Computational Access and Use of Texts and Data behind Paywalls

Challenges and Resources

Katie Zimmerman, kbzimmer@mit.edu
Ye Li, yel@mit.edu

January 25, 2022
Outline

- Introduction to computational access and use
- Copyright and computational use
- Contracts for access
Which category does your field of study or interest belong to?

Arts and Humanities
Science, Engineering, and Technology
Social Sciences
Business and Management
Medicine and Health Sciences
Other
Why are you interested in "Computational Access and Use of Texts and Data behind Paywalls"?
Data and Texts

Traditional Use

READ
Digest and processed in human brain

Computational Use

Computer processed first
Reveal pattern and insights through generating analytics and applying algorithms
Computational research has been rising fast since 2010 and applied to many disciplines.

Computational Access and Use
Access and use (mostly aggregated) data and texts through computational means, often secondary use
An Example of Chemical Reactions


ASKCOS model available at [https://github.com/connorcoley/ASKCOS](https://github.com/connorcoley/ASKCOS)

Robotically reconfigurable platform
Performing multistep chemical syntheses planned by AI
Computational Use

Types of Data and Texts

(Texts) Raw Data

(Bibliographic) Metadata

—

Texts
Narratives

Artworks
Music

Numeric

(A+2)^2

Others

Survey Data

Experimental Data
“Small” datasets from individual lab studies
Batch datasets from high-throughput experiments

Simulated Data
Texts and Data Curation

Text mining
Data curation

Converted to machine-readable (e.g. OCR)

Indexed

Aggregated

Evaluated

Structured

Enriched with metadata and provenance

Tagged with a linked standardized vocabulary / ontologies / schema

In a standardized format

Value-added!
Sources of Texts and Data

- Converted to machine-readable (e.g. OCR)
- Indexed
- Aggregated
- Evaluated
- Structured
- Registered
- Enriched with metadata and provenance
- Tagged with a linked standardized vocabulary / ontologies / schema
- In a standardized format

Sources of Texts and Data

- Books
- Journal articles
- Multimedia
- Reference works (Handbooks, Encyclopedia etc.)
- Reference datasets
- Metadata databases
- Databases of curated / extracted data
- Texts and data repositories
Providers for Sources of Texts and Data

Increasing concerns on protection of intellectual property and business interests

More likely behind paywalls!
Challenges in meeting needs for computational access and use of data and texts

- Justify and seek for additional funding
- Solve technological issues in facilitating computational access
- Address licensing issues involving intellectual property (IP) and contract laws
- Polish negotiation skills to reach agreement without giving in
### Example 1: CrossRef - Metadata Retrieval

<table>
<thead>
<tr>
<th>Feature / Option</th>
<th>Metadata Search</th>
<th>Simple Text Query</th>
<th>REST API</th>
<th>XML API</th>
<th>OAI-PMH</th>
<th>OpenURL</th>
<th>Metadata Plus (OAI-PMH + REST API)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface for people or machines?</td>
<td>People</td>
<td>People</td>
<td>People (low volume and occasional use) and machines</td>
<td>Machines</td>
<td>Machines</td>
<td>Machines</td>
<td>Machines</td>
</tr>
<tr>
<td>Output format</td>
<td>Text, JSON</td>
<td>Text</td>
<td>JSON</td>
<td>XML</td>
<td>XML</td>
<td>XML</td>
<td>JSON, XML</td>
</tr>
<tr>
<td>Suitable for citation matching?</td>
<td>Yes (low volume)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Supports volume downloads?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Suitable for usage type</td>
<td>Frequent and occasional</td>
<td>Frequent and occasional</td>
<td>Frequent and occasional</td>
<td>Frequent</td>
<td>Frequent</td>
<td>Frequent</td>
<td>Frequent and occasional</td>
</tr>
<tr>
<td>Free or cost?</td>
<td>Free</td>
<td>Free</td>
<td>Free and cost options</td>
<td>Free and cost options</td>
<td>Cost for full service, more options available</td>
<td>Free</td>
<td>Cost</td>
</tr>
<tr>
<td>Includes all available metadata?</td>
<td>In JSON only</td>
<td>DOIs only</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Bibliographic only</td>
<td>Yes</td>
</tr>
<tr>
<td>Documentation</td>
<td>Metadata Search</td>
<td>Simple Text Query</td>
<td>REST API</td>
<td>XML API</td>
<td>OAI-PMH</td>
<td>OpenURL</td>
<td>Metadata Plus (OAI-PMH + REST API)</td>
</tr>
</tbody>
</table>

