Annotation guidelines

FOSSIL dataset · Source: mpilhlt/fossil, docs/guidelines.md at commit f0b90da

Contents

General strategy **2**

Layered model architecture **3**

Data correction **3**

Document segmentation model **4**

- Tags 5
- Analysis 6
  - Start of the document (cover page and header) 6
  - Additional information 8
  - Elements interrupting the document body: headnotes and footnotes 8
  - Tables and figures 11
  - Hidden characters 11
  - Page breaks 11

Reference segmentation model **12**

- Examples 12
- Example from the demo 14

Citation model **15**

- Basic bibliographic elements 15
- Legal and Humanities Scholarship 20
  - Genre of the cited work 20
  - The issuing authority (<authority>) 20
  - Pinpoint citations (<citedRange>) 20
  - Docket and official identifiers 22
  - Case short-name and statute short-title 22
  - German legal commentaries (Kommentare) 22
  - Mentions of legal texts 23
  - Data embedded in a title or URL 24
  - Intra-footnote references 24
  - Citation-signal phrases 25
  - Quotations 26
  - A real-world example: multiple references with multiple comments 26
  - The reference-segmentation examples, annotated 27

Worked example: the full annotation pipeline **29**

- Stage 1: document segmentation model 30
- Stage 2: reference segmentation model 31
- Stage 3: citation (references) model 31

The annotation guidelines are a collection of best practices and indications for annotating training material in the FOSSIL dataset. They complement what is described in the related models in the Grobid documentation.

The official Grobid documentation is written with documents that have a conventional, structured bibliography in mind — a numbered or alphabetical list of references gathered at the end of the text. These guidelines instead focus specifically on legal scholarship and the humanities disciplines that share its citation culture (history, philology, area studies), where the citation apparatus lives almost entirely in footnotes instead.

> **Note**
>
> These guidelines assume that annotators:
>
> - are familiar with the type of documents to be annotated
> - work with the PDF-TEI-Editor — though annotation can always be done with any XML editor instead

The PDF-TEI-Editor has a visual mode in which annotation is done by marking a span of text and picking the matching tag from a chip palette, rather than by typing XML by hand. Even so, annotators must first familiarize themselves with the raw XML markup described in this document, since it is what defines the underlying schema — the chip palette only makes that schema faster to apply; it does not replace understanding it.

General strategy

The way annotations are expressed is not always unique, and different interpretations from different people (with perhaps different experience and background) are expected. For this reason these guidelines are a "living" document and may change over time.

There are different strategies for annotation:

- blind annotations + review: annotators work on the same documents and one reviewer consolidates the different annotations
- double round + review: each annotator works on a portion of the dataset, and then revises other annotators' work; the reviewer checks the final result

The work will be organised in iterations where the team performs the following tasks on a batch of documents:

1. automatic extraction from the source PDF
2. annotation (either blind or double)
3. review, feedback and corrections, update of guidelines if necessary
4. training/evaluation of the model
5. proceed with 1. on a new batch

The input files are PDF documents that are pre-annotated by the ML models. The annotators then have to correct the pre-annotated output from the model.

Layered model architecture

Grobid does not annotate a document in one pass. It works through three separate models, one after the other, each one handed only the part of the document that the previous model marked out for it. First, the document segmentation model looks at the whole document and divides it into broad zones — the header, the body, the footnotes, and so on. Next, the reference segmentation model takes only the zones marked as footnotes and splits that raw text into individual citations. Finally, the citation model takes each of those individual citations and works out its fine-grained structure — who the authors are, what the title is, which year it was published, and so on.

The practical consequence of this pipeline is worth keeping in mind while annotating: each model can only "see" what the model before it correctly identified and handed down. If a piece of text is not tagged correctly at an earlier stage — for example, if a footnote is not marked as such by the segmentation model — it never reaches the later models at all, and no amount of careful annotation further down the pipeline can fix that. This is why getting the earlier, coarser distinctions right (what is a footnote, what is body text, what is a header) matters just as much as getting the later, more detailed ones right.

Data correction

The most important principle when correcting the pre-annotated training data is to keep the stream of text untouched. Only the tags can be moved; the text itself shall not be modified or corrected. The stream of text present in the training file after extraction of the content of the PDF is similar to the stream of text Grobid will have to process once the models are (re)created. It is thus important to have Grobid trained on this real-world input, even if it contains OCR errors, noise, unknown unicode characters, etc. Therefore, **never** correct mistakes in the text — leave everything as it is.

There are three exceptions to this main rule:

1. Actual end-of-lines from the PDF files are indicated by the element `<lb/>`. These tags are not annotation tags, but should be considered part of the stream of text. They therefore must not be moved or removed with respect to the overall text stream.
2. Line breaks and changes of indentation are accepted outside the main tags to make the formatting more readable.

> **Tip**
>
> Example:
>
> ```
> <text>
>     <body>...</body><listBibl>...</listBibl><page>...</page>
> </text>
> ```
>
> can safely be reformatted to
>
> ```
> <text>
>     <body>...</body>
>
>     <listBibl>...</listBibl>
>
>     <page>...</page>
> </text>
> ```

3. In the TEI/XML files, an end-of-line is equivalent to a space character — it is thus possible to add or remove end-of-line characters as long as the spacing is preserved.

   > **Tip**
   >
   > See this example from the Grobid documentation:
   >
   > ```
   > <title level="a">In XML training files, end-of-line and space are the
   >   same</title> <lb/> <author>Kermitt Jr</author> <lb/> <date>2017</date>
   > ```
   >
   > is equivalent to
   >
   > ```
   > <title level="a">In XML training files, end-of-line and space are
   >     the same</title> <lb/>
   >     <author>Kermitt Jr</author>
   >     <lb/>
   >
   >     <date>2017</date>
   > ```

4. Never place a non-whitespace character outside of the direct child elements of `<text>`. All text content must be inside of one of them.

Document segmentation model

> **Note**
>
> Model name in the PDF-TEI-Editor: `grobid.training.segmentation`, training files end in `.training.segmentation.tei.xml`

This section complements what's described in the related model in the Grobid documentation.

The general structure of the files is:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- document metadata -->
  </teiHeader>
  <text>
    <titlePage>
```

```
      Cover page, or pages that are leftover from the journal that are not fully related
  to the article
    </titlePage>

    <div type="toc">
      Table of content
    </div>

    <front>
      Front matter content. This includes:
        - title
        - abstract
        - footnotes containing affiliations and acknowledgements
      line breaks: <lb/>
    1<lb/>
    </front>

    <!-- more <front> tags as needed -->

    <note place="headnote">Headers, if any, such as journal name or edition, author name,
  title of the article from Page 2<lb/></note>

    <body>
      Body content: this is the text of the article / chapter without the footnotes,
  including section headings, illustrations, etc.<lb/>
    </body>

    <listBibl>
      All footnotes that contain reference data<lb/>
    </listBibl>

    <note place="footnote">Any footnotes that do not contain references, or footers
  containing repeating text, similar to the headnote<lb/></note>

    <!-- more <note>, <body> and <listBibl> tags -->

    <note place="footnote">Endnotes, if any, which do not contain references, otherwise
  use listBibl<lb/></note>

    <div type="acknowledgment|availability|funding|annex">
      Any type of annex for acknowledgement, availability, funding or other information
  that is not in the header but attached to the end of the article
    </div>

  </text>
</TEI>
```

Tags

The following TEI elements are used by the segmentation model:

- `<titlePage>` for the cover page
- `<front>` for the document header
- `<note place="headnote">` for the page header
- `<note place="footnote">` for the page footer and numbered footnotes
- `<body>` for the document body
- `<listBibl>` for the bibliographical section
- `<page>` to indicate page numbers
- `<div type="toc">` for table of content

- `<div type="conflict">` for conflict of interest statements (when not placed in the header)
- `<div type="contribution">` for author contribution statements (when not placed in the header)
- `<div type="acknowledgement">` for acknowledgment statement in the annex (note that both British and American spelling are accepted by Grobid)
- `<div type="availability">` for data and code availability statement annex (when not placed in the header)
- `<div type="funding">` for funding information annex (when not placed in the header)
- `<div type="annex">` for any other annexes

These substructures must be identified whenever they interrupt the flow of the `<body>`. Figures and tables (including their potential titles, captions and notes) are considered part of the body, so they are contained by the `<body>` element.

Note that the markup overall follows the TEI conventions.

> **Tip**
> It is highly recommended to study the existing training documents for the segmentation model first, to see examples of how these elements should be used.

Analysis

The following sections provide detailed information and examples on how to handle certain typical cases.

Start of the document (cover page and header)

Cover page (`<titlePage>`)

An optional cover page — usually added by the publisher to summarize the bibliographical and copyright information — might be present, and needs to be entirely identified by the `<titlePage>` element. The cover page is considered an addition to a standalone well-formed article. The content of a cover page is usually redundant with the bibliographical information found in the article header. A cover page usually corresponds to the content of the entire first page.

Header (`<front>`)

The header section typically contains bibliographical information, such as the document's title, author(s) possibly with affiliations, abstract, keywords, container journal title, etc. The header usually covers everything until the start of the article body (e.g. until reaching the introduction of the article). While a cover page is optional, an article should always include a header, even if limited to the title.

> **Note**
> For the segmentation model, there are no `<title>` or `<author>` elements, because they are handled by the `header` model, which runs in cascade at the next stage on the content the segmentation model identified as "header".

All this material should be contained within the `<front>` element. In addition, any footnotes that are referenced from within the header (for example when author affiliations and addresses are expressed in footnotes) should also be annotated under a `<front>` element. Furthermore, the footer that contains

the first page number should also be treated as part of the header, since the first page number is useful, common bibliographical information.

In general, we expect to find all the bibliographical information of the document as part of the header. This principle should be followed in every document in order to ensure homogeneity of the "header" content across the training data.

The following bibliographical-metadata lines, when they appear as a footnote on the first page of the document, should be contained inside a `<front>` element:

- Received: [date]
- Revised: [date]
- Accepted: [date]

However, any footnotes referenced from within the `<body>` should remain outside the header element, even if they are on the first page or surrounded by `<front>` fragments.

There should be as many `<front>` elements as necessary to contain all the content identified as "front content" (bibliographical information), not necessarily limited to the first pages. `<front>` can contain items that are not always at the beginning of the document, such as:

- Copyright information / Open Access licence and statement
- Correspondence information
- Detailed affiliation and address information
- Submission information: when the document was received, approved and published

These items often appear at the very end of an article or just after the document body. However, for consistency, they should be annotated under `<front>` because they are bibliographical information covered by the header model.

The following information blocks sometimes appear inside the article header, so they should be annotated as `<front>`:

- Author contributions
- Ethics and competing interests
- Funding
- Data / code availability statement

However, when they instead appear as annexes after the document body, they should be annotated as `<div type="annex">` (see next section), not under `<front>`.

It is possible that the position of a title in the text flow of a document differs from the visual layout of the document.

The following TEI XML annotation shows the presence of a `<front>` element surrounding both the topic and the title:

