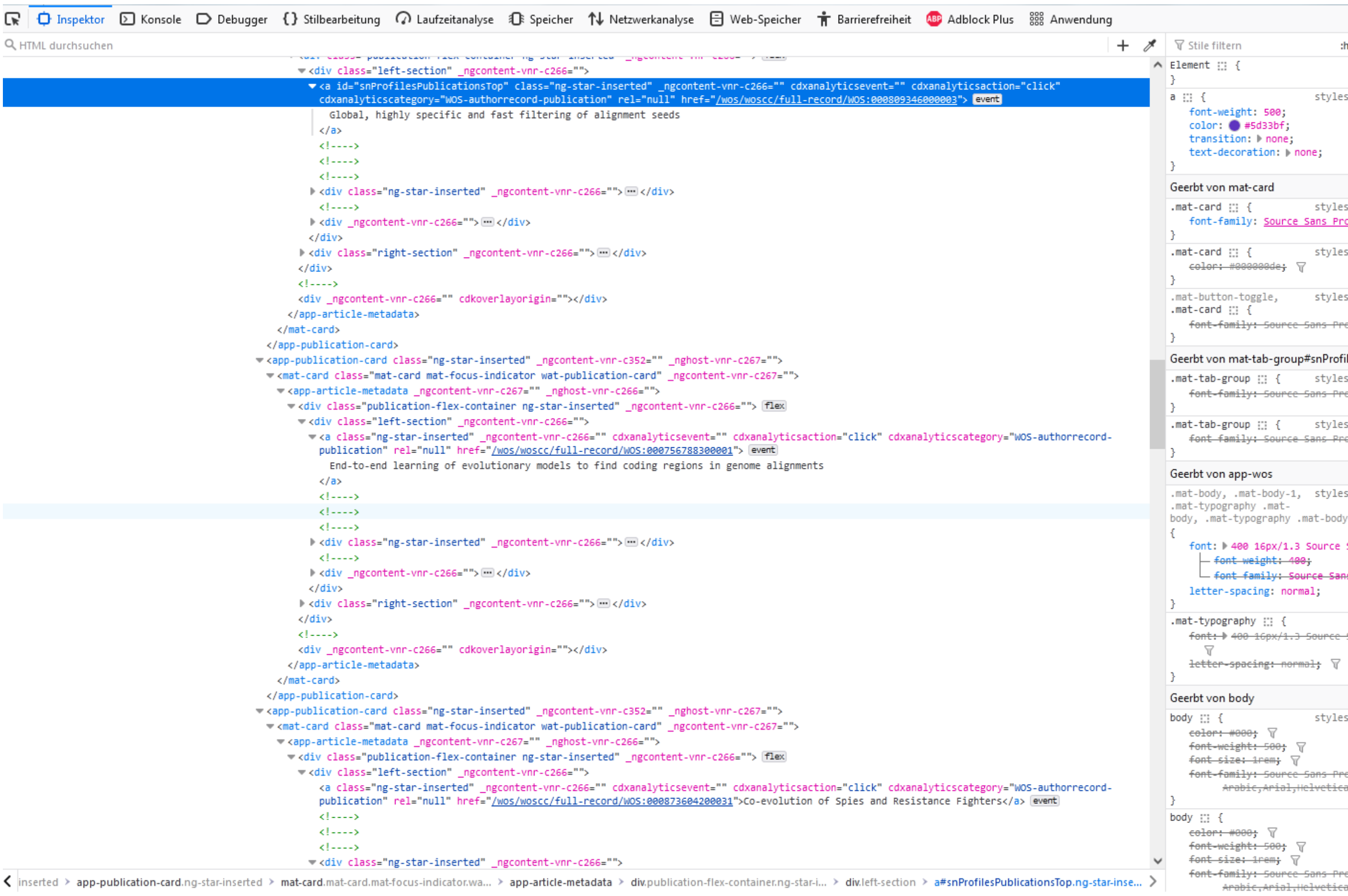
Since it is unlikely that the profile ID is known from the beginning, some scripting will have to be done.

On

```
https://www.webofscience.com/wos/author/search
```

a form will have to be filled out to initiate the search process. If the name is unambiguous, a link pattern such as in the first example will be followed. Additionally it has to be noted that a session detector will prevent web crawling without being logged in, which is a downside, because it could lead to complications during scripting.