**API Documentation**
https://api.crossref.org/help

https://www.crossref.org/services/metadata-retrieval/
Accessed July 2021
Example 2: JSTOR -- Text Mining

Platform to explore, analyze, and request dataset of fulltext journal articles published by JSTOR.

https://constellate.org/builder

Free to access with 1000 article download limitation; Exploring a tiered pricing model for participating institutions to access datasets larger than 25,000 items.

Text and data mining (TDM)
Machine-ready articles to support your R&D’s digital future

What can we provide?

We have made our full journals catalogue available for use with TDM and machine learning applications, supplied as XML with tables and accompanying images. Bulk delivery of archive material can be supplemented by secure delivery of new publications.

We have practical licensing options, and a dedicated data support team on hand to help with queries.

Example 3: RSC - text and data mining
What is Entellect?

Entellect is a Life Sciences data integration platform that enables insights from data to drive effective innovation. It brings together your data, Elsevier’s data, and data from global researchers.

With a modern architecture that adheres to FAIR Principles, Entellect facilitates access to clean, reusable data and metadata, enabling R&D scientists to optimize decision making and cycle time. It enables better data governance and helps drive accurate AI/ML based discovery.

Example 4: Elsevier Entellect - Data Integration Platform

Technical Requirements for Texts and Data

Quantity:
- Small amount at once
- Large amount at once

Format:
- Print
- Electronic
- Electronic Standardized

Structure:
- Appeal to human eyes
- Embedded and linked

Access:
- Manual retrieval, point-and-click
- Automated retrieval, programmatic interaction
Format and Structure Requirements

Format
- Print
- Electronic
  - Standardized

Structure
- Appeal to human eyes
  - Embedded and linked
- Print Only
  - Digitized
    - Digitized with OCR
      - Born-Digital
        - Born-Digital with Metadata and in Standardized Format
      - Born-Digital with Metadata and in Standardized Format
    - Born-Digital with Open and Standardized Metadata and in Open and Standardized Format
Computational Access Requirements

Access
- Manual retrieval
  - Point-and-click
  - No digital copy
- Automated retrieval
  - Programmatic interaction
  - API service
- Copy / Paste
- Batch Download
- Web crawler
- Point-and-click Export
- Batch Delivery (e.g. Email, FTP etc.)
Computational Access Requirements

Access

Manual retrieval
Point-and-click

Automated retrieval
Programmatic interaction

No digital copy

Copy / Paste
Batch Download
Web crawler

Point-and-click Export
Batch Delivery (e.g. Email, FTP etc.)

New trend - Platform allowing Interaction with data and texts without downloading them

API service

- Most interactive, flexible, and secure
- Requiring most structured and interoperable data

Resources

- ProQuest TDM Studio: https://about.proquest.com/products-services/research-tools/TDM-Studio.html
- Elsevier Entellect: https://www.elsevier.com/solutions/entellect
- HathiTrust Research Center: https://analytics.hathitrust.org/
API (Application Programming Interface) Services

- Enabling users to retrieve data without knowing the server structure
- Can carry tokens for authentication
- Query as an URL, or Python, Ruby, R codes etc.
- Response in structured and tagged format (e.g. JSON, XML etc.)
How familiar are you with using API?