```
virus. <lb/>But is the role of LGP2 in CD8 + T cell <lb/>survival and function cell
intrinsic <lb/></body>

<front>A N T I V I R A L I M U N I T Y <lb/>LGP2 rigs CD8 + T cells for
survival <lb/></front>

<body> or extrinsic? T cell receptor-and <lb/>IFNβ-mediated signalling in CD8 + T
```

> **Note**
> In general, whether the `<lb/>` (line break) element is inside or outside `<front>` or other elements is of no importance. However, as indicated in Data correction, the `<lb/>` element should not be removed and should follow the stream of text.

Sometimes an article starts mid-page, with the end of the preceding one occupying the upper first third of the page. As this content does not belong to the article in question, do not add any elements, and remove any `<front>` or `<body>` elements that could appear in the preceding article.

Additional information `<div type="...">`

Additional and supporting information sections, which are located **after the body** of the article (typically after the conclusion), should be annotated under `<div type="annex">` or the following more specific annex types:

- `<div type="acknowledgment">` for acknowledgment annex (including funding/grant acknowledgement when inside an acknowledgement section)
- `<div type="availability">` for data and code availability statement annex
- `<div type="funding">` for funding information annex

> **Note**
> Different annex-type sections should be segmented into separate `<div type="annex">` elements to capture the start and end of each section block.

Supplementary texts, supplementary figures and tables, and any similar appendices should all be encoded under `<div type="annex">`.

Elements interrupting the document body: headnotes and footnotes

Any information appearing in the page header needs to be surrounded by a `<note place="headnote">`.


Vol 443|12 October 2006|doi:10.1038/nature05123 nature

LETTERS

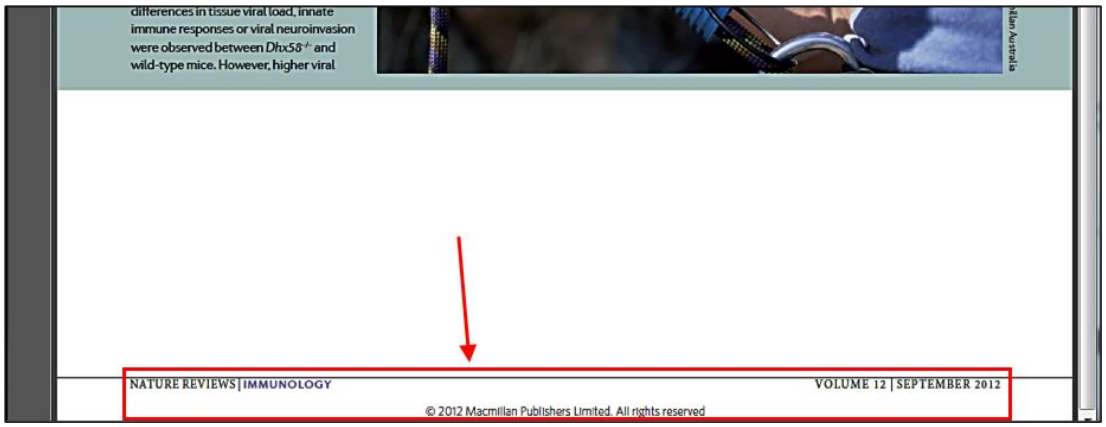
Fast neurotransmitter release triggered by Ca influx through AMPA-type glutamate receptors

Andrés E. Chávez[1], Joshua H. Singer[1]† & Jeffrey S. Diamond[1]

Example of a headnote

The contents of the grey band in the screenshot above should be surrounded by a `<note place="headnote">`, except on the first page, where this type of information would be inside the `<front>` element.

Any information appearing in the page footer needs to be put inside a `<note place="footnote">`, as shown in the following example:



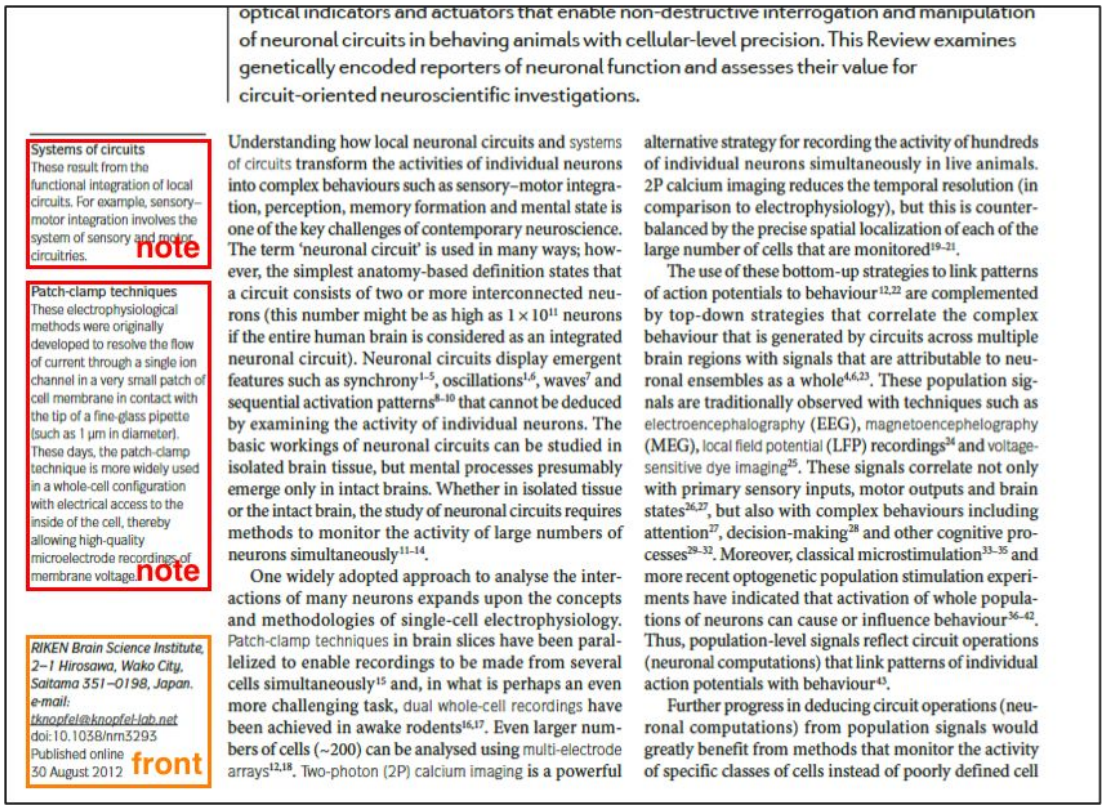
Example of a footnote

Corresponding TEI XML:

```
<note place="footnote">NATURE REVIEWS | IMMUNOLOGY <lb/>VOLUME 12 |
    SEPTEMBER 2012 <lb/>© 2012 Macmillan Publishers Limited. All rights reserved</note>
```

The `<page>` element, which contains the page number, should be outside any of the above `<note>` elements.

Any notes to the left of the main body text are to be encoded as `<note>` if they are related to an element of the `<body>`; if they concern header elements, they go into a `<front>` element. See this screenshot as an example:



Example of different note types

References are also expected to appear in footnotes, in four different styles:

1. Header information, such as affiliation or publication information.

   > Prof. Dr. Eric Steinhauer*

> **Bibliotheksrecht als Zugangsrecht und die Herausforderungen der Digitalisierung**
>
> …
>
> * Der Verfasser ist stellvertretender Direktor der Universitätsbibliothek Hagen und Honorarprofessor am Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin.
>
> 1 Vgl.: *Gantert*, Bibliothekarisches Grundwissen, 9. Aufl. 2016, S. 3.
>
> 2 Vgl.: *Genge*, Art. „Sekretierte Bücher“, in: Corsten/Füssel/Pflug, Lexikon des gesamten Buchwesens, Bd. 7, 2. Aufl. 2007, S. 49; *Kellner*, Der „Giftschrank“, 2002, S. 1 ff.

The `*` note should be annotated as `<front>...</front>` because it is part of the bibliographic information header.

> **The Power of Teaching Police through the Prism of Human Rights**[1]
>
> *Sean Molloy*[2]
>
> …
>
> [1] The author wishes to think Professor Colin Murray, Professor Chris Ashford, Dr Conall Mallory, and Dr Helene Tyrell for comments on earlier drafts. Any errors that remain are my sole responsibility.
>
> [2] Sean Molloy is a NUAcT Fellow in Law and Newcastle Law School, Newcastle University.

Both references 1 and 2 are affiliation and acknowledgment, and should be annotated as `<front>...</front>`.

2. A pure footnote with a comment related to content in the text, annotated as `<note place="footnote">.....</note>`.

> [4] Feinstein, S., & Wood, R. (1995). *History of law-related education* at 10. (ED401163). ERIC. https://perma.cc/KCL2-8VFM
>
> [5] *Id.*
>
> [6] American Bar Association Division for Public Education. (n.d.). *Isidore Starr Award for Excellence in Law-Related Education.* https://perma.cc/4SZ2-47A9
>
> [7] Feinstein & Wood, *supra* note 3 at 12.
>
> [8] I have prepared a Timeline of Juvenile Law and Related Topics at https://civicsguide.com/chronology/.
>
> [9] Arbetman, L. (2018). Street Law, Inc.: context, history and future. *International Journal of Public Legal Education*, 2(1), https://perma.cc/P8RH-2E4D

"see reference 8" should be annotated as `<note place="footnote">`.

3. A pure reference, with the data of the referred article, annotated as `<listBibl>....</listBibl>`.

   Example: footnotes 4, 5, 6, 7 and 9 in the screenshot above should be annotated as `<listBibl>`.

4. Mixed content, with a comment related to the body and a reference annexed to it, annotated as `<listBibl>....</listBibl>` as well (see footnote 9 in the same screenshot). See Reference segmentation model below for how the commentary and the citation(s) it introduces are annotated together.

> **Warning**
>
> - Due to the way text is extracted from the PDF, `<note place="headnote">`, `<note place="footnote">`, and `<page>` elements are not necessarily where you expect them — sometimes they are all placed at the bottom of the page. Just annotate, do not move the elements around.
> - The page text extraction algorithm also sometimes strips the numbering of the headers and places it elsewhere. Do not correct this either.

Tables and figures

Figures and tables belong to the main body structure: they are not specifically encoded at the segmentation level.

Figures and tables, including captions, appearing after the references but related to the body (e.g. a list of figures in preprints), should be under `<body>`. If a figure or table appears inside an annex of an article, it should remain inside the `<div type="annex">` element. If a figure or table appears in an abstract (which is rare but might happen), it should remain within the `<front>` element.

Hidden characters

It happens that Grobid picks up hidden text that is present but not visible on the PDF's page for the reader. Such content should not be surrounded by any element, to indicate to Grobid that it should be ignored.

