

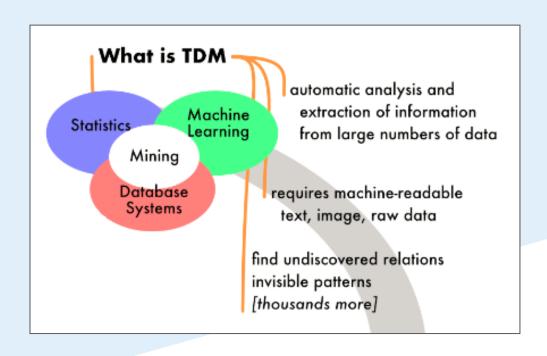
TDM Text and Data Mining

Text and Data Mining

Text and Data Mining (TDM) is the (semi-)automated process of selecting and analyzing large amounts of text or data resources for purposes such as

- searching,
- finding patterns,
- discovering relationships,
- semantic analysis and
- learning how content relates to ideas and needs

in a way that can provide valuable information needed for studies, research, etc.



Text and Data Mining => Use Cases

- Searching and archiving
- Metadata display
- Knowledge retrieval
- Multisource library
- Online portals
- Apps, products and solutions
- Fraud detection (banking, insurance) and spam filtering
- Some build in our APIs into their local search facility to simply improve their researchers' discovery experience, some others do specific term or data mining.
- Some go very deep into semantic text mining, some others want to retrieve hidden research data structures as part of given science programs.
- Some even may use TDM to investigate competitor activities and staff.

Text and Data Mining => A TDM Framework for Springer Nature

Text and Data Mining => A TDM Framework for Springer Nature



https://www.springernature.com/text-and-data-mining

Text and Data Mining

- A growing part of Springer Nature's journal articles is published open access.
- TDM is usually allowed without restrictions for these publications since the majority of Springer Nature open access content is licensed under CC-BY.

Text and Data Mining at CrossRef

- Springer Nature is participating in the Crossref TDM working group and is recommending Crossref services for pan-publisher TDM.
- For further information see http://tdmsupport.crossref.org/

Text and Data Mining for Subscribers

- For subscribers Springer Nature offers a large variety of TDM tools such as metadata and fulltext APIs, applicable to both open access and subscribed resources.
- See https://api.springernature.com



Text and Data Mining for Non-Subscribers

- Non-subscribers are offered a variety of TDM tools for our Open Access resources, such as our Open Access fulltext API (see https://api.springernature.com).
- TDM requirements from non-subscribers for pay-walled content are treated on a case-by-case basis.
- Please contact tdm@springernature.com.

Image Mining

At the moment Springer Nature does not offer any API for image mining.

Argumentation Mining

- "Argumentation mining aims to automatically detect, classify and structure argumentation in text (...) i.e. understanding the content of serial arguments, their linguistic structure, the relationship between the preceding and following arguments (...)." (Mochales, R. & Moens, MF. Artif Intell Law (2011) 19: 1. https://doi.org/10.1007/s10506-010-9104-x)
- Argumentation mining can be considered as a subset of text mining. If you are
 planning to locally store non-open-access content during an argumentation mining
 project, please get in contact with tdm@springernature.com to discuss options.

Springer Nature APIs

- **metadata**: general xml metadata using PAM standard (Prism Aggregate format); uses prism and dc (dublincore) namespaces and elements. In use but fixed (i.e., no changes will be made).
- meta/v1: based on metadata API. Has PAM standard but additional elements added for various business reasons. Fluid API to which changes can be easily made.
- **openaccess**: complete, "fulltext" (i.e., body) xml where available; when fulltext/body not available, xml metadata is provided.
- **integro**: journal-level information. Based on Casper journal xml (e.g., http://www.live.springer.com/api/journal/10653.xml).
- xmldata: JATS format for fulltext as available; else metadata.
 - Throttle: recommended crawling rate of 150 hits/minute and 7500/day.
 - Key: special key provided for TDM.
- All APIs have xml and json except for xmldata and integro.
- A full picture of Springer Nature's API offerings with detailed information, examples and API key sign-up can be looked up under https://api.springernature.com.

Springer Nature APIs

- Citations APIs (internal)
 - Article/chapter count: returns the number of citations for a given DOI.
 - Show citations: returns actual citations for a given DOI.
 - Book-level count: returns total count of citations for all the chapters in a book.
 - Journal-level count: returns total count of citations for all the articles in a journal.
 - Recent citations: provides recent citation data from Crossref.

