Crawling full texts, metadata, and media on thieme-connect.com

File Versioning

Version	Changes	Author	Date
1.1	Added graphical_abstract to article interface	Selina Andrews	2024-03-25
1.0	Initial document version	Selina Andrews	2024-03-25

Table of Contents

Background: Crawling full texts and metadata

Thieme Connect offers full texts and metadata for licensed content to its partners. Access to back content published before the partnership was established is provided via the crawling interface described in this documentation.

Back content is used to describe all articles and books published in the past, as these are not included in the regular export processes.

Full texts are provided where available in XML and/or PDF format. The XML schemas and DTDs used are described in the appendix of this document. Some older content (before the year 2002) and some congress abstracts (before the year 2014) only provide an abstract XML, while most recent articles and books provide a full-text XML.

Metadata is provided in the XML document per article or book. The XML schemas and DTDs used are described in the appendix of this document.

Media is provided as direct links to the media elements like images, videos, and supporting information.

All elements can be selectively crawled through an XML tagging system in the crawling interface, which is described in this documentation.

Important: Please only crawl the files you need to archive by using the tagging system. This reduces traffic on both your servers and ours.

The crawling interface provides multiple starting points: journals can be found by crawling the whole list of journals in Thieme Connect or by DOI (journal, issue, and article level), while books can be crawled by book or chapter DOI. If you need to crawl book back content, please contact export-ejournals@thieme.de to request a list of your licensed back content books.

To give your crawler access to your licensed data, our export team will be in touch with you requesting the necessary authentication data. We need you to provide:

- Name and country of your organization / institution
- Name and e-mail address of your contact person
- IP addresses / ranges for crawling
- If possible, a time frame in which you expect the back content crawling to be completed

If you encounter any problems while crawling or have further questions, please contact export-ejournals@thieme.de for help.

Technical Constraints

Note: This documentation does not include a detailed description on how to implement a crawler, but focuses on the technical details of the data Thieme Connect provides.

When implementing the crawler, please take the following **technical constraints** into account:

- max. 1 request / second
- include the session cookie received from the first response in every further request

Not complying with these constraints might lead to your requests being denied to keep our system stable.

1 Crawling Structure

Thieme Connect supports the following crawling approaches for journal articles and books, respectively:

Approach	Journal Articles	Books
Crawling by title list	yes	no
Crawling by DOI	yes *(journal, issue, article)*	yes *(book, chapter)*

Crawling by title list requires no prior knowledge of the data to crawl and can be used for an initial crawl of the whole licensed journal content.

- for articles: see 1.1.1 Home Page

Crawling by DOI requires the crawler to be supplied with a list of DOIs to call the corresponding URLs. This approach can be started on any layer of the hierarchical DOI structure (journal > issue > article / book > chapter).

- for articles: see 1.1.2 Journal Page, 1.1.3 Journal Year Page, 1.1.4 Issue Page, 1.1.5 Article Page
- for books: see 1.2.1 Book Page, 1.2.2 Chapter Page

Trying to access unlicensed data leads to an HTTP FORBIDDEN 403 error.

Please make sure to limit your crawling activity to one request / second and to include the session cookie with every request.

1.1 Journal Crawling Structure

Journal articles can be crawled **by title list** (starting at 1.1.1 Home Page) or **by DOI** (starting at 1.1.2 Journal Page with the journal DOI, 1.1.3 Journal Year Page with the journal DOI and year, 1.1.4 Issue Page with the issue DOI, or 1.1.5 Article Page with the article DOI).

Example: Crawling journal articles by title list

1. Call home page (https://www.thieme-connect.com/products/all/home.html) and identify journal link to follow
2. Call journal link (ex. https://www.thieme-connect.com/products/ejournals/journal/10.1055/s-00000001) and identify year link to follow
3. Call journal year link (ex. https://www.thieme-connect.com/products/ejournals/journal/10.1055/s-00000001/2024) and identify issue to follow
4. Call issue link (ex. https://www.thieme-connect.com/products/ejournals/issue/10.1055/s-014-59080) and find metatag article_list_url with link to issue interface
5. Call issue interface (see 2.1 Journal Crawling Interfaces)