The profile pages themselves contain up to 50 publications (abstracts only) per page, which I deem sufficient for our purposes, as seen in the inspector

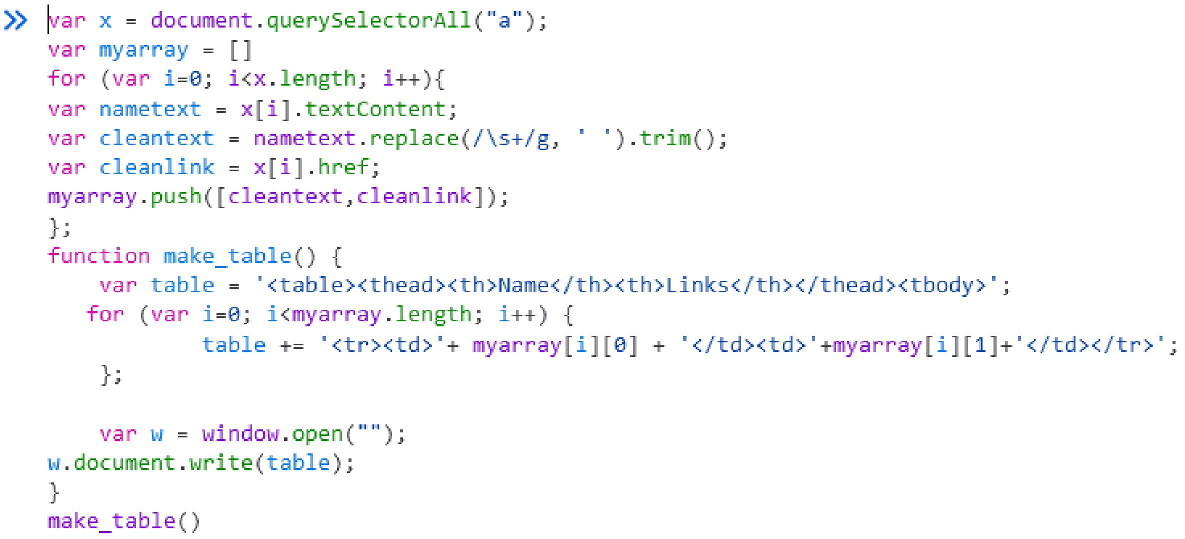

Inspektor
Konsole
Debugger
Stilbearbeitung
Laufzeitanalyse
Speicher
Netzwerkanalyse
Web-Speicher
Barrierefreiheit
Adblock Plus
Anwendung
HTML durchsuchen
Stile filtern
Geerbt von mat-card
Geerbt von mat-tab-group#snProfil
Geerbt von app-wos
Geerbt von body
event
flex
inserted > app-publication-card.ng-star-inserted > mat-card.mat-card.mat-focus-indicator.wa... > app-article-metadata > div.publication-flex-container.ng-star-i... > div.left-section > a#snProfilesPublicationsTop.ng-star-inse...

```
      <div class="left-section" _ngcontent-vnr-c266="">
        <a id="snProfilesPublicationsTop" class="ng-star-inserted" _ngcontent-vnr-c266="" cdxanalyticsevent="" cdxanalyticsaction="click"
        cdxanalyticscategory="WOS-authorrecord-publication" rel="null" href="/wos/woscc/full-record/WOS:000809346000003"> event
          Global, highly specific and fast filtering of alignment seeds
        </a>
        <!---->
        <!---->
        <!---->
        <div class="ng-star-inserted" _ngcontent-vnr-c266="">…</div>
        <!---->
        <div _ngcontent-vnr-c266="">…</div>
      </div>
      <div class="right-section" _ngcontent-vnr-c266="">…</div>
    </div>
    <!---->
    <div _ngcontent-vnr-c266="" cdkoverlayorigin=""></div>
  </app-article-metadata>
 </mat-card>
</app-publication-card>
<app-publication-card class="ng-star-inserted" _ngcontent-vnr-c352="" _nghost-vnr-c267="">
 <mat-card class="mat-card mat-focus-indicator wat-publication-card" _ngcontent-vnr-c267="">
  <app-article-metadata _ngcontent-vnr-c267="" _nghost-vnr-c266="">
    <div class="publication-flex-container ng-star-inserted" _ngcontent-vnr-c266=""> flex
      <div class="left-section" _ngcontent-vnr-c266="">
        <a class="ng-star-inserted" _ngcontent-vnr-c266="" cdxanalyticsevent="" cdxanalyticsaction="click" cdxanalyticscategory="WOS-authorrecord-
        publication" rel="null" href="/wos/woscc/full-record/WOS:000756788300001"> event
          End-to-end learning of evolutionary models to find coding regions in genome alignments
        </a>
        <!---->
        <!---->
        <!---->
        <div class="ng-star-inserted" _ngcontent-vnr-c266="">…</div>
        <!---->