```
visible in lane 10 (longer exposure), where anti-rabbit secondary antibodies<lb/> were
  used. <lb/></body>

print ncb1110 17/3/04 2:58 PM Page 309 <lb/>

<note place="footnote">© 2004 Nature Publishing Group <lb/></note>
```

Page breaks

The PDF-TEI-Editor marks the start of every PDF page with a `<pb n="..."/>` element, where `n` is the page number in the PDF (counted from 1, not the number printed on the page). It is placed before everything on the new page: if the page starts with a new element (for example `<page>` with the page number, or a headnote), the `<pb/>` stands before that element, directly inside `<text>`. If the page starts in the middle of an element (for example body text that continues from the previous page), it stands inside that element before the first word of the page. GROBID ignores `<pb/>` during training. Like `<lb/>`, it records the layout of the PDF and is not an annotation: do not move or delete it, and leave it where it is when moving the boundaries of an element.

```
continued on the next page. <lb/></body>
<pb n="2"/>
<page>2 <lb/></page>
<note place="headnote">Doe and Roe, A Sample Article <lb/></note>
```

A page that starts in the middle of the body:

```
<body>... the argument is <lb/><pb n="3"/>continued here. <lb/></body>
```

Reference segmentation model

> **Note**
> Model name in the PDF-TEI-Editor: `grobid.training.references.referenceSegmenter`, training files end in `.training.references.referenceSegmenter.tei.xml`

The reference segmenter is a model that takes the content of the `<listBibl>` label produced by the segmentation model and does a more fine-grained segmentation, which can then be parsed by the final citation (`references`) model.

The reference segmenter model is trained with two labels only:

- `<label>` — the reference number, e.g. "[1]" in a numbered bibliography, or the footnote number (as in "^1^ This is the first footnote.")
- `<bibl>` — encloses an individual reference
- `<bibl type="footnote">` — encloses an individual comment or note that does not refer to any reference. This is to catch incorrect classification in the upstream "segmentation" annotation.

For footnote references, we use one `<bibl>` element per reference. The `references` model (see Citation model below) then parses the content of each `<bibl>` and extracts the reference metadata.

If there is commentary (i.e. content that is not bibliographic metadata), including introductory signal words ("See also", "Vgl. z.B.", "Anderer Meinung:", etc.), include it with the `<bibl>` it belongs to semantically. The comment will mostly precede but sometimes also follow the reference (see below for examples). You need to look closely; if in doubt, ask.

> **Note**
> The `references` model has no notion of a labeled span sitting *between* two `<bibl>`s. When one comment or signal phrase introduces several references, it is annotated only once, inside the `<bibl>` it is closest to (see Citation model for how that same convention carries through into `<seg type="citationContext">` at the next stage).

Examples

Here are some examples that we will expand with more cases as needed.

Several references in one footnote

> [2] Ronald H. Coase, *The Problem of Social Cost*, 56 J.L & ECON 837 (1960); GUIDO CALABRESI, THE COSTS OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS (1970); RICHARD A. POSNER, THE ECONOMICS OF JUSTICE (1981); RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW (9th ed. 2014); ROBERT COOTER & THOMAS ULEN, LAW & ECONOMICS (2000); ROBIN P. MALLOY, LAW AND MARKET ECONOMY: REINTERPRETING THE VALUES OF LAW AND ECONOMICS (2000); NICHOLAS MERCURO & STEVEN G. MEDEMA, ECONOMICS AND THE LAW: FROM POSNER TO POST-MODERNISM AND BEYOND (2d ed. 2006).

```
<bibl><label>2</label> Ronald H. Coase, The Problem of Social Cost, 56 J.L &amp;amp; ECON 837
  (1960); </bibl>
<bibl>GUIDO CALABRESI, THE COSTS<lb/> OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS (1970);
  </bibl>
<bibl>RICHARD A. POSNER, THE ECONOMICS OF JUSTICE<lb/> (1981); </bibl>
<bibl>RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW (9th ed. 2014); </bibl>
<bibl>ROBERT COOTER &amp;amp; THOMAS<lb/> ULEN, LAW &amp;amp; ECONOMICS (2000); </bibl>
<bibl>ROBIN P. MALLOY, LAW AND MARKET ECONOMY: REINTERPRETING THE<lb/> VALUES OF LAW AND
  ECONOMICS (2000); </bibl>
<bibl>NICHOLAS MERCURO &amp;amp; STEVEN G. MEDEMA, ECONOMICS AND THE<lb/> LAW: FROM POSNER TO
  POST-MODERNISM AND BEYOND (2d ed. 2006).<lb/> </bibl>
```

German legal citation

> 2 Dazu etwa: *Hofmann*, ZGE 2016, 482, 498; *Becker*, ZGE 2016, 239, 273; *Raue*, ZGE 2014, 387, 389.

```
<bibl><label>2</label> Dazu etwa: Hofmann, ZGE 2016, 482, 498; </bibl>
<bibl>Becker, ZGE 2016, 239, 273; </bibl>
<bibl>Raue, ZGE 2014,<lb/> 387, 389.<lb/> </bibl>
```

Introductory comment

> 3 Zur Frage, ob der Erwerb eines (digitalen) Werkexemplars mit dem Erwerb eines dinglichen Genussrechts verbunden ist, vgl.: *Kuschel*, Der Erwerb digitaler Werkexemplare zur privaten Nutzung, 2019.

```
<bibl><label>3</label> Zur Frage, ob der Erwerb eines (digitalen) Werkexemplars mit dem
  Erwerb eines dinglichen<lb/> Genussrechts verbunden ist, vgl.: Kuschel, Der Erwerb
  digitaler Werkexemplare zur privaten<lb/> Nutzung, 2019.<lb/> </bibl>
```

> [15] Arguing that national corporate due diligence laws potentially breach the principle of consent in international law and the sovereignty of host States, and perpetuate power imbalances of colonial derivation, see e.g.: C. OMARI LICHUMA, *(Laws) Made in the 'First World'*, cit., pp. 517-518; F. DEHBI, O. MARTIN-ORTEGA, *An integrated approach to corporate due diligence from a human rights, environmental, and TWAIL perspective*, in *Regulation & Governance*, 2023, n. 17, pp. 927-943, in particular pp. 932-935. In contrast, affirming that such national legislations are to be regarded as instruments facilitating home States' compliance with their international obligations, rather than as breaches of the host States' sovereignty, see: A. BONFANTI, *Imprese multinazionali, diritti umani e ambiente. Profili di diritto internazionale pubblico e privato*, Milano, 2012, p. 136.

```
<bibl><label>15</label> Arguing that national corporate due diligence laws potentially
  breach the principle of consent in<lb/> international law and the sovereignty of host
  States, and perpetuate power imbalances of colonial<lb/> derivation, see e.g.: C. OMARI
  LICHUMA, (Laws) Made in the 'First World', cit., pp. 517-518; </bibl>
<bibl>F. DEHBI, O.<lb/> MARTIN-ORTEGA, An integrated approach to corporate due diligence
  from a human rights, environmental,<lb/> and TWAIL perspective, in Regulation &amp;amp;
  Governance, 2023, n. 17, pp. 927-943, in particular pp. 932-935.<lb/> </bibl>
<bibl>In contrast, affirming that such national legislations are to be regarded as
  instruments facilitating home<lb/> States' compliance with their international
  obligations, rather than as breaches of the host States'<lb/> sovereignty, see: A.
  BONFANTI, Imprese multinazionali, diritti umani e ambiente. Profili di diritto<lb/>
  internazionale pubblico e privato, Milano, 2012, p. 136.<lb/> </bibl>
```

Here, the signal phrase ("Arguing that..., see e.g.:") semantically belongs to the first and second references, but this cannot be expressed at the referenceSegmenter stage, so it is put into the first

`<bibl>` only. (At the `references` stage, this same phrase is what becomes `<seg type="citationContext">` — see Citation-signal phrases below.)

Trailing comment

> [9] See e.g.: Inter-American Court of Human Rights, judgment of 16 February 2020, *Indigenous Communities of the Lhaka Honhat (Our Land) Association v. Argentina*, par. 254, quoting the *amicus curiae* intervention of the former UN Special Rapporteur on the Right to Food, De Schutter: «Many indigenous peoples understand the right to adequate food as a collective right. They often see subsistence activities such as hunting, fishing and gathering as essential not only to their right to food, but to nurturing their cultures, languages, social life and identity». The cultural value of food is recognized – with respect to indigenous peoples in particular – also under art. 27 of the International Covenant on Civil and Political Rights (ICCPR). See *infra*, section 3.2.