Example: Crawling journal articles by article DOI

1. Identify article DOI (ex. 10.1055/a-2238-8971)
2. Call article page with article DOI (ex. https://www.thieme-connect.com/products/ejournals/abstract/10.1055/a-2238-8971) and find metatag article_crawling_links_url with link to article interface
3. Call article interface (see 2.1.2 Article Interface)

1.1.1 Home Page

URL	https://www.thieme-connect.com/products/all/home.html

The home page provides a list of all journals in the HTML version at the XPath provided below:

XPath	Journals
//*[@id="tabContent-Ejournals"]/div/ul/li/div/span/h2/a/@href	Active Journals
//*[@id="tabContent-Ejournals"]/div/ul/div/ul/li/a/@href	Discontinued / Transferred Journals

The href attribute contains a link to the journal page (see 1.1.2 Journal Page) while the link text is the journal title.

1.1.2 Journal Page

URL	https://www.thieme-connect.com/products/ejournals/journal/[Journal DOI]

The journal page can either be accessed through the link from the home page (see 1.1.1 Home Page) or via the URL above with the journal DOI inserted at the end.

The journal page provides a link to all journal year pages in the HTML version at the XPath provided below:

XPath	//*[@id="yearSelectContainer"]/ul/li/a

The href attribute contains a link to the journal year page (see 1.1.3 Journal Year Page) while the link text is the year linked to.

1.1.3 Journal Year Page

URL	https://www.thieme-connect.com/products/ejournals/issues/[Journal DOI]/[Year]

The journal year page can either be accessed through the link from the journal page (see 1.1.2 Journal Page) or via the URL above with the journal DOI and year inserted at the marked positions.

The journal year page provides links to all issue pages in the HTML version at the XPath provided below:

XPath	//*[@id="content"]/div[2..]/a

Starting at div[2], the href attribute contains a link to the issue page (see 1.1.4 Issue Page) while the link text provides the issue number as well as the first and last page of the issue in the format “#issue: first_page-last_page”.

1.1.4 Issue Page

URL	https://www.thieme-connect.com/products/ejournals/issue/[Issue DOI]

The issue page can either be accessed through the link from the journal year page (see 1.1.3 Journal Year Page) or via the URL above with the issue DOI inserted at the end.

The issue page provides a link to the issue interface through the following metatag:

Metatag	<meta name="article_list_url" content="[URL]">

The content attribute contains a link to the issue interface (see 2.1.1 Issue Interface).

1.1.5 Article Page

URL	https://www.thieme-connect.com/products/ejournals/abstract/[Article DOI]

The article page can be accessed via the URL above with the article DOI inserted at the end.

The article page provides a link to the article interface through the following metatag:

Metatag	<meta name="article_crawling_links_url" content="[URL]" article-type="[Type]" publication_type="[Publication Type]">

The content attribute contains a link to the article interface (see 2.1.2 Article Interface).

The article-type attribute contains one of the following article types to filter by:

- congress-abstract
- erratum
- evaluation
- magazine
- promotional
- scientific

The publication_type attribute contains one of the following publication types to filter by:

- efirst[1]
- issue
- am
- continuous-publication

Both article-type and publication_type can be used to select specific subsets of articles to crawl.

1.2 Book Crawling Structure

Books can be crawled by either book DOI (starting at 1.2.1 Book Page) or chapter DOI (starting at 1.2.2 Chapter Page).

Currently, books cannot be crawled by title list. If you need to crawl book back content, please contact export-ejournals@thieme.de to request a list of your licensed back content books.

Example: Crawling books by book DOI

1. Identify book DOI from provided list (ex. 10.1055/b000000221)
2. Call book page with book DOI (ex. https://www.thieme-connect.com/products/ebooks/book/10.1055/b000000221) and find metatag chapter_list_url with link to book interface
3. Call book interface (see 2.2.1 Book Interface)

1.2.1 Book Page

URL	https://www.thieme-connect.com/products/ebooks/book/[Book DOI]

The book page can be accessed via the URL above with the book DOI inserted at the end.

The book page provides a link to the book interface through the following metatag:

Metatag	<meta name="chapter_list_url" content="[URL]">

The content attribute contains a link to the book interface (see 2.2.1 Book Interface).

1.2.2 Chapter Page

URL	https://www.thieme-connect.com/products/ebooks/lookinside/[Chapter DOI]

The chapter page can be accessed via the URL above with the chapter DOI inserted at the end.