        <div _ngcontent-vnr-c266="">…</div>
      </div>
      <div class="right-section" _ngcontent-vnr-c266="">…</div>
    </div>
    <!---->
    <div _ngcontent-vnr-c266="" cdkoverlayorigin=""></div>
  </app-article-metadata>
 </mat-card>
</app-publication-card>
<app-publication-card class="ng-star-inserted" _ngcontent-vnr-c352="" _nghost-vnr-c267="">
 <mat-card class="mat-card mat-focus-indicator wat-publication-card" _ngcontent-vnr-c267="">
  <app-article-metadata _ngcontent-vnr-c267="" _nghost-vnr-c266="">
    <div class="publication-flex-container ng-star-inserted" _ngcontent-vnr-c266=""> flex
      <div class="left-section" _ngcontent-vnr-c266="">
        <a class="ng-star-inserted" _ngcontent-vnr-c266="" cdxanalyticsevent="" cdxanalyticsaction="click" cdxanalyticscategory="WOS-authorrecord-
        publication" rel="null" href="/wos/woscc/full-record/WOS:000873604200031">Co-evolution of Spies and Resistance Fighters</a> event
        <!---->
        <!---->
        <!---->
        <div class="ng-star-inserted" _ngcontent-vnr-c266="">
```

```
Element {
}
a {                                   styles
   font-weight: 500;
   color: #5d33bf;
   transition: none;
   text-decoration: none;
}
.mat-card {                           styles
   font-family: Source Sans Pro
}
.mat-card {                           styles
   ~~color: #000000de;~~
}
.mat-button-toggle,                   styles
.mat-card {
   ~~font-family: Source Sans Pro~~
}
.mat-tab-group {                      styles
   ~~font-family: Source Sans Pro~~
}
.mat-tab-group {                      styles
   ~~font-family: Source Sans Pro~~
}
.mat-body, .mat-body-1,               styles
.mat-typography .mat-
body, .mat-typography .mat-body
{
   font: 400 16px/1.3 Source S
     ├ ~~font-weight: 400;~~
     └ ~~font-family: Source Sans~~
   letter-spacing: normal;
}
.mat-typography {
   ~~font: 400 16px/1.3 Source S~~
   ~~letter-spacing: normal;~~
}
body {                                styles
   ~~color: #000;~~
   ~~font-weight: 500;~~
   ~~font-size: 1rem;~~
   ~~font-family: Source Sans Pro~~
       ~~Arabic,Arial,Helvetica~~
}
body {
   ~~color: #000;~~
   ~~font-weight: 500;~~
   ~~font-size: 1rem;~~
   ~~font-family: Source Sans Pro~~
       ~~Arabic,Arial,Helvetica~~
```

Figure 1: Excerpt of the html body of the example profile page using the Firefox code inspector