```
<bibl><label>9 </label>See e.g.: Inter-American Court of Human Rights, judgment of 16
  February 2020, Indigenous Communities<lb/> of the Lhaka Honhat (Our Land) Association
  v. Argentina, par. 254, quoting the amicus curiae intervention<lb/> of the former UN
  Special Rapporteur on the Right to Food, De Schutter: «Many indigenous peoples<lb/>
  understand the right to adequate food as a collective right. They often see subsistence
  activities such as<lb/> hunting, fishing and gathering as essential not only to their
  right to food, but to nurturing their cultures,<lb/> languages, social life and
  identity». The cultural value of food is recognized -with respect to indigenous<lb/>
  peoples in particular -also under art. 27 of the International Covenant on Civil and
  Political Rights<lb/> (ICCPR). See infra, section 3.2.<lb/> </bibl>
```

Court decisions

> [6] See, *ex multis*: African Commission on Human and Peoples Rights, decision of 27 October 2001, *Social and Economic Rights Action Center (SERAC) and Center for Economic and Social Rights (CESR) v. Nigeria*; Inter-American Court of Human Rights, judgment of 27 June 2012, *Case of the Kichwa Indigenous People of Sarayaku v. Ecuador*.

```
<bibl><label>6</label> See, ex multis: African Commission on Human and Peoples Rights,
  decision of 27 October 2001, Social<lb/> and Economic Rights Action Center (SERAC) and
  Center for Economic and Social Rights (CESR) v.<lb/> Nigeria; </bibl>
<bibl>Inter-American Court of Human Rights, judgment of 27 June 2012, Case of the Kichwa
  Indigenous<lb/> People of Sarayaku v. Ecuador.<lb/> </bibl>
```

References to other references (`id.`, `op. cit.`, `a.a.O.`, `ders.`, *supra*, *infra*, "N. 22", "above") are, at the referenceSegmenter stage, simply part of the `<bibl>` they occur in — no separate label exists for them here. They are only broken out into their own `<ref type="...">` elements at the `references` stage; see Intra-footnote references below.

Example from the demo

Sometimes the references are in separate sentences rather than separated by semicolons:

> [1] DANIEL J. BOORSTIN, THE MYSTERIOUS SCIENCE OF THE LAW, at vii (Peter Smith 1973) (1941); *see also* Albert W. Alschuler, *Rediscovering Blackstone*, 145 U. PA. L. REV. 1, 2 (1996). The *Commentaries* was hugely popular in the eighteenth and nineteenth centuries. *See* WILFRID PREST, WILLIAM BLACKSTONE 307 (2008).

```
<bibl><label>1</label> DANIEL J. BOORSTIN, THE MYSTERIOUS SCIENCE OF THE LAW, at vii
  (Peter Smith 1973) (1941);<lb/> </bibl>
<bibl>see also Albert W. Alschuler, Rediscovering Blackstone, 145 U. PA. L. REV. 1, 2
  (1996). The<lb/> Commentaries was hugely popular in the eighteenth and nineteenth
  centuries. See WILFRID<lb/> PREST, WILLIAM BLACKSTONE 307 (2008).<lb/> </bibl>
```

Citation model

> **Note**
>
> Model name in the PDF-TEI-Editor: `references`, training files end in `.training.references.tei.xml`

This section complements what's described in the related model in the Grobid documentation. The `references` model takes each `<bibl>` produced by the reference segmenter and labels its internal structure: author names, titles, dates, extents, identifiers, and so on. Every element below (except `<label>` and `<lb/>`) can occur any number of times, in any order, inside a `<bibl>`.

Most of the corpus is annotated with this base vocabulary alone. A smaller but significant part of the corpus — legal scholarship and the humanities disciplines that share its citation culture — needs a further, domain-specific layer on top of it, which is documented separately in Legal and Humanities Scholarship below.

The annotation principle: content is the statement, attributes are the normalization. Each element contains the whole statement as it appears in the source, including any language-specific words or abbreviations that are part of it — "pp. 67-82", "Rn. 5", "2nd ed.", "Zweite, verbesserte Auflage", "8.9.2016". The attributes carry what the statement means in a normalized, machine-readable form: the kind of unit in `@unit`, a page, volume or other position in `@from`/`@to`, an edition or footnote number in `@n`, an ISO date in `@when`. The two layers do different jobs. The element content records *where* a piece of information is in the text and how the source expresses it; the attributes make it comparable across languages and citation styles without re-reading the text.

```
<biblScope unit="page" from="67" to="82">pp. 67-82</biblScope>
<edition n="2">Zweite, verbesserte Auflage</edition>
<date type="decision" when="2016-09-08">8.9.2016</date>
```

`<biblScope>` and `<citedRange>` always carry `@from`/`@to`, even when the content is a bare number. For other elements, content that is already in normalized form — a plain year, a bare edition number — needs no attribute, since the content is its own value. The rules below apply this principle element by element; where they say what stays *outside* an element, it is because that text is not part of the statement the element denotes (punctuation between fields, connecting words, a role marker after a name).

Basic bibliographic elements

Authors and editors. `<author>` and `<editor>` hold a complete sequence of names as running text. Where the source clearly separates given name / surname, it is fine (but not required) to decompose them with `<persName>`, `<forename>`, `<surname>`:

```
<author>J. Smith</author>
<author><persName><forename>J.</forename> <surname>Smith</surname></persName></author>
```

When a reference lists several people or organizations in the same role — several authors, several editors, several institutional co-authors — each one gets its own element, even when the source packs them into a single comma-separated (or "and"/"&"-joined) run of names. This holds for any number of names, so that every person or organization is captured individually rather than the whole run being left as one approximated string. The rule applies to `<author>` and `<editor>` here, and equally to `<orgName>` (see Institutional authorship) and `<authority>` (see The issuing authority) elsewhere in a `<bibl>`:

```
<author>J. Jett</author>, <author>B. Capitanu</author>, <author>D. Kudeki</author>,
<author>T. Cole</author>, <author>Y. Hu</author>, <author>P. Organisciak</author>,
<author>T. Underwood</author>, <author>E. Dickson Koehl</author>,
<author>R. Dubnicek</author>, <author>J. S. Downie</author>

<editor>H. Müller</editor>, <editor>K. Schmidt</editor> (eds.)
```

Titles. `<title level="...">` classifies the title by the kind of work it names:

`@level`	Meaning
`a`	Article or chapter title (analytic)
`j`	Journal title, including the name of a blog
`m`	Monograph, proceedings, book, or thesis title, including the name of a whole website
`s`	Series title
`u`	Unpublished work title

Web sources follow the same split: a website cited as a whole is `level="m"`, a blog is `level="j"`, and a single post or page within either is `level="a"`.

An optional `@key` holds a short resolvable key for a named work (e.g. an abbreviation), so it can later be linked to an external database.

Translated titles. References to works in other languages often give a translation of the title next to the original, typically in square brackets. Annotate the original and the translation as two separate `<title>` elements with the same `@level`, and leave the brackets outside them. Both titles then carry `@xml:lang` with the language code (BCP 47, e.g. `pl`, `en`) of their text. `@xml:lang` is only added when a translated title is present; a title without a translation gets no `@xml:lang`.

```
<bibl><author>J. Chlebny</author>,
  <title level="a" xml:lang="pl">Standardy Rady Europy i Europejskiego Trybunału Praw
  Człowieka w procedurze administracyjnej i sądowoadministracyjnej</title>
  [<title level="a" xml:lang="en">Standards of the Council of Europe and of the European
  Court of Human Rights in Administrative and Administrative Court Procedure</title>],
  [in:] <editor>Z. Kmieciak</editor> (ed.),
  <title level="m" xml:lang="pl">Postępowanie administracyjne w Europie</title>
  [<title level="m" xml:lang="en">Administrative Procedure in Europe</title>],
  <pubPlace>Warszawa</pubPlace>: <publisher>Wolters Kluwer Polska</publisher>
  <date>2010</date>,
  <biblScope unit="page" from="31" to="42">pp. 31-42</biblScope>.</bibl>
```

Dates. `<date type="publication" when="...">` marks the publication date; `@when` holds the machine-readable ISO form and the text content the date as it appears in the source. A bare year on its own (`2019`) is already that ISO form, so no `@when` is needed. As soon as the text is a fuller or non-ISO

date — a day-month-year date, a date with a month name, an abbreviated form like `8.9.2016` — the content is no longer machine-readable on its own, so set `@when` to the converted ISO value. Only omit `@when` on such a date when the source genuinely does not pin down a resolvable date, e.g. "n.d.".

Date ranges. When the cited date spans more than one day — a multi-day court session, a range given as e.g. "1-2.03.2012" — use `@from`/`@to` instead of `@when`, each holding the ISO form of its end of the range, with the text content kept as it appears in the source:

```
<date from="2012-03-01" to="2012-03-02">1-2.03.2012</date>
```

`@from`/`@to` and `@when` are mutually exclusive: `@when` marks a single point in time, `@from`/`@to` a span.

Extent. `<biblScope unit="...">` carries the cited work's own extent — its volume, issue, or overall page range:

`@unit`	Matches
`page`	Full page range of the article
`volume`	Volume number
`issue`	Issue / number

Designators go inside the element. A word or abbreviation that names the unit or kind of a value — `p.`, `pp.`, `S.`, `Vol.`, `Bd.`, `No.`, `n.`, `Heft`, `Rn.`, `§`, `Abs.`, `para`, `ed.`, `Aufl.`, `note`, or the equivalent in any other language — is annotated together with the value it introduces. This applies the annotation principle stated at the start of this section: the element content covers the whole expression as it appears in the source, and the attributes (`@unit`, `@from`/`@to`) hold the normalized, language-independent value, just as `@when` does for **Dates** above:

```
<biblScope unit="volume" from="1" to="1">Vol. 1</biblScope>, <biblScope unit="issue"
  from="2" to="2">No. 2</biblScope>,
<biblScope unit="page" from="67" to="82">pp. 67-82</biblScope>
<citedRange unit="page" from="79" to="79">p. 79</citedRange>
<citedRange unit="margin" from="5" to="5">Rn. 5</citedRange>
<citedRange unit="section" from="64" to="64">para 64</citedRange>
```

There are three reasons for this rule:

- It follows the TEI convention. The TEI Guidelines' own `<biblScope>` and `<edition>` examples include the designator in the element.
- It loses nothing. The bare value can always be derived from content plus attributes, but the extent of the original expression cannot be reconstructed from a value-only annotation.
- Designators are not always unambiguous. `S.` is a page (*Seite*) in `S. 123` but a sentence (*Satz*) in `§ 1 Abs. 2 S. 1`; `ed.` is an edition after a number but an editor after a name. Keeping the designator inside the element records which reading the annotator chose, through `@unit` or the element name.

What stays **outside** the element:

- Punctuation that separates fields — commas, semicolons, brackets. The abbreviation's own period (`pp.`, `Rn.`) belongs to the designator and stays inside.
- Connecting words that do not name a unit, such as "at" in "233, at 240", "in particular", or "in:".
- Role markers after names, such as "(ed.)", "(eds.)", "(Hrsg.)". `<author>` and `<editor>` hold names, and a role marker is not part of the name.

When one designator introduces a list of values, it goes into the first element only:

```
<citedRange unit="margin" from="5" to="5">Rn. 5</citedRange>, <citedRange unit="margin"
  from="7" to="7">7</citedRange>
```

This differs from Grobid's own convention, which leaves page abbreviations outside `<biblScope>`. Converting to that convention is a matter for the export to Grobid's training format, not for annotation; annotators always follow the rule given here.

Always set `@from`/`@to`. Every `<biblScope>` holds its normalized value in `@from` and `@to`. A range gets its two endpoints; a single value repeats the same value in both attributes. This holds even when the content is already a bare number, so that consumers of the data always read the value from the same place. Roman numerals and other non-Arabic forms are normalized to Arabic numbers:

```
<biblScope unit="volume" from="3" to="3">3</biblScope>
<biblScope unit="volume" from="2" to="2">II</biblScope>
<biblScope unit="volume" from="2" to="2">Bd. 2</biblScope>
<biblScope unit="page" from="79" to="79">p. 79</biblScope>
<biblScope unit="volume" from="3" to="5">3–5</biblScope>
<biblScope unit="page" from="517" to="518">pp. 517-518</biblScope>
```

The one exception is a range whose end the text does not give — an article cited by its first page alone, or an open-ended `ff.` (see Pinpoint citations). Set `@from` alone and omit `@to`, since the end point is unknown rather than identical to the start:

```
<biblScope unit="page" from="795">795</biblScope>
```

Pinpoint extent. When the citing text names a more specific location than `<biblScope>`'s own extent — the common "first page, at pinpoint page" pattern, e.g. "233, at 240" — mark that pinpoint with `<citedRange unit="page">`, a sibling of `<biblScope>`, never nested inside it:

```
<biblScope unit="page" from="233" to="250">233–250</biblScope>, at
<citedRange unit="page" from="240" to="240">240</citedRange>
```

`<citedRange>`'s `@unit` list also includes several pinpoint levels specific to legal and humanities citation practice (statute sections, marginal numbers, recitals, and more) — see Pinpoint citations for the full vocabulary.

Publisher and place. `<publisher>` and `<pubPlace>` hold the publisher's name (also used for corporate authors such as web pages) and place of publication.

Edition. `<edition>` marks the complete edition statement, including the generic designator (`ed.`, `edn.`, `Aufl.`, `éd.`) and any qualifiers ("revised", "verbesserte"). Where an edition number can be determined, record it as an Arabic number in `@n`. An edition statement without a number gets no `@n`:

```
<edition n="2">2nd ed.</edition>
<edition n="2">Zweite, verbesserte und vermehrte Auflage</edition>
<edition n="42">42. Ed.</edition>
<edition>rev. ed.</edition>
```

A date that follows the edition statement stays a separate `<date>` next to `<edition>`, not nested inside it (see the commentary examples under German legal commentaries).

Basic identifiers. `<idno type="...">` accepts, among the values also documented under Legal and Humanities Scholarship: `DOI`, `ISSN`, `arXiv`, `report` (technical/institutional report number).

Web pointer. `<ptr type="web" target="...">` holds a URL as text content, excluding prefixes like "URL:" and trailing periods. An identifier that occurs only inside the URL may be tagged in place with a nested `<idno>` (see Data embedded in a title or URL).

If the URL itself contains a DOI (e.g. `https://doi.org/10.1162/coli_a_00345` or `https://dx.doi.org/...`) and the reference does not also give the DOI separately as its own `<idno type="DOI">`, mark up the DOI portion of the URL text with an embedded `<idno type="DOI">`, nested inside the `<ptr type="web">`:

```
<ptr type="web" target="https://doi.org/10.1162/coli_a_00345">https://doi.org/<idno
  type="DOI">10.1162/coli_a_00345</idno></ptr>
```

If a separate, standalone `<idno type="DOI">` is already present elsewhere in the reference, leave the URL's `<ptr type="web">` as plain text instead — do not annotate the same DOI twice.

Miscellaneous note. `<note>` marks any content not covered by another tag.

- `<note type="report">` is used to indicate the type or form of the publication (e.g. "Ph.D. thesis", "Technical Report").
- `<note>` without any attribute is the fallback for any remaining non-bibliographic data, such as interspersed or trailing comments.

Institutional authorship. `<orgName type="...">` names an institution that is not itself the work's issuing authority (see The issuing authority for that case) — for example the university behind a thesis, or a project consortium behind a technical report:

`@type`	Meaning
`institution`	An institution (e.g. a university)
`collaboration`	Project-based collaboration acting as an author group
`department`	A department within an institution
`laboratory`	A laboratory or research group

A non-legal example combining several of these:

```
<bibl><author>J. Smith</author> and <author>A. Doe</author>,
  <title level="a">A Study of Neural Machine Translation</title>,
  <title level="j">Journal of Computational Linguistics</title>
  <biblScope unit="volume" from="45" to="45">45</biblScope>(<biblScope unit="issue"
  from="2" to="2">2</biblScope>),
  <biblScope unit="page" from="123" to="145">123–145</biblScope>
  (<date type="publication">2019</date>).
  <idno type="DOI">10.1162/coli_a_00345</idno>
</bibl>

<bibl><author>M. Müller</author>, <title level="m">Diffusion Models for Generative
  Art</title>
  (Ph.D. thesis, <orgName type="institution">ETH Zürich</orgName>,
  <date type="publication">2022</date>).
  <note type="report">Ph.D. thesis</note>
</bibl>
```

Legal and Humanities Scholarship

Legal scholarship and the humanities disciplines that share its citation culture (history, philology, area studies) rely on the basic vocabulary above plus a further layer: courts and legislatures cited as the "author" of a work, pinpoint citations below the level of a page (a statute section, a marginal number, a recital), elliptical back- and forward-references within the same footnote apparatus (`id.`, `op. cit.`, `a.a.O.`, `ders.`, *supra*, *infra*), and citation-signal phrases (`see`, `cf.`, `vgl.`) whose polarity and force is itself a meaningful research signal. This section documents that layer.

Genre of the cited work

```
<bibl type="legislation">...</bibl>
<bibl type="decision">...</bibl>
```

`legislation` marks a statute citation, `decision` a court ruling / judicial decision citation. `type` is omitted for an ordinary reference, and reserved to `footnote` for a comment that is not a bibliographic reference at all (see Reference segmentation model above).

The issuing authority (`<authority>`)

```
<authority type="court">EuGH</authority>, Urt. v. <date type="decision" when="2016-09-
  08">8.9.2016</date> –
<idno type="docket">C-160/15</idno>, …
```

`<authority type="court"|"legislature">` names the institution responsible for the cited work — a court, a legislature, or (outside the legal domain) a regulator or other issuing body. Use it instead of `<author>` (an institution is not a person) and instead of a bare `<orgName>` (which has no "responsible for this work" role).

`<orgName>` **may** still be nested inside `<authority>` to decompose a compound institutional name, e.g. a court plus its deciding panel:

```
<authority type="court"><orgName>Oberlandesgericht Frankfurt am Main</orgName>, 5.
  Zivilsenat</authority>
```

The `<date type="...">` values `decision` (date a court decision was issued) and `enacted` (date a statute was enacted) are likewise specific to this section — outside it, `<date>` only uses `type="publication"` (which is the default when omitted).

Pinpoint citations (`<citedRange>`)

`<biblScope>`, introduced above, carries the cited work's *own* extent. In legal and humanities citation practice, the citing text often also names a *pinpoint* into that work — a level of granularity `<biblScope>`'s own `@unit` list has no room for, such as a statute section or a court decision's marginal number. `<citedRange unit="...">` carries that pinpoint. If a work has no pagination at all (a decision cited only by marginal number, say), there is simply no `<biblScope unit="page">` — only `<citedRange>`.

`@unit`	Matches
`section`	`§ 19a`, `Art. 5`, `Sec. 2`
`subSection`	`Abs. 2`
`sentence`	`S. 1`, `Satz 1`

`@unit`	Matches
`number`	`Nr. 3`
`letter`	`lit. b`, `Buchst. b`
`margin`	`Tz. 24`, `Rn. 20`, `Rdnr. 7`
`recital`	`ErwGr. 21` (EU recitals)
`page`	an in-text pinpoint page, e.g. the `(240)` in `233 (240)`