- Not sure
- Never used API
- Know basics / used API lightly
- Interacted with APIs fluently
  - Used API extensively, including integration with other system
Activity I: A Glance at API access

- Experience another way of access
  - See how results can come to you dynamically
  - See possibilities to integrate into your daily workflow and other applications
- You do not have to know any programming language
- You do not have to become an expert of API to understand what’s important


The activity will take 5 - 7 min and we will remind you when 1 min is left.
Assess FAIRness of Data and Texts Not in the Open


GoogleForm version can be accessed at: https://forms.gle/xGGJahtBrTa7PGL76
## Accessible

### What and How accessible?

<table>
<thead>
<tr>
<th>Access</th>
<th>No access to data or metadata</th>
<th>Access to metadata only</th>
<th>Unspecified conditional access</th>
<th>Conditional access specified in individual license agreement</th>
<th>Conditional access specified in general term of agreement</th>
<th>Publicly accessible</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Additional fee?

<table>
<thead>
<tr>
<th>Access</th>
<th>Additional one-time cost to build infrastructure and workflow</th>
<th>Additional cost for individual research projects and researchers</th>
<th>Additional cost for campus licensing</th>
<th>No additional fee beyond campus-wide subscription</th>
<th>Free</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Access Mechanism?

<table>
<thead>
<tr>
<th>Access</th>
<th>No access to data through electronic means</th>
<th>By individual arrangement</th>
<th>File downloadable or crawlable from an online location</th>
<th>Platform allowing interaction without downloading</th>
<th>Non-standard web service</th>
<th>Standard web service API</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Metadata availability even if the data is no longer available?

<table>
<thead>
<tr>
<th>Access</th>
<th>No</th>
<th>Not sure</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## R1.1 License and usage right

<table>
<thead>
<tr>
<th>Approaches?</th>
<th>In-kind sponsorship</th>
<th>General term of use for all users</th>
<th>Individual license agreement with researchers and/or per project</th>
<th>Addendum to campus license for traditional use</th>
<th>Openly licensed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage?</td>
<td>No local storage, remote interaction only</td>
<td>Controlled local storage with time limitation</td>
<td>Local storage with necessary security measure for individual access specified in the license</td>
<td>Local storage allowing campus-wide access</td>
<td></td>
</tr>
<tr>
<td>Redistribution?</td>
<td>No re-distribution</td>
<td>Sample portion to re-disseminate with research outputs for validation</td>
<td>Re-distribute with research output</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resulting IP?</td>
<td>Shared IP</td>
<td></td>
<td></td>
<td>Belongs to researchers and their institutions; can be openly licensed</td>
<td></td>
</tr>
</tbody>
</table>

## R1.2 Provenance captured?

| No provenance information recorded | Partially recorded | Fully recorded in a text format | Fully recorded in a machine-actionable format |

## R1.3 Meet domain-relevant community standard?

| No domain-relevant community standard available | Standard in development | Do not meet standard | Meet standard |
Copyright and Computational Use
When does copyright apply?

17 U.S.C. § 102. Subject matter of copyright: In general

(a) Copyright protection subsists, in accordance with this title, in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device. Works of authorship include the following categories:

(1) literary works;
(2) musical works, including any accompanying words;
(3) dramatic works, including any accompanying music;
(4) pantomimes and choreographic works;
(5) pictorial, graphic, and sculptural works;
(6) motion pictures and other audiovisual works;
(7) sound recordings; and
(8) architectural works.
When does copyright apply?

17 U.S.C. § 102. Subject matter of copyright: In general (cont.)

(b) In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.
Uncopyrightable

- Facts
- Ideas
- Insufficiently creative combinations of facts and ideas
- Federal government works
- Works that have aged into the public domain
Copyrightability

Google Books Ngram Viewer

https://books.google.com/ngrams
Copyrightability

Google Books Ngram Viewer

https://books.google.com/ngrams

A787305

airbreathers_NOUN

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>393791</td>
<td>apocalypse_NOUN</td>
<td>1988</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>393792</td>
<td>apocalypse_NOUN</td>
<td>1989</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>393793</td>
<td>apocalypse_NOUN</td>
<td>1990</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>393794</td>
<td>apocalypse_NOUN</td>
<td>1991</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>393795</td>
<td>apocalypse_NOUN</td>
<td>1992</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>393796</td>
<td>apocalypse_NOUN</td>
<td>1993</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>393797</td>
<td>apocalypse_NOUN</td>
<td>1994</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>393798</td>
<td>apocalypse_NOUN</td>
<td>1995</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>393799</td>
<td>apocalypse_NOUN</td>
<td>1996</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>393800</td>
<td>apocalypse_NOUN</td>
<td>1997</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>393801</td>
<td>apocalypse_NOUN</td>
<td>1998</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>393802</td>
<td>apocalypse_NOUN</td>
<td>1999</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>393803</td>
<td>apocalypse_NOUN</td>
<td>2000</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>393804</td>
<td>apocalypse_NOUN</td>
<td>2001</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>
In the US, text mining is a fair use