Linked Open Data Publishing So Far



Welcome to Springer Nature SciGraph, our Linked Open Data offering which aggregates data sources from Springer Nature and key partners from the scholarly domain. The Linked Open Data platform collates information from across the research landscape, for example funders, research projects, conferences, affiliations and publications.

Additional data, such as citations, patents, clinical trials and usage numbers will follow over time. This high quality data from trusted and reliable sources provides a rich semantic description of how information is related, as well as enabling innovative visualizations of the scholarly domain.

By doing so, Springer Nature SciGraph overcomes former boundaries by relating comprehensive information about the research landscape. It represents a further step in data integration and it will continue to grow organically. Our aim is to increase the discoverability of high quality data as larger parts of our datasets are being made available under CC-BY licensing.



Any questions?
Please contact us. (>)

Latest Datasets Download

Data Explorer

Licensing Information

At a glance:

- 1,5B+ triples
- CC-BY license

Metadata about:

- All articles + abstracts
- All books + chapters
- Grants (200k)
- Journals
- Subjects / Taxonomies
- Core Ontology

www.springernature.com/scigraph

Springer Nature APIs => General Constraints

Example Queries

- ISBN: http://api.springer.com/meta/v1/pam?q=isbn:978-0-387-79148-7%

Springer Nature APIs => General Constraints

- Example Queries (cont'd)

metadata API

- - General Elements
 - identifier (doi)
 - title (of article/chapter)
 - creator (author)
 - publicationName (book, journal title)
 - issn/isbn (electronic)
 - genre (ArticleType/ChapterType)
 - journalID
 - volume
 - number (issue)
 - startingPage
 - openAccess
 - doi
 - publisher
 - publicationDate
 - url (resolver)
 - copyright
 - abstract

```
<pam:article>
                <dc:identifier>doi:10.1007/s10404-009-0428-3</dc:identifier>
                <dc:title>Numerical analysis of electrokinetic transport in micro-nanofluidic
                <dc:creator>Wang, Yi</dc:creator>
                <dc:creator>Pant, Kapil</dc:creator>
                <dc:creator>Chen, Zhijian</dc:creator>
                <dc:creator>Wang, Guiren</dc:creator>
                <dc:creator>Diffey, William F.</dc:creator>
                <dc:creator>Ashley, Paul</dc:creator>
                <dc:creator>Sundaram, Shivshankar</dc:creator>
                sm:publicationName>Microfluidics and Nanofluidics
                sm:issn>1613-4990
                sm:genre>OriginalPaper
                <journalId>10404</journalId>
                sm:volume>7
                sm:number>5
                <prism:startingPage>683</prism:startingPage>
                <openAccess>false</openAccess>
                sm:doi>10.1007/s10404-009-0428-3
                <dc:publisher>Springer</dc:publisher>
                sm:publicationDate>2009-10-31/prism:publicationDate>
                <prism:url>http://dx.doi.org/10.1007/s10404-009-0428-3</prism:url>
                copyright>@2009 Springer-Verlag
             </pam:article>
        </xhtml:head>
        <xhtml:body>
          <h1>Abstract</h1>
          The phenomenon of enrichment of charged analytes due to the presence of an electr;
             <CitationRef
CitationID="CR17">2005b</CitationRef>): (1) ion transport of background electrolytes (BGE) and (
          </xhtml:body>
     </pam:message>
```

meta API

- - General Elements
 - (Same as metadata, except as below)
 - Casper pdf and html urls
 - issueType (Regular, Supplement)
 - topicalCollection (ArticleCollection; e.g., topicalcollection:Topics in Flow Control)
 - Can query with openaccess:True

```
<pam:article>
             <dc:identifier>doi:10.1007/s11671-010-9562-9</dc:identifier>
             cprism:url format="html"
platform="web">http://link.springer.com/openurl/fulltext?id=doi:10.1007/s11671-010-9562-9
             cprism:url format="pdf"
platform="web">http://link.springer.com/openurl/pdf?id=doi:10.1007/s11671-010-9562-9</pris
             <dc:title>Template Synthesis of Carbon Nanofibers Containing Linear Mesocad
             <dc:creator>Wang, Yongwen</dc:creator>
             <dc:creator>Zheng, Mingbo</dc:creator>
             <dc:creator>Lu, Hongling</dc:creator>
             <dc:creator>Feng, Shaoging</dc:creator>
             <dc:creator>Ji, Guangbin</dc:creator>
             <dc:creator>Cao, Jieming</dc:creator>
             prism:publicationName>Nanoscale Research Letters
             sm:issn>1556-276X
             prism:genre>Nano Express
             <journalId>11671</journalId>
             <prism:volume />
             <prism:number />
             <issueType />
             <topicalCollection />
             <prism:startingPage>1</prism:startingPage>
             <openAccess>true</openAccess>
             <prism:doi>10.1007/s11671-010-9562-9</prism:doi>
             <dc:publisher>Springer</dc:publisher>
             prism:publicationDate>2010-05-09
             <prism:url>http://dx.doi.org/10.1007/s11671-010-9562-9</prism:url>
             sm:copyright>@2010 The Author(s)
           </pam:article>
    </xhtml:head>
    <xhtml:body>
       <h1>Abstract</h1>
       Carbon nanofibers containing linear mesocage arrays were prepared via evaporat.
    </xhtml:bodv>
  </pam:message>
```