```
<bibl type="legislation">
  <citedRange unit="section" from="19a" to="19a">§ 19a</citedRange>
  <citedRange unit="subSection" from="2" to="2">Abs. 2</citedRange>
  <title level="m" type="legislation" key="UrhG">UrhG</title>
</bibl>
```

Always set `@from`/`@to` — the same rules as for `<biblScope>` above, including keeping the designator (`p.`, `S.`, `Rn.`, `§`, `Abs.`, `lit.`) inside the tag (see Designators go inside the element). A single pinpoint repeats its value in `@from` and `@to`, also when the content is a bare number; a range gets its two endpoints. Values that are not plain numbers, such as `19a`, are recorded as they stand:

```
<citedRange unit="page" from="240" to="240">240</citedRange>
<citedRange unit="page" from="240" to="240">S. 240</citedRange>
<citedRange unit="margin" from="12" to="12">Rn. 12</citedRange>
<citedRange unit="section" from="19a" to="19a">§ 19a</citedRange>
<citedRange unit="page" from="932" to="935">pp. 932-935</citedRange>
```

Only an open-ended pinpoint (`ff.`, `et seqq.`, covered below) sets `@from` alone.

Telling an item's own extent apart from a pinpoint. A single page number or a small page range (`p. 5`, `pp. 23–27`) has exactly the same shape whether it names the *analytic item's own extent* — a short chapter, a commentary or dictionary entry — or a *pinpoint* into a longer work already cited (or about to be cited) by a larger extent; the two cannot be told apart by form alone. Decide by what the number actually bounds: if it is the full extent of the analytic item (`<title level="a">`) inside its monographic container (`<title level="m">`), annotate it as `<biblScope>`; if it locates a passage inside a work whose own extent is given separately, annotate it as `<citedRange>`. An open-ended continuation marker — `f.`/`ff.` or `et seq.`/`et seqq.`, covered next — is a reliable sign of the second case: a work's own extent has a definite end, so no one describes it as "p. 233 ff."; that phrasing only makes sense when pointing into a passage whose end the citing author hasn't bothered to name. Annotate any `f.`/`ff.`/`et seq.`/`et seqq.` pinpoint as `<citedRange>`, never as `<biblScope>`. NOTE: Since chapters are usually longer than 1-4 pages, a general heuristic is that **a small page range in most cases is a pinpoint citation and should be annotated as `<citedRange>`, with the exception of encyclopedia or dictionary entries.**

Open-ended continuation pinpoints (`f.`, `ff.`, `et seq.`, `et seqq.`). The citing text sometimes marks a pinpoint as continuing onto subsequent pages or sections instead of naming an explicit end point, using the German abbreviations `f.`/`ff.` or the Latin `et seq.`/`et seqq.`. These differ in exactly how many further units they claim:

- **`f.`** (German, singular) and **`et seq.`** (Latin, *et sequens* — "and the following [one]") both claim exactly *one* further unit. Treat this as an explicit range: set `@from` to the cited unit and `@to` to the immediately following one.

- **`ff.`** (German, plural) and **`et seqq.`** (Latin, *et sequentia* — "and the following [ones]") claim an unspecified, open-ended number of further units. Since the end point is not actually given, set only `@from` and omit `@to`.

This applies to whichever `@unit` the pinpoint is in (page, margin number, section, etc.), not just `page`:

```
<citedRange unit="page" from="233" to="234">233 f.</citedRange>
<citedRange unit="page" from="233" to="234">233 et seq.</citedRange>
<citedRange unit="page" from="233">233 ff.</citedRange>
<citedRange unit="page" from="233">233 et seqq.</citedRange>
<citedRange unit="margin" from="12" to="13">Rn. 12 f.</citedRange>
<citedRange unit="margin" from="12">Rn. 12 ff.</citedRange>
```

Docket and official identifiers

```
<idno type="docket">C-160/15</idno>
<idno type="ECLI">ECLI:EU:C:2016:644</idno>
<idno type="CELEX">32019L0790</idno>
<idno type="actNumber">2019/790</idno>
```

`actNumber` is a legislative act's own official year/number designation (e.g. the "2019/790" in "Directive (EU) 2019/790"). It is not a volume or issue: no periodical is involved. These are the legal-specific values of `<idno type="...">`, alongside the general ones listed under Basic bibliographic elements.

Case short-name and statute short-title

```
<title level="a" type="caseName">GS Media/Sanoma</title>
<title level="m" type="legislation" key="UrhG">Urhebergesetz</title>
```

These are the legal-specific presets of `<title>`, on top of the plain `@level` values listed under Basic bibliographic elements. Tagging the case short-name explicitly, rather than leaving it as trailing free text, is what stops the model from treating `– GS Media/Sanoma` as the start of a new reference.

German legal commentaries (Kommentare)

A commentary (e.g. *BeckOK UrhR*, *Dreier/Schulze*) is organized by the provisions of the statute it explains. A heading such as `UrhG § 60c` therefore names an *entry within the commentary*, not the statute itself. Annotate it as an analytic title (`<title level="a">`), and the commentary as the monograph-level title (`<title level="m">`). The marginal number into the entry is `<citedRange unit="margin">`. The `<bibl>` gets no `type="legislation"`, and the heading is not split into `<citedRange unit="section">` and a statute title.

German legal scholarship cites commentaries in a standardized short form in footnotes (the full citation is reserved for the bibliography). Two names can appear:

- **The commentary's short title.** Many commentaries are named after their founding editors (`Dreier/Schulze`, `Grüneberg`). These names are part of the title, not authors, and are annotated with it as `<title level="m">`.
- **The commentator** (*Bearbeiter*) who wrote the cited entry. This is the only person annotated as `<author>`. The commentator is optional and often omitted.

Two forms are common, where "Rn." (*Randnummer*) is the marginal number:

Form	Example
`Author, in: Short title, § X Rn. Y`	`Westermann, in: MünchKomm-BGB, § 985 Rn. 12`
`Short title/Author, § X Rn. Y`	`MünchKomm-BGB/Westermann, § 985 Rn. 12`

In the slash form, the commentator is the name after the *last* slash; everything before it is the short title (`Dreier/Schulze/Dreier` is the commentary *Dreier/Schulze*, entry by Dreier). In both forms, the commentator is `<author>` and the short title is `<title level="m">`; the slash, "in:" and commas stay outside the tags. When the statute is already part of the commentary's title (`MünchKomm-BGB`), the entry heading is just `§ 985`, still annotated as `<title level="a">`.

```
<bibl><author>Westermann</author>, in: <title level="m">MünchKomm-BGB</title>,
  <title level="a">§ 985</title> <citedRange unit="margin" from="12" to="12">Rn.
  12</citedRange></bibl>

<bibl><title level="m">MünchKomm-BGB</title>/<author>Westermann</author>,
  <title level="a">§ 985</title> <citedRange unit="margin" from="12" to="12">Rn.
  12</citedRange></bibl>

<bibl><author>Spindler</author>, in: <title level="m">BeckOK-BGB</title>,
  <title level="a">§ 823</title> <citedRange unit="margin" from="45" to="45">Rn.
  45</citedRange></bibl>

<bibl><title level="m">BeckOK UrhR</title>/<author>Grübler</author>,
  <edition n="42">42. Ed.</edition> <date type="publication" when="2024-05-
  01">1.5.2024</date>,
  <title level="a">UrhG § 60c</title> <citedRange unit="margin" from="5" to="5">Rn.
  5</citedRange>;</bibl>

<bibl><title level="m">Dreier/Schulze</title>/<author>Dreier</author>,
  <edition n="7">7. Aufl.</edition> <date type="publication">2022</date>,
  <title level="a">UrhG § 60c</title> <citedRange unit="margin" from="1" to="1">Rn.
  1</citedRange>.</bibl>
```

Single-volume commentaries named after their editor (e.g. *Grüneberg*, formerly *Palandt*) are often cited without a commentator, e.g. `Grüneberg, BGB, § 280 Rn. 8`. Here the name is part of the commentary's title, not an `<author>`:

```
<bibl><title level="m">Grüneberg, BGB</title>, <title level="a">§ 280</title>
  <citedRange unit="margin" from="8" to="8">Rn. 8</citedRange></bibl>
```

Mentions of legal texts

Annotate a statute or other legal text only when it is itself the source cited for a statement, i.e. when the `<bibl>` refers to it. A legal text that is merely mentioned — inside a citation-signal phrase or inside the title of a cited work — is left as part of that phrase or title, without `<citedRange>` or a legislation `<title>`:

```
<bibl><seg type="citationContext">For a more detailed analysis of § 60 d UrhG,</seg>
  <author>G. Iacino</author>, <author>P. Kamocki</author>, <author>P. Leinen</author>,
  <title level="m">Assessment of the Impact of the DSM-Directive on Text+</title>.</bibl>
```

Here neither `§ 60 d UrhG` nor `DSM-Directive` is annotated: the cited source is the report, not the statute or the directive.

Data embedded in a title or URL

Court-decision citations usually present the authority, date and docket number as separable fields, which are tagged as flat siblings inside the `<bibl>` (see The issuing authority). Formal titles of legislative acts often do not: the act's number, the issuing body and the enactment date are grammatically part of the title string itself. In that case, keep the title as one continuous `<title>` span in the original word order, and tag the embedded data points in place inside it. `<title>` accepts `<idno>`, `<authority>`, `<date>` and `<ref>` as children for this purpose.

> Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, DSM Directive, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790