Four fair use factors

- Purpose and character of your use
- Nature of their work
- Amount taken
- Market effect of your use
HathiTrust and Google Books cases
The Crucible: Drama in Two Acts - Page 32
Arthur Miller - 1954

FOUND INSIDE – PAGE 32
But . . . surely you know what a jabberer she is? Did you tell them that? MARY. Mister Proctor, in open court she near to choked us all to death. PROCTOR. JH. How choked you? MARY. She sent her spirit out. ELIZABETH. Oh, Mary, Mary, surely . . .

books.google.com › books
Lao She and the Chinese Revolution - Page 62
Ranbir Vohra - 1974

FOUND INSIDE – PAGE 62
There was no way out; Lao She had to follow his own instincts and be a loner. The year 1931 was also important to Lao She for another, more personal reason. In this year, at the age of 33, he married Miss Hu Hsieh-
So, what does that get us?

Create your corpus → Do your analysis → Publish the dataset?

Publish the results
Who owns what?
Contracts for access
Contract v. Copyright

Clause 3. No text and data mining allowed...
3.3 **No Diminution of Rights.** Nothing in this Agreement … shall be interpreted to diminish the rights and privileges of the Licensee or Authorized Users with respect to any of the Licensed Materials, including exceptions or limitations to the exclusive rights of copyright owners, such as fair use, under Section 107 of the U.S. Copyright Act. ([LibLicense](LibLicense))
Restrictions on use

1.4 Restrictions on Use of Subscribed Products.
Except as expressly stated in this Agreement or otherwise permitted in writing by vendor, the Subscriber and its Authorized Users may not:

- abridge, modify, translate or create any derivative work based on the Subscribed Products, except to the extent necessary to make them perceptible on a computer screen to Authorized Users;

- remove, obscure or modify in any way any copyright notices, other notices or disclaimers as they appear in the Subscribed Products;

- use any robots, spiders, crawlers or other automated downloading programs, algorithms or devices to continuously and automatically search, scrape, extract, deep link, index or disrupt the working of the Subscribed Products; except as may otherwise be expressly permitted through a separate API vendor makes available under separate agreement with individual developers; or

- substantially or systematically reproduce, retain or redistribute the Subscribed Products.
Additional terms with implications for computational use

- Who can access the content? What about research collaborators?
- What are the terms around sharing the content?
Contract Review
Access & Storage

“The text and data mining must be performed using a Publisher approved API service, and the Licensee must abide by any limits on load rates and download rates conveyed to Licensee from time to time.”
Access & Storage I - What potential issues do you see in this contract section?

“The text and data mining must be performed using a Publisher approved API service, and the Licensee must abide by any limits on load rates and download rates conveyed to Licensee from time to time.”
Access & Storage

“In no event will the Licensee or any of its Authorized Users use or share the results from the text and data mining in a manner that competes with the Platform or the Licensed Content, for the use or benefit of any third party, or otherwise harm the economic interest of the Publisher, including, without limitation, by posting the results on the Internet for the public to access, regardless of whether the results can be reconstructed to its original, human readable form.”
Access & Storage II - What potential issues do you see in this contract section?

“In no event will the Licensee or any of its Authorized Users use or share the results from the text and data mining in a manner that competes with the Platform or the Licensed Content, for the use or benefit of any third party, or otherwise harm the economic interest of the Publisher, including, without limitation, by posting the results on the Internet for the public to access, regardless of whether the results can be reconstructed to its original, human readable form.”
Access & Storage

“For the sake of clarity, Licensee shall not share the results from the text and data mining with any third party, including, without limitation, Licensee or Participating Institutions (if they are not also the Licensee).”
Access & Storage III - What potential issues do you see in this contract section?