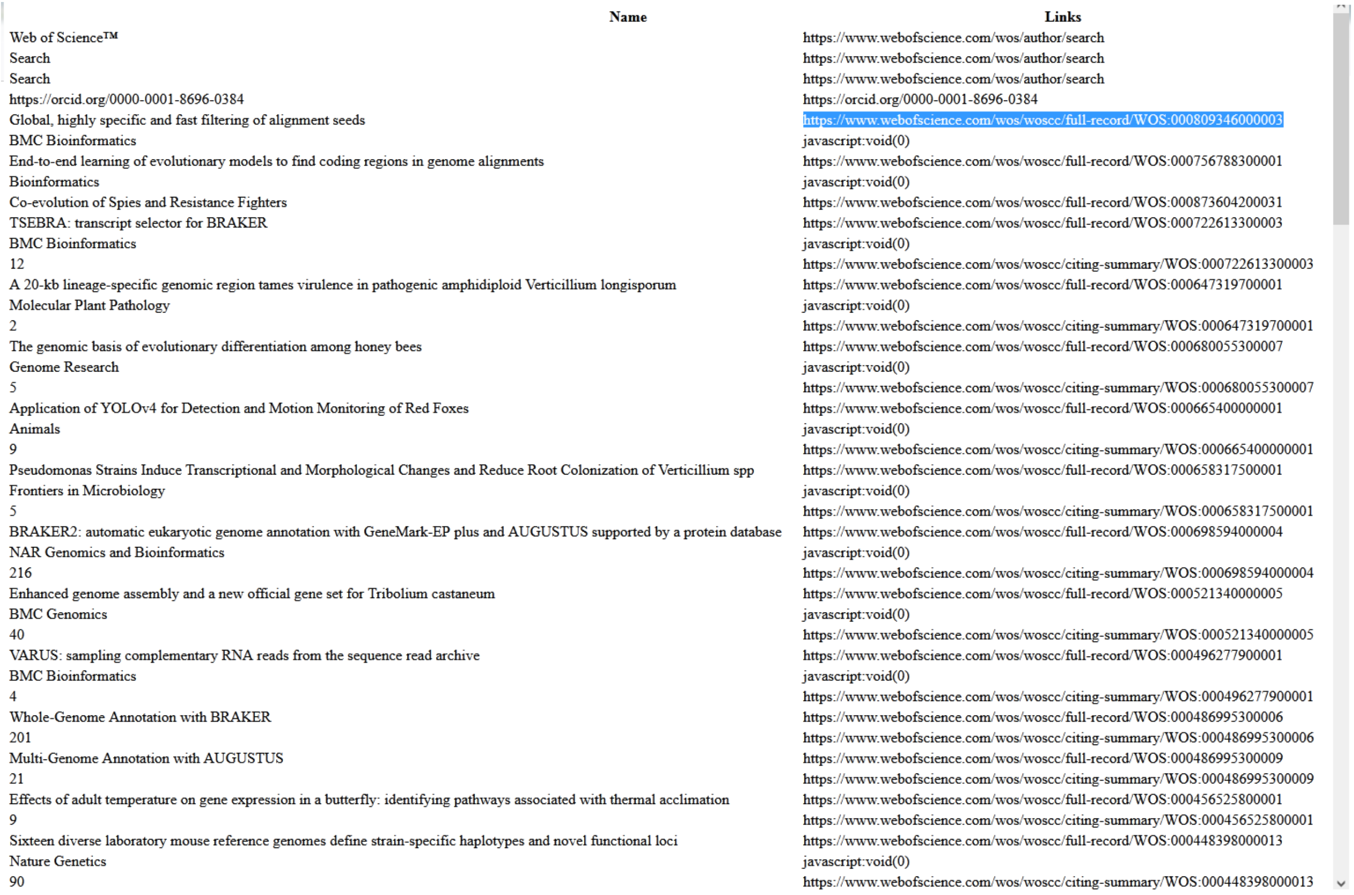
A great advantage is the way the individual publications are linked. They follow a tight pattern and are self-contained on the page. A small javascript code snippet can be used to reliably extract name of the publication and link and save it to array, which contents can be saved to a text file of desired format with minor alterations and filtering for desired links only.



```
>> var x = document.querySelectorAll("a");
   var myarray = []
   for (var i=0; i<x.length; i++){
   var nametext = x[i].textContent;
   var cleantext = nametext.replace(/\s+/g, ' ').trim();
   var cleanlink = x[i].href;
   myarray.push([cleantext,cleanlink]);
   };
   function make_table() {
       var table = '<table><thead><th>Name</th><th>Links</th></thead><tbody>';
      for (var i=0; i<myarray.length; i++) {
               table += '<tr><td>'+ myarray[i][0] + '</td><td>'+myarray[i][1]+'</td></tr>';
       };

       var w = window.open("");
   w.document.write(table);
   }
   make_table()
```