Open Access API

- Returns fulltext JATS xml for Open Access articles and chapters

```
<Section1 ID="Sec2">
                              <Heading>Case Report/Heading>
                              <Para>A 59-year-old man was admitted to the vascular surgery department as a result of acute
                                 <InternalRef</pre>
RefID="Fig1">1</InternalRef>). A decision was made to perform surgical embolectomy and endovascular intervention (EVAR). Th
                                 <Figure Category="Standard" Float="Yes" ID="Fig1">
                                    <Caption Language="En">
                                       <CaptionNumber>Fig. 1</CaptionNumber>
                                       <CaptionContent>
                                          <SimplePara>Angio-CT axial scan, aortic aneurysm below the level of the renal art
Type="Bold">A</Emphasis> Within the anterior aortic wall there is a narrow, irregular soft tissue infiltration that suggest
                                              <Emphasis Type="Italic">arrow</Emphasis>).
                                             <Emphasis Type="Bold">B</Emphasis> Axial scan at the level of femoral arteries
                                             <Emphasis
Type="Italic">arrows</Emphasis>) reveals an embolism of the left artery. The right femoral artery is patent
                                          </51mplePara>
                                       </CaptionContent>
```

xmldata/fulltext API => General Constraints

Equals constraints

DOI: http://api.springer.com/xmldata/jats?q=doi:10.1007/s10741-007-9067-5-2

Subject:

Keyword:

Pub:

Year:

Country:

xmldata/fulltext API => General Constraints

- Equals constraints (cont'd)

xmldata/fulltext API

- Querying constraints
 - doi
 - journalid
 - issn
 - journal
 - isbn
 - book
 - pub (publication)
 - title
 - name
 - subject
 - date
 - year
 - keyword
 - type (Book or Journal)
 - country
 - JournalOnlineFirst (true or false)
 - Orgname
 - empty (no constraint, general search)

- Returns fulltext JATS xml as available; for internal use (except as indicated elsewhere).

```
</ChapterHeader>
 <Body>
    <Section1 ID="Sec1">
       <Heading>Introduction</Heading>
       <Para
o a thorough analysis of the industrial needs and of the limitations of current simulation
          <CitationRef CitationID="CR1">1</CitationRef>], which represents the thermal-
his comprehensive program, aims at building a new software platform for advanced two-phase
tions meeting the industrial needs. The different modeling strategies can be taken into ac
       </Para>
       <Para
r consists of a general presentation of the model implemented in NEPTUNE CFD code, origina
    </Section1>
    <Section1 ID="Sec2">
       <Heading>Multi-Fluid Volume Averaged Models/Heading>
       <Section2 ID="Sec3">
          <Heading>General Equations</Heading>
id volume averaged models play a central role for numerically handling multi-phase flows h
ds to a set of separate mean balance equations for each phase [
             <CitationRef CitationID="CR2">2</CitationRef>,
             <CitationRef
CitationRef>] which looks like to the single-phase conservation equations of mass, momentu
          </Para>
```

near query

- The purpose of the "near" query is to find terms within x number of words from each other. In the highlighted case, we use "engineering" which can have many meanings and contexts:
 - The "engineering NEAR/3 locomotive" API call

yields four results where "engineering is within 3 words of "locomotive" which is what was specified in the API call.

The "engineering NEAR/5 metals" API call

yields four results where "engineering is within 5 words of "metals".