“For the sake of clarity, Licensee shall not share the results from the text and data mining with any third party, including, without limitation, Licensee or Participating Institutions (if they are not also the Licensee).”
Access & Storage

“The extracted or raw data and the text and data mining results may only be stored and analyzed within the Licensee’s secure network that is not accessible by any unauthorized third parties.

The text and data mining may only be performed during the Authorized Window set forth above. During the Authorized Window, if the Publisher, in its sole discretion, determines that the text and data mining is causing a burden on the Publisher’s network or interferes with other users’ access to the Platform and the Licensed Content, Licensee will temporarily suspend such activity upon Publisher’s request, and only resume once Publisher notifies the Licensee to do so.”
Access & Storage IV - What potential issues do you see in this contract section?

"The extracted or raw data and the text and data mining results may only be stored and analyzed within the Licensee’s secure network that is not accessible by any unauthorized third parties.

The text and data mining may only be performed during the Authorized Window set forth above. During the Authorized Window, if the Publisher reasonably determines that the text and data mining is causing an unreasonable burden on the Publisher’s network or substantially interferes with other users’ access to the Platform and the Licensed Content, Licensee will temporarily suspend such activity upon Publisher’s request, and only resume once Publisher notifies the Licensee to do so.”
Audits

“The Publisher will be entitled, at its sole discretion, costs and expense, to audit the Subscriber’s and any engaged third party vendor’s, entire TDM environment, which would include the extracted Data, the TDM Output, the hardware environment, software deployed and security controls in place and to ensure compliance with the terms and conditions of this Agreement.”
Audits - What potential issues do you see in this contract section?

“The Publisher will be entitled, at its sole discretion, costs and expense, to audit the Subscriber’s and any engaged third party vendor’s, entire TDM environment, which would include the extracted Data, the TDM Output, the hardware environment, software deployed and security controls in place and to ensure compliance with the terms and conditions of this Agreement.”
“Upon termination of any grant of license rights to MIT under this Addendum, or upon termination of the entire Access Agreement, howsoever arising, MIT shall procure the destruction of any copies of the Publisher Metadata and Abstracts and Citations (and PDF copies of fully articles, if any) if locally loaded for purposes of Text and Data Mining. MIT shall provide evidence of destruction to the Publisher’s Publications no later than 30 days after termination signed by an authorized officer of MIT.”
After the Party - What potential issues do you see in this contract section?

"Upon termination of any grant of license rights to MIT under this Addendum, or upon termination of the entire Access Agreement, howsoever arising, MIT shall procure the destruction of any copies of the Publisher Metadata and Abstracts and Citations (and PDF copies of fully articles, if any) if locally loaded for purposes of Text and Data Mining. MIT shall provide evidence of destruction to the Publisher’s Publications no later than 30 days after termination signed by an authorized officer of MIT."
IP Rights of Research Results

“To the extent that Subscriber Results contain [vendor’s] intellectual property rights, such rights shall be of [vendor] and may not be distributed or commercialized by Subscriber without [vendor’s] prior written approval. [Vendor] is granted the right of first refusal to purchase from Subscriber the rights in and to any or all of the Subscriber Results.”
“To the extent that Subscriber Results contain [vendor’s] intellectual property rights, such rights shall be of [vendor] and may not be distributed or commercialized by Subscriber without [vendor’s] prior written approval. [Vendor] is granted the right of first refusal to purchase from Subscriber the rights in and to any or all of the Subscriber Results.”
Local Resources

1. **Your Subject Specialist!**
2. Library Contact for Access Questions: [textmine@mit.edu](mailto:textmine@mit.edu)
3. Library Contact for Data Support: [data-management@mit.edu](mailto:data-management@mit.edu)
4. Support for individual contracts at MIT Office of Strategic Alliances & Technology Transfer: [https://osatt.mit.edu/](https://osatt.mit.edu/)

Links with list of specific resources

- [Text and Data Mining at MIT](https://libraries.mit.edu/services/text-data-mining)
- [APIs for Scholarly Resources](https://libraries.mit.edu/services/scholarly-api)