The chapter page provides a link to the chapter interface through the following metatag:

Metatag	<meta name="chapter_crawling_links_url" content="[URL]">

The content attribute contains a link to the chapter interface (see 2.2.2 Chapter Interface).

[1] Referring to a VoR publication not yet in an issue.

2 Crawling Interfaces

The crawling interfaces provide XML and/or PDF full text files as well as media elements (where available) for download. The links to these files are provided through XML interfaces described in the following chapter.

Please make sure to limit your crawling activity to 1 request / second and to include the session cookie with every request.

2.1 Journal Crawling Interfaces

To crawl journal full texts and media, two interfaces are provided:

- The issue interface allows for downloading data for all articles within an issue as identified by issue DOI.
- The article interface allows for downloading data for a single article identified by article DOI.

Issues might contain conference abstracts. These are contained on a separate interface page which can be reached by following the links in the conference elements as described in 2.1.1 Issue Interface.

2.1.1 Issue Interface

URL (issue)	https://www.thieme-connect.com/products/ejournals/issue-xml/[Issue DOI]
URL (eFirst[2])	https://www.thieme-connect.com/products/ejournals/issue-xml/eFirst/[Journal DOI]

The issue interface can be reached through the meta tag on the issue page (see 1.1.4 Issue Page) or directly via the first URL above with the issue DOI inserted at the end.

The issue interface for eFirst articles can be reached via the second URL above with the journal DOI inserted at the end. If a journal has no eFirst articles, this URL returns an HTTP ERROR 404.

The issue interface provides a list of document elements per article within a root documents element. The document elements available per article are described in 2.1.2 Article Interface.

Root Element <documents>

XPath	//documents

The root element documents contains all elements of the issue and article interface listed below. It also links articles to the issue they are contained in via the parent_doi attribute (see attribute list below).

Attribute	Description	Example
parent_doi	Issue DOI for the articles included in the document elements	parent_doi="10.1055/s-011-52611"

Element <conferences>

XPath	//documents/conferences

The conferences element groups all individual conference elements (described below) linking to the different groupings of conference abstracts within an issue.

Element <conference>

XPath	//documents/conferences/conference

The conference element links to the groupings of conference abstracts. A single conference might contain multiple groupings, therefore multiple conference elements might exist.

[2] Referring to a VoR publication not yet in an issue.

To get access to the conference abstract texts, the URLs contained in the url attributes of the conference element need to be called. This might require multiple calls down the hierarchy until no more conference elements are found. The conference abstracts themselves follow the same structure as the article interface (see 2.1.2 Article Interface).

Attribute	Description	Example
url	URL linking to the next hierarchy layer of the conference	url="http://www.thieme-connect.de/products/ejournals/issue-xml/10.1055/s-012-52706/grouping/082883/10.1055/s-00000094"

2.1.2 Article Interface

URL	https://www.thieme-connect.com/products/ejournals/article-xml/[Article DOI]

The article interface can be reached through the meta tag on the article page (see 1.1.5 Article Page) or directly via the URL above with the article DOI inserted at the end.

The article interface provides a document element for the article within a root documents element. The root documents element is the same as the documents element on the issue interface (see 2.1.1 Issue Interface).

Element <document>

XPath	//documents/document

The element document contains all elements of the article interface belonging to the same article as identified in the doi attribute (see attribute list below).

The other attributes can be used to selectively crawl only articles of a given article or publication type.

Attribute	Description	Example
doi	Article DOI	doi="10.1055/a-1589-7568"
article_type	Article types to filter by: • congress-abstract • erratum • evaluation • magazine • promotional • scientific	article_type="scientific"
publication_type	Publication types to filter by: • efirst • issue • am • continuous-publication	publication_type="issue"

Element <abstract>

XPath	//documents/document/abstract

The abstract element links to the abstract XML in its url element.

This element is only available for congress abstracts as some older congresses are only accessible via this element. All other articles contain the abstract in their full text XML as defined by the respective DTD.

Attribute	Description	Example

url	URL linking to abstract XML	url="http://www.thieme-connect.de/products/ejournals/xml/10.1055/s-0041-1740682.xml"
format_fulltext	Format of the abstract XML (compare Appendix: XML Schemas / DTDs)	format_fulltext="congressdtd"

Element <fulltext>

XPath	//documents/document/fulltext

The fulltext element groups the available full texts for an article. This might be a full text XML, a full text PDF, or both (see next two elements).