```
<bibl type="legislation"><title level="m" type="legislation" key="DSM
  Directive">Directive (EU)
  <idno type="actNumber">2019/790</idno> of <authority type="legislature">the European
  Parliament and of the Council</authority> of <date type="enacted" when="2019-04-17">17
  April
  2019</date> on copyright and related rights in the Digital Single Market and amending
  Directives <ref type="work">96/9/EC</ref> and <ref
  type="work">2001/29/EC</ref></title>,
  <title level="m" type="legislation" subtype="short" key="DSM Directive">DSM
  Directive</title>,
  <ptr type="web">https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:<idno
  type="CELEX">32019L0790</idno></ptr>.</bibl>
```

- **Short title.** `subtype="short"` marks an abbreviated or alternate form of a title that is already tagged in full in the same `<bibl>`, typically one that later footnotes use on its own ("DSM Directive"). It combines with any `@level`/`@type` preset. Give both forms the same `@key` so they can be linked when the citation is resolved.
- **Identifier inside a URL.** When an identifier appears only as part of a URL (here the CELEX number in the query string), tag it with a nested `<idno>` inside the `<ptr>`. Do not repeat it outside the `<ptr>`: annotation only adds tags, it never adds or changes text.
- **Other works mentioned in a title.** `<ref type="work">` marks a passing mention of another work that does not get a citation of its own (here the two amended directives). Tagging these is optional.

Nesting does not replace the flat pattern: where the source presents authority, date and identifier as separate fields, keep tagging them as siblings of `<title>`. Only nest when the source text itself nests them.

Intra-footnote references

Legal and humanities footnotes constantly avoid repeating a citation just given, using devices such as *id.*, *op. cit.*, *ibid.*, *a.a.O.*, *ders.*, *supra*, *infra*, "above", or "N. 22". `<ref type="...">` labels these:

`@type`	Meaning	Examples
`precedingWork`	the work just cited	*supra*, *above*, `id.`, `op. cit.`, `a.a.O.`
`subsequentWork`	a work cited later	*infra*, *below*
`precedingAuthor`	the author just named, work unspecified	`ders.`, `idem`

@type	Meaning	Examples
footnote	an explicit cross-reference to another footnote, carrying @n	n. 7, N. 22, "oben N. 22"
work	a passing mention of another work that is not cited in its own <bibl>	"amending Directives 96/9/EC" (see Data embedded in a title or URL)

A footnote cross-reference follows the same rule as <biblScope> and <citedRange>: the designator (n., N., note, Fn.) is part of the <ref type="footnote">, and @n holds the footnote number.

```
<bibl><author>Vogel</author>, <ref type="precedingWork">id.</ref>,
  <citedRange unit="page" from="79" to="79">p. 79</citedRange></bibl>

<bibl><author>Vogel</author>, <ref type="precedingWork">op. cit.</ref>,
  <ref type="footnote" n="7">n. 7</ref>,
  <citedRange unit="page" from="80" to="81">pp. 80–1</citedRange>.</bibl>

<bibl><ref type="precedingAuthor">Ders.</ref>, <ref type="precedingWork">a.a.O.</ref>,
  <citedRange unit="page" from="123" to="123">S. 123</citedRange>.</bibl>
```

When two of these devices sit next to each other but are lexically distinct — e.g. Kaiser (oben N. 22) 89 — split them into two adjacent <ref>s rather than fusing them into one:

```
<author>Kaiser</author>
(<ref type="precedingWork">oben</ref> <ref type="footnote" n="22">N. 22</ref>)
<biblScope unit="page" from="89" to="89">89</biblScope>
```

The same pattern applies to a *forward*-pointing reference, combining subsequentWork with footnote, e.g. "infra note 24":

```
<ref type="subsequentWork">infra</ref> <ref type="footnote" n="24">note 24</ref>
```

@target is available to hold a resolved pointer once the citation is resolved in post-processing; it is not required for training annotation.

Citation-signal phrases

Mark the text that introduces a citation with a signal word ("see", "see also", "cf.", "but see", "contra", "vgl.", "anderer Ansicht" …) with <seg type="citationContext">:

- The <seg> starts where the preceding citation (or the footnote label) ends and ends where the cited work itself begins (usually its author, editor or title). It contains all text in between, which can be more than one sentence, including words after the signal ("see, for example, the classic text by").
- The <seg> holds plain text only: titles or names inside it are not tagged.

- When several citations share one signal phrase, put the `<seg>` only in the first `<bibl>` it introduces (see the pipeline example and the referenceSegmenter's "introductory comment" examples).

```
<!-- correct -->
<seg type="citationContext">In the Polish literature -cf.</seg> <editor>Z.
  Kmieciak</editor>, …
<!-- incorrect: the leading text is outside the <seg> -->
In the Polish literature <seg type="citationContext">-cf.</seg> <editor>Z.
  Kmieciak</editor>, …
```

Quotations

`<quote>` pairs a direct quotation from the body text with the `<bibl>` that supports it. No attributes: quotation marks and trailing punctuation stay in the source text.

```
<bibl><seg type="citationContext">…</seg><quote>„The U. S. law and development movement
  …"</quote>, <author>Merryman</author>, …</bibl>
```

A real-world example: multiple references with multiple comments

Footnotes routinely mix several of the devices above in one paragraph — multiple citations, each with its own signal phrase or `supra`-style back-reference:

> [23] *See* 1 WILLIAM BLACKSTONE, COMMENTARIES ON THE LAWS OF ENGLAND 37 (The Legal Classics Library 1983) (1765-70) (quoting with approval John Fortescue's endorsement that lay students "trace[] up the principles and grounds of the law, even to their original elements"). Blackstone's use of natural law in the *Commentaries* may also reflect his own introduction to the law. It was St. German's *Doctor and Student*, a work integrating theological and legal thought, that drew Blackstone to the law. Cook, *supra* note 12, at 170. Perhaps for similar reasons, the *Commentaries* turned students from theology to law. McKnight, *supra* note 2, at 401. Note that the pagination of the various editions of the *Commentaries* has become more or less standardized. This move is not without its problems, however. *See* Alschuler, *supra* note 1, at 3 n.4; *cf.* Carli N. Conklin, *The Origins of the Pursuit of Happiness*, 7 WASH. U. JURIS. REV. 195, 200 (2015) (selecting for use the first edition, as does the present article); Alan Watson, *The Structure of Blackstone's Commentaries*, 97 YALE L.J. 795, 801 (1988) (same).

```
<bibl><label>23</label> <seg type="citationContext">See</seg> <author>1 William
  Blackstone</author>,
  <title level="m">Commentaries on the Laws of England</title>
  <citedRange unit="page" from="37" to="37">37</citedRange> (The Legal Classics Library
  1983) (1765-70))
  (quoting with approval John Fortescue's endorsement that lay students "trace[] up the
  principles and grounds of the law, even to their original elements"). Blackstone's use
  of natural law in the <title level="m">Commentaries</title> may also reflect his own
  introduction to the law. It was St. German's <title level="m">Doctor and
  Student</title>,
  a work integrating theological and legal thought, that drew Blackstone to the law.
</bibl>
<bibl><author>Cook</author>, <ref type="precedingWork">supra</ref> <ref type="footnote"
  n="12">note 12</ref>,
  <citedRange unit="page" from="170" to="170">170</citedRange>. Perhaps for similar
  reasons, the
  <title level="m">Commentaries</title> turned students from theology to law.
</bibl>
<bibl><author>McKnight</author>, <ref type="precedingWork">supra</ref> <ref
  type="footnote" n="2">note 2</ref>,
  <citedRange unit="page" from="401" to="401">401</citedRange>.
</bibl>
<bibl><seg type="citationContext">Note that the pagination of the various
  editions of the Commentaries has become more or less
```

```
  standardized. This move is not without its problems, however. See</seg>
  <author>Alschuler</author>,
  <ref type="precedingWork">supra</ref> <ref type="footnote" n="1">note 1</ref>,
  <citedRange unit="page" from="3" to="3">3</citedRange> n.4;
</bibl>
<bibl><seg type="citationContext">cf.</seg> <author>Carli N. Conklin</author>,
  <title level="a">The Origins of the Pursuit of Happiness</title>,
  <title level="j">7 Wash. U. Juris. Rev.</title>
  <biblScope unit="page" from="195">195</biblScope>, <citedRange unit="page" from="200"
  to="200">200</citedRange> (<date type="publication">2015</date>)
  (selecting for use the first edition, as does the present article);
</bibl>
<bibl><author>Alan Watson</author>, <title level="a">The Structure of Blackstone's
  Commentaries</title>,
  <title level="j">97 Yale L.J.</title> <biblScope unit="page"
  from="795">795</biblScope>, <citedRange unit="page" from="801"
  to="801">801</citedRange> (<date type="publication">1988</date>) (same).
</bibl>
```

The point is the pattern: one comment/signal phrase per `<bibl>` it introduces, `supra note N` split into `<ref type="precedingWork">supra</ref>` + `<ref type="footnote" n="N">note N</ref>` exactly as in Intra-footnote references, and no special nesting for the multiple citations sharing one footnote. (This annotation validates against `schema/grobid.training.references.rng`.)

The reference-segmentation examples, annotated

The reference segmentation model examples above show each footnote only up to the point of splitting it into raw, unparsed `<bibl>` citations. Picking most of them up again here shows what the `references` model does with that same raw text next — the same vocabulary introduced above, applied to real footnotes rather than invented ones. (The several references in one footnote example is skipped: it is a plain, non-legal citation list with nothing further to annotate beyond ordinary `<author>`/`<title>`/`<date>`, already covered under Basic bibliographic elements.)

This is the **German legal citation** example:

> 2 Dazu etwa: *Hofmann*, ZGE 2016, 482, 498; *Becker*, ZGE 2016, 239, 273; *Raue*, ZGE 2014, 387, 389.

