CASE STUDY: Content Extraction of Web of Science
SUBMISSION BY: Karen E Brown, Collection Development Librarian
DATE: November 2017

Synopsis/Executive Summary:

A lecturer in the Faculty of Information approached his Subject Librarian for assistance with a research project which involved text and data mining of the Web of Science database. The Subject Librarian approached Information Resources for assistance and the Collection Development Librarian received the following outline from the lecturer:

To confirm, these are the specifications of our research project:
1. Requests from Web Of Science (WOS): To download for 20 journals -- the abstract + BibTeX (bibinfo) + keywords. (*no fulltext needed*)
   Username is my monash.edu email (i.e. this email).

2. Methodology: once downloaded as offline .BIB (BibTeX) files, to apply natural language processing / classification on the resulting bibinfo files' metadata.

3. Start time: anytime next week will be good (dates flexible).

4. Computer / IP: will be from an internal Monash staff office workstation, in the Clayton Faculty of IT.

Please let me know if there is anything else you need from us at this stage.

On behalf of our research team from [Various National and International Institutions]
many thanks!

No further information was needed before discussing with the licensor, Clarivate Analytics.

Findings:

Clarivate Analytics and the Collection Librarian discussed the request and the following four options were presented to Monash University for this particular query:

1. WoS Platform direct query searching
   Not data-mining as such, but regular direct search in the Web of Science platform will most likely be manageable for small projects like the one proposed, with an export limited to 500 records at a time.

2. Use of API(s)
   Monash (either the Library or researcher directly) may run queries on Clarivate’s APIs. These are currently available for the Web of Science Core Collection, as used by Monash to bring citation data into Pure. Please see the attached file “Web Services Expanded and Links AMR”. Current Terms of Use for APIs restrict access to Monash University’s own publications, though depending on the scope of the project, Clarivate can consider an amendment.
3. Clarivate custom data project
   This can be a custom data project where the licensor exports the raw data for a
   custom selection, then the raw data may be analysed directly by Monash University.
   An example of a custom data project is attached with University of Melbourne looking
   at the use of words related to “Global Warming” in Scientific literature.

4. Outright purchase of the Web of Science data
   This gives licensee everything the licensor has for all indices in XML format and the
   data archive is typically shipped to the licensee on a hard disk. This is used in
   benchmarking and comparison of WoS to other large datasets, but could be used for
   other projects.

Discussion:

The simple but effective outline from the lecturer had all the relevant information, with the
exception of whether or not they planned to mine non-Monash University publications, which
sits outside of the current allowable license terms. The Collection Librarian assumed the
researcher did not want to limit the mining to Monash University-only publications and so a
possible amendment will be signed as a result.

After presenting only options 1 and 2 to the researcher for simplicity, the Collection Librarian
gleaned further information relating to the researcher’s mode of thought around his
responsibilities with the data. He indicated that he read an online user agreement, looking for
a specific mention of how he could extract and use the data. He most likely read our Library
“terms of use” before authenticating via Ezproxy.

The researcher also mentioned he couldn’t find anything in particular limiting his mining, but
he thought there must be some restrictions around “fair use”. He then contacted the Subject
Librarian to check with the Library.

Conclusions:
The researcher went through the appropriate steps to try and ascertain the process and restrictions around data mining. The Collection Development Librarian did have to speak to the vendor to ascertain licensing conditions/process.