JATS API

Industry standard XML schema. Development completed May 2017 for journals and books.

```
<records>
  <article article-type="research-article" dtd-version="1.1" xml:lang="en">
     <front>
        <journal-meta>
            <journal-id journal-id-type="publisher-id">11583</journal-id>
           <journal-title-group>
              <journal-title>Il Nuovo Cimento Series 10</journal-title>
              <abbrev-journal-title abbrev-type="publisher">Nuovo Cim</abbrev-journal-title>
           </journal-title-group>
           <issn pub-type="ppub">0029-6341</issn>
           <issn pub-type="epub">1827-6121</issn>
           <publisher>
              <publisher-name>Società Italiana di Fisica</publisher-name>
              <publisher-loc>Bologna
           </publisher>
           <custom-meta-group>
              <custom-meta>
                 <meta-name>toc-levels</meta-name>
                 <meta-value>0</meta-value>
              </custom-meta>
              <custom-meta>
                 <meta-name>volume-type</meta-name>
                 <meta-value>Regular</meta-value>
              </custom-meta>
              <custom-meta>
                 <meta-name>journal-subject-primary</meta-name>
                 <meta-value>Physics</meta-value>
              </custom-meta>
              <custom-meta>
                 <meta-name>journal-subject-secondary</meta-name>
                 <meta-value>Physics, general</meta-value>
              </custom-meta>
```

Citations APIs (for Journal Articles)

- Article/chapter count: returns the number of citations for a given DOI
 - http://citationsapi.nkb3.org/article/citations/count?doi=10.1007/s00216-007-1222-2
- Show citations: returns actual citations for a given DOI.
 - http://citationsapi.nkb3.org/article/citations?doi=10.1007/s00216-007-1222-2

 Data for citations APIs comes from the responses to requests we make to Crossref to find out who is citing our data.

```
-<citation>
   <target doi>10.1007/s00216-007-1222-2</target doi>
 - < journal_cite fl_count="0" submissionId="" datetimeRecevied="2014-07-15T22:23:36.619273+02:00">
     <issn type="print">0028-8233</issn>
     <issn type="electronic">1175-8775</issn>
     <journal title>New Zealand Journal of Agricultural Research
     <iournal abbreviation>New Zealand Journal of Agricultural Research
   -<article title>
       Analysis of major fatty acids in milk produced from high-quality grazed pasture
     </article title>
   -<contributors>
      -<contributor first-author="true" sequence="first" contributor role="author" citation formatted na
          <given name>I</given name>
          <surname>Rugoho</surname>
       </contributor>
      - <contributor first-author="false" sequence="additional" contributor role="author" citation format
          <given name>Y</given name>
         <surname>Liu</surname>
       </contributor>
     - <contributor first-author="false" sequence="additional" contributor role="author"
       citation_formatted_name="Dewhurst, RJ">
          <given name>RJ</given name>
          <surname>Dewhurst</surname>
       </contributor>
     </contributors>
     <first page>1</first page>
     <year>2014</year>
     <publication type>full text/publication type>
     <doi type="journal article">10.1080/00288233.2014.899505</doi>
```

Citations APIs (for Book Chapters and Journals)

- Book-level count: returns total count of citations for all the chapters in a book.
 - http://citationsapi.springerservice.com/book/citations/count?doi=10.1007/978-3-642-24040-9
- Journal-level count: returns total count of citations for all the articles in a journal.
 - http://citationsapi.springerservice.com/journal/367/citations/count
- Citations Combination Count and Display Book API:
 - http://citationsapi.springerservice.com/book/citations?doi=10.1007/978-3-540-30299-5%s=51%p=20

```
<citations-result><doi>10.1007/978-3-540-30299-5</doi>\doi>\doi>\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi\doi<pre
         <contributor first-author="true" sequence="first" contributor role="author">
          <given name>B.</given name>
          <surname>Buchholz</surname>
         </contributor>
         <contributor first-author="false" sequence="additional" contributor role="author">
          <given name>B.</given name>
          <surname>Kühnreich</surname>
         </contributor>
         <contributor first-author="false" sequence="additional" contributor role="author">
          <given name>H. G. J.</given name>
          <surname>Smit</surname>
         </contributor>
         <contributor first-author="false" sequence="additional" contributor role="author">
          <given name>V.</given name>
          <surname>Ebert</surname>
```