```
<bibl><label>2</label> <seg type="citationContext">Dazu etwa:</seg>
  <author>Hofmann</author>,
  <title level="j">ZGE</title> <date type="publication">2016</date>,
  <biblScope unit="page" from="482">482</biblScope>, <citedRange unit="page" from="498"
  to="498">498</citedRange>;
</bibl>
<bibl><author>Becker</author>, <title level="j">ZGE</title> <date
  type="publication">2016</date>,
  <biblScope unit="page" from="239">239</biblScope>, <citedRange unit="page" from="273"
  to="273">273</citedRange>;
</bibl>
<bibl><author>Raue</author>, <title level="j">ZGE</title> <date
  type="publication">2014</date>,
  <biblScope unit="page" from="387">387</biblScope>, <citedRange unit="page" from="389"
  to="389">389</citedRange>.
</bibl>
```

Each citation follows the German convention of *first page*, *pinpoint page* ("482, 498"), so the first page becomes `<biblScope unit="page">` (the article's own extent — only `@from` is set, since the article's actual last page is never given) and the pinpoint becomes `<citedRange unit="page">` — the same page/pinpoint distinction as `233 (240)` in the real-world example below.

This is **introductory comment, example 1** (Kuschel):

> 3 Zur Frage, ob der Erwerb eines (digitalen) Werkexemplars mit dem Erwerb eines dinglichen Genussrechts verbunden ist, vgl.: *Kuschel*, Der Erwerb digitaler Werkexemplare zur privaten Nutzung, 2019.

```
<bibl><label>3</label> <seg type="citationContext">Zur Frage, ob der Erwerb eines
  (digitalen) Werkexemplars mit dem Erwerb eines dinglichen Genussrechts verbunden ist,
  vgl.:</seg>
  <author>Kuschel</author>, <title level="m">Der Erwerb digitaler Werkexemplare zur
  privaten Nutzung</title>,
  <date type="publication">2019</date>.
</bibl>
```

This is **introductory comment, example 2** (Lichuma / Dehbi & Martin-Ortega / Bonfanti). Note how `cit.` — the Italian-style short form for "already cited above" — is tagged the same way as `supra`/`a.a.O.` in Intra-footnote references, and how the third citation carries its own, separate signal phrase rather than sharing the first one:

> [15] Arguing that national corporate due diligence laws potentially breach the principle of consent in international law and the sovereignty of host States, and perpetuate power imbalances of colonial derivation, see e.g.: C. OMARI LICHUMA, *(Laws) Made in the 'First World'*, cit., pp. 517-518; F. DEHBI, O. MARTIN-ORTEGA, *An integrated approach to corporate due diligence from a human rights, environmental, and TWAIL perspective*, in *Regulation & Governance*, 2023, n. 17, pp. 927-943, in particular pp. 932-935. In contrast, affirming that such national legislations are to be regarded as instruments facilitating home States' compliance with their international obligations, rather than as breaches of the host States' sovereignty, see: A. BONFANTI, *Imprese multinazionali, diritti umani e ambiente. Profili di diritto internazionale pubblico e privato*, Milano, 2012, p. 136.

```
<bibl><label>15</label> <seg type="citationContext">Arguing that national corporate due
  diligence laws potentially breach the principle of consent in international law and the
  sovereignty of host States, and perpetuate power imbalances of colonial derivation, see
  e.g.:</seg>
  <author>C. Omari Lichuma</author>, <title level="a">(Laws) Made in the 'First
  World'</title>,
  <ref type="precedingWork">cit.</ref>, <biblScope unit="page" from="517" to="518">pp.
  517-518</biblScope>;
</bibl>
<bibl>
  <author>F. Dehbi</author>, <author>O. Martin-Ortega</author>,
  <title level="a">An integrated approach to corporate due diligence from a human rights,
  environmental, and TWAIL perspective</title>,
  <title level="j">Regulation &amp;amp; Governance</title>, <date
  type="publication">2023</date>,
  <biblScope unit="issue" from="17" to="17">n. 17</biblScope>, <biblScope unit="page"
  from="927" to="943">pp. 927-943</biblScope>,
  in particular <citedRange unit="page" from="932" to="935">pp. 932-935</citedRange>.
</bibl>
<bibl>
  <seg type="citationContext">In contrast, affirming that such national legislations are
  to be regarded as instruments facilitating home States' compliance with their
  international obligations, rather than as breaches of the host States' sovereignty,
  see:</seg>
  <author>A. Bonfanti</author>, <title level="m">Imprese multinazionali, diritti umani e
  ambiente. Profili di diritto internazionale pubblico e privato</title>,
  <pubPlace>Milano</pubPlace>, <date type="publication">2012</date>, <biblScope
  unit="page" from="136" to="136">p. 136</biblScope>.
</bibl>
```

This is the **trailing comment** example. It combines an `<authority>` (a court, not a person), a `<title type="caseName">`, and a `<quote>` — with two separate `<seg type="citationContext">` spans, since the signal phrase and the phrase introducing the quotation are two distinct devices. The closing

"See infra, section 3.2." points to a section of the *citing* article itself, not to another footnote or work, so it falls outside the `<ref>` vocabulary documented above and is left as plain text:

> [9] See e.g.: Inter-American Court of Human Rights, judgment of 16 February 2020, *Indigenous Communities of the Lhaka Honhat (Our Land) Association v. Argentina*, par. 254, quoting the *amicus curiae* intervention of the former UN Special Rapporteur on the Right to Food, De Schutter: «Many indigenous peoples understand the right to adequate food as a collective right. They often see subsistence activities such as hunting, fishing and gathering as essential not only to their right to food, but to nurturing their cultures, languages, social life and identity». The cultural value of food is recognized – with respect to indigenous peoples in particular – also under art. 27 of the International Covenant on Civil and Political Rights (ICCPR). See *infra*, section 3.2.

```
<bibl><label>9</label> <seg type="citationContext">See e.g.:</seg>
  <authority type="court">Inter-American Court of Human Rights</authority>, judgment of
  <date type="decision" when="2020-02-16">16 February 2020</date>,
  <title level="a" type="caseName">Indigenous Communities of the Lhaka Honhat (Our Land)
  Association v. Argentina</title>,
  <citedRange unit="number" from="254" to="254">par. 254</citedRange>,
  <seg type="citationContext">quoting the amicus curiae intervention of the former UN
  Special Rapporteur on the Right to Food, De Schutter:</seg>
  <quote>«Many indigenous peoples understand the right to adequate food as a collective
  right. They often see subsistence activities such as hunting, fishing and gathering as
  essential not only to their right to food, but to nurturing their cultures, languages,
  social life and identity».</quote>
  The cultural value of food is recognized – with respect to indigenous peoples in
  particular – also under
  art. 27 of the International Covenant on Civil and Political Rights (ICCPR). See infra,
  section 3.2.
</bibl>
```

This is the **court decisions** example:

> [6] See, *ex multis*: African Commission on Human and Peoples Rights, decision of 27 October 2001, *Social and Economic Rights Action Center (SERAC) and Center for Economic and Social Rights (CESR) v. Nigeria*; Inter-American Court of Human Rights, judgment of 27 June 2012, *Case of the Kichwa Indigenous People of Sarayaku v. Ecuador*.

```
<bibl><label>6</label> <seg type="citationContext">See, ex multis:</seg>
  <authority type="court">African Commission on Human and Peoples Rights</authority>,
  decision of
  <date type="decision" when="2001-10-27">27 October 2001</date>,
  <title level="a" type="caseName">Social and Economic Rights Action Center (SERAC) and
  Center for Economic and Social Rights (CESR) v. Nigeria</title>;
</bibl>
<bibl>
  <authority type="court">Inter-American Court of Human Rights</authority>, judgment of
  <date type="decision" when="2012-06-27">27 June 2012</date>,
  <title level="a" type="caseName">Case of the Kichwa Indigenous People of Sarayaku v.
  Ecuador</title>.
</bibl>
```

(All five annotations above validate against `schema/grobid.training.references.rng`.)

Worked example: the full annotation pipeline

The example below is meant to show how one passage of body text and its footnotes move through all three models in sequence, from raw PDF text to a fully annotated citation — including a backward-pointing reference (`supra`) *and* a forward-pointing one (`infra`).

The passage of body text reads:

> Comparative criminology has increasingly examined the transnational transfer of legal concepts.[8]. [...] This same methodological caution applies with particular force to law-and-development scholarship.[19]

Footnote 8 gives a full citation that footnote 19 later cites back to with `supra`. Footnote 19 is the composite footnote itself — several citations joined by signal phrases and a quotation, with a `supra` back-reference to note 8 and an `infra` forward-reference to note 24. Footnote 24, quoted only in part below, is what that forward reference points to.

Stage 1: document segmentation model

At this stage the footnote text is not yet split into individual citations — each footnote is simply wrapped in its own `<listBibl>` as raw text, exactly as extracted from the PDF:

```
<body>
  Comparative criminology has increasingly examined the transnational transfer of
  legal concepts.<lb/>8 [...] This same methodological caution applies with particular
  force to law-and-development scholarship.<lb/>19
</body>

<listBibl>
8 Günther Kaiser, Kriminologie: Ein Lehrbuch (3d ed. 1996).<lb/>
</listBibl>

<listBibl>
19 See, e.g., Neil J. Smelser, Theory of Collective Behavior 175-76 (1962). On<lb/>
criminology, see Kaiser, supra note 8, at 89, as well as Blazicek &amp;amp; Janeksela,
  Some<lb/>
Comments on Comparative Methodologies in Criminal Justice, 6 Int'l J. Comp. &amp;amp;
  Applied<lb/>
Crim. Just. 233, 240 (1978). The uncritical transfer of such concepts to countries
  of<lb/>
the Third World has proven especially dangerous, as discussed further infra note 24.<lb/>
Commentators reached the following conclusion: "The U.S. law and development