Figure 2: JS code snippet to extract all links and corresponding names into an array and display results in a table. Source: https://towardsdatascience.com/quickly-extract-all-links-from-a-web-page-using-javascript-and-the-browser-console-49bb6f48127b



Name	Links
Web of Science™	https://www.webofscience.com/wos/author/search
Search	https://www.webofscience.com/wos/author/search
Search	https://www.webofscience.com/wos/author/search
https://orcid.org/0000-0001-8696-0384	https://orcid.org/0000-0001-8696-0384
Global, highly specific and fast filtering of alignment seeds	https://www.webofscience.com/wos/woscc/full-record/WOS:000809346000003
BMC Bioinformatics	javascript:void(0)
End-to-end learning of evolutionary models to find coding regions in genome alignments	https://www.webofscience.com/wos/woscc/full-record/WOS:000756788300001
Bioinformatics	javascript:void(0)
Co-evolution of Spies and Resistance Fighters	https://www.webofscience.com/wos/woscc/full-record/WOS:000873604200031
TSEBRA: transcript selector for BRAKER	https://www.webofscience.com/wos/woscc/full-record/WOS:000722613300003
BMC Bioinformatics	javascript:void(0)
12	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000722613300003
A 20-kb lineage-specific genomic region tames virulence in pathogenic amphidiploid Verticillium longisporum	https://www.webofscience.com/wos/woscc/full-record/WOS:000647319700001
Molecular Plant Pathology	javascript:void(0)
2	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000647319700001
The genomic basis of evolutionary differentiation among honey bees	https://www.webofscience.com/wos/woscc/full-record/WOS:000680055300007
Genome Research	javascript:void(0)
5	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000680055300007
Application of YOLOv4 for Detection and Motion Monitoring of Red Foxes	https://www.webofscience.com/wos/woscc/full-record/WOS:000665400000001
Animals	javascript:void(0)
9	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000665400000001
Pseudomonas Strains Induce Transcriptional and Morphological Changes and Reduce Root Colonization of Verticillium spp	https://www.webofscience.com/wos/woscc/full-record/WOS:000658317500001
Frontiers in Microbiology	javascript:void(0)
5	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000658317500001
BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP plus and AUGUSTUS supported by a protein database	https://www.webofscience.com/wos/woscc/full-record/WOS:000698594000004
NAR Genomics and Bioinformatics	javascript:void(0)
216	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000698594000004
Enhanced genome assembly and a new official gene set for Tribolium castaneum	https://www.webofscience.com/wos/woscc/full-record/WOS:000521340000005
BMC Genomics	javascript:void(0)
40	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000521340000005
VARUS: sampling complementary RNA reads from the sequence read archive	https://www.webofscience.com/wos/woscc/full-record/WOS:000496277900001
BMC Bioinformatics	javascript:void(0)
4	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000496277900001
Whole-Genome Annotation with BRAKER	https://www.webofscience.com/wos/woscc/full-record/WOS:000486995300006
201	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000486995300006
Multi-Genome Annotation with AUGUSTUS	https://www.webofscience.com/wos/woscc/full-record/WOS:000486995300009
21	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000486995300009
Effects of adult temperature on gene expression in a butterfly: identifying pathways associated with thermal acclimation	https://www.webofscience.com/wos/woscc/full-record/WOS:000456525800001
9	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000456525800001
Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci	https://www.webofscience.com/wos/woscc/full-record/WOS:000448398000013
Nature Genetics	javascript:void(0)
90	https://www.webofscience.com/wos/woscc/citing-summary/WOS:000448398000013

Figure 3: Table generated by code in Fig.2 for page https://www.webofscience.com/wos/author/record/1177820

Now the individual pages of the publications can be accessed and downloaded as text files, which after a little bit of trimming could directly be used for tokenization. In summary this procedure would result in up to 50 abstracts for each person as per sEt1.

Google Scholar

The same basic priciples as for webofscience are applicable. The main advantage is that no login is required to view profiles. Since it is still necessary to enter names in the search engine, thereafter clicking on a profile to view it, webcrawling will still have to be done.

It's still unclear whether Google Scholar or webofscience is content-wise better suited to be used for data acquisition.