Element <fulltext_xml>

XPath	//documents/document/fulltext/fulltext_xml

The fulltext_xml element contains a link to the full text XML in its url attribute.

The DTD used for this XML is specified in the format_fulltext attribute. This allows for selectively crawling texts of a specific DTD and helps in matching the DTDs described in Appendix: XML Schemas / DTDs to the full text files.

Attribute	Description	Example
url	URL linking to full text XML	url="http://www.thieme-connect.de/products/ejournals/xml/10.1055/s-0041-1740682.xml"
format_fulltext	Format of the full text XML (compare Appendix: XML Schemas / DTDs)	format_fulltext="jats1.3"

Element <fulltext_pdf>

XPath	//documents/document/fulltext/fulltext_pdf

The fulltext_pdf element contains a link to the full text PDF in its url attribute.

Attribute	Description	Example
url	URL linking to the full text PDF	url="http://www.thieme-connect.de/products/ejournals/pdf/10.1055/a-1712-0389.pdf"

Element <graphical_abstract>

XPath	//documents/document/graphical_abstract

The graphical_abstract element contains a link to the graphical abstract image in its url attribute.

Attribute	Description	Example
url	URL linking to the graphical abstract	url="http://www.thieme-connect.de/bilder/synthesis/200723/e184_ga"

Element <media>

XPath	//documents/document/media

The media element contains the media linked to an article in its child media_element elements. If an article has no media attached, this element does not exist.

Element <media_element>

XPath	//documents/document/media/media_element

The media_element element contains a link to the media element in its url attribute.

The media type is specified in the type attribute to allow for selective downloads of specific media types.

Important: Video, audio, and podcast files can be large and thus create a lot of traffic both on the providing and the receiving end. We request you consider whether you really need to archive them before including them in your crawling process.

Attribute	Description	Example
type	Media type to filter by: • Image • Video • Audio • Podcast • Look inside • Supplementary material	type="Image"
url	URL linking to the media element	url="https://thieme-connect.de/media/ains/202201/10-1055-a-1712-0430-iai01.jpg"

2.2 Book Crawling Structure

To crawl book full texts and media, two interfaces are provided:

- The book interface allows for downloading data for the whole book (identified by its book DOI) as well as all chapters within.
- The chapter interface allows for downloading data for a single chapter (identified by its chapter DOI).

If you need to crawl book back content, please contact export-ejournals@thieme.de to request a list of your licensed back content books and their respective DOIs.

2.2.1 Book Interface

URL	https://www.thieme-connect.com/products/ebooks/book-xml/[Book DOI]

The book interface can be reached through the meta tag on the book page (see 1.2.1 Book Page) or directly via the URL above with the book DOI inserted at the end.

The book interface provides the book full text in XML and PDF format (if available), and a list of document elements for the book's chapters within a root documents element. The document elements available per chapter are described in 2.2.2 Chapter Interface.

Root Element <documents>

XPath	//documents

The root element documents contains all elements of the book and chapter interface listed below. It also links chapter to the book they are contained in via the parent_doi attribute (see attribute list below).

Attribute	Description	Example
parent_doi	Book DOI for the chapters included in the document elements	parent_doi="10.1055/b000000239"

Element <fulltext_xml_book>

XPath	//documents/fulltext_xml_book

The fulltext_xml_book element contains a link to the full text XML of the entire book in its url attribute. To crawl book XMLs by chapter, reference 2.2.2 Chapter Interface.

The DTD used for this XML is specified in the format_fulltext attribute. All books are marked as “kisdtd”. Different KIS DTD version can be differentiated from the XML content. For more details, see A.2 Book DTDs.

Attribute	Description	Example
url	URL linking to full text XML	url="http://www.thieme-connect.de/products/ebooks/bookxml/10.1055/b000000239.xml"
format_fulltext	Format of the full text XML (compare Appendix: XML Schemas / DTDs)	format_fulltext="kisdtd"

Element <fulltext_pdf_book>

XPath	//documents/fulltext_pdf_book

The fulltext_pdf element contains a link to the full text PDF of the entire book in its url attribute. To crawl book PDFs by chapter, reference 2.2.2 Chapter Interface.