Citations APIs

- Recent citations: live feed receiving daily Crossref data; not just about recent Springer Nature articles; this is about ANY new article citing (/citation) a Springer Nature article (/item) from any time.
- http://citations.springer.com/xml/recent_citations/2015-02-03?start=1&end=10
- http://citations.springer.com/xml/recent_citations/2015-02-03?maxResults=10

```
<?xml version="1.0" encoding="UTF-8"?>citations total="115154"></pre
<target doi>10.1007/978-1-61737-960-4 2</target doi><journal cite fl count="1" submissionId="1375909629"</pre>
datetimeRecevied="2015-02-03T02:35:27.3877+01:00"><issn type="print">19326254</issn><journal title>Journal
of Tissue Engineering and Regenerative Medicine</journal title><journal abbreviation>J Tissue Eng Regen
Med</journal abbreviation><article title>Explant culture: a simple, reproducible, efficient and economic
technique for isolation of mesenchymal stromal cells from human adipose tissue and
lipoaspirate</article title><contributors><contributor first-author="true" citation formatted name="Priya,
Nancy"><given_name>Nancy</given_name><surname>Priya</surname></contributor><contributor first-
author="false" citation formatted name="Sarcar, Shilpita"><given name>Shilpita</given name>
<surname>Sarcar</surname></contributor><contributor first-author="false"</pre>
citation_formatted_name="Majumdar, Anish Sen"><given_name>Anish Sen</given_name>
<surname>Majumdar</surname></contributor><contributor first-author="false"</pre>
citation formatted name="SundarRaj, Swathi"><given name>Swathi</given name><surname>SundarRaj</surname>
</contributor></contributors><volume>8</volume><issue>9</issue><first_page>706</first_page>
<year>2014full text</publication type><doi
type="journal article">10.1002/term.1569</doi><textRefs></textRefs></journal cite></citation><item></item>
</recent citation><recent citation><citation><target doi>10.1007/978-3-642-25501-4 1</target doi>
<issn type="print">1793-5245</issn><issn type="electronic">1793-7159</issn><journal title>International
Journal of Biomathematics</journal title><journal abbreviation>Int. J. Biomath.</journal abbreviation>
<article title>On the complexity of average path length for biological networks and
```

Search on a Portal

- User runs search on portal, e. g. hip cartilage
 - Though the user is searching the portal our API is being called and the results are returned on the fly.

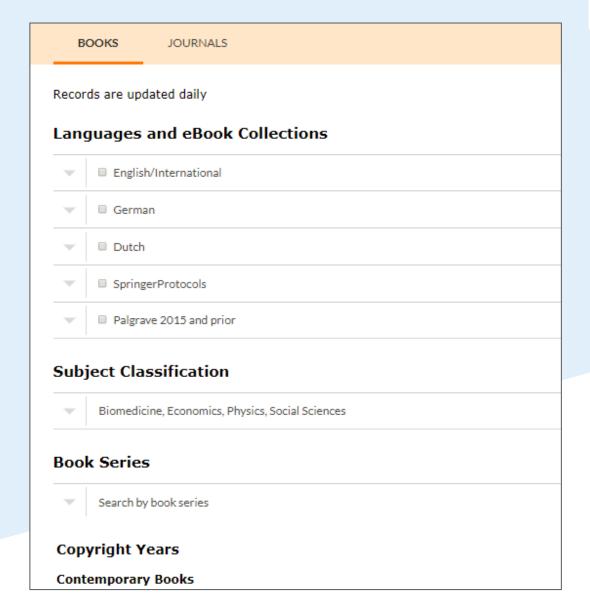




Metadata Delivery

- Springer Nature also offers direct metadata delivery options in various formats, such as JATS, Dublin Core, ONIX, or MARC records, using different protocols (ftp/ftps, sftp) including for metadata harvesting (OAI-PMH).
- For direct metadata download features via the metadata downloader see https://metadata.springernature.com.
- If you have requests for metadata delivery please contact dds.support@springernature.com.

Metadata Delivery



Content and Data Enrichment

- We are continuously enriching our Springer Nature content with PIDs and other smart data annotation.
- Collaboration with InfoChem => We just finished the annotation of 4 million Springer Nature documents (articles, book chapters, major reference works) with 604,000 unique chemical compounds associated.
- These resources are ready to be data mined.



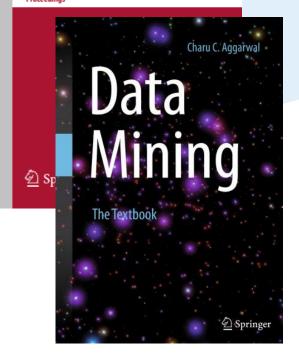
Springer Nature Publications About TDM







Second International Conference, DMBD 2017 Fukuoka, Japan, July 27 – August 1, 2017 Proceedings





Thank you!

Henning Schoenenberger
henning.schoenenberger@springernature.com

Text and Data Mining at Springer Nature
https://www.springernature.com/text-and-data-mining

Contact:

tdm@springernature.com