Attribute	Description	Example
url	URL linking to the full text PDF	url="http://www.thieme-connect.de /products/ebooks/bookpdf/10.1055/b000000239.pdf"

2.2.2 Chapter Interface

URL	https://www.thieme-connect.com/products/ebooks/chapter-xml/[Article DOI]

The chapter interface can be reached through the meta tag on the chapter page (see 1.2.2 Chapter Page) or directly via the URL above with the chapter DOI inserted at the end.

The chapter interface provides a document element for the chapter within a root documents element. The root documents element is the same as the documents element on the book interface (see 2.2.1 Book Interface). It also contains the fulltext_xml_book and fulltext_pdf_book elements described above linking to the full text of the entire book.

Element <document>

XPath	//documents/document

The element document contains all elements of the chapter interface belonging to the same chapter as identified in the doi attribute (see attribute list below).

Attribute	Description	Example
doi	Article DOI	doi="10.1055/b-0043-193641"

Element <fulltext>

XPath	//documents/document/fulltext

The fulltext element groups the available full texts for a chapter. This might be a full text XML, a full text PDF, or both (see next two elements).

Element <fulltext_xml>

XPath	//documents/document/fulltext/fulltext_xml

The fulltext_xml element contains a link to the full text XML of the chapter in its url attribute.

The DTD used for this XML is specified in the format_fulltext attribute. All books are marked as “kisdtd”. Different KIS DTD version can be differentiated from the XML content. For more details, see A.2 Book DTDs.

Attribute	Description	Example
url	URL linking to full text XML	url="http://www.thieme-connect.de/products/ebooks/chapterxml/10.1055/b-0043-193641.xml"
format_fulltext	Format of the full text XML (compare Appendix: XML Schemas / DTDs)	format_fulltext="kisdtd"

Element <fulltext_pdf>

XPath	//documents/document/fulltext/fulltext_pdf

The fulltext_pdf element contains a link to the full text PDF of the chapter in its url attribute.

Attribute	Description	Example
url	URL linking to the full text PDF	url="http://www.thieme-connect.de/products/ebooks/chapterpdf/10.1055/b-0043-193641.pdf"

Element <media>

XPath	//documents/document/media

The media element contains the media linked to a chapter in its child media_element elements. If a chapter has no media attached, this element does not exist.

To crawl all media for a book, the media elements of all document elements need to be crawled. There is no separate section containing all media.

Element <media_element>

XPath	//documents/document/media/media_element

The media_element element contains a link to the media element in its url attribute.

The media type is specified in the type attribute to allow for selective downloads of specific media types.

Important: Video, audio, and podcast files can be large and thus create a lot of traffic both on the providing and the receiving end. We request you consider whether you really need to archive them before including them in your crawling process.

Attribute	Description	Example
type	Media type to filter by: • Image • Video • Audio • Podcast • Look inside • Supplementary material	type="Image"
url	URL linking to the media element	url="https://thieme-connect.de/media/10.1055-b000000239/lookinside/10-

	1055_b000000239_bookfrontmatter-1.jpg"

Appendix: XML Schemas / DTDs

To allow you to parse the crawled full text and abstract XMLs, Thieme provides the schemas / DTDs used. This appendix is intended to help matching the correct schema / DTD version to the XML files.

A.1 Journal Article Schemas / DTDs

Journal articles are marked with the format_fulltext attribute on both the abstract and fulltext_xml elements to help identify the schema / DTD they are falling under. This allows for filtering and grouping the articles during the crawling process.

JATS 1.3

In use since July 2020

format_fulltext	jats1.3

Schema 2.0

In use mainly 2016-2020

format_fulltext	schema2.0

Thieme DTD

In use prior to 2016

format_fulltext	thiemedtd

Chemistry DTD

In use prior to 2016

format_fulltext	chemistrydtd

Congress DTD

In use prior to 2016

format_fulltext	congressdtd

A.2 Book DTDs

format_fulltext	kisdtd

All book and chapter XMLs are marked with the format_fulltext attribute of “kisdtd”. Differentiating between the versions of the KIS DTD outlined below requires checking the book XML file. All chapters of the same book follow the same KIS DTD as the full book.

KIS DTD 3.2

In use since March 3, 2021

KIS DTD 3.1

In use 2020-2021

KIS DTD 3.0

In use 2016-2020

KIS DTD 2.3

In use prior to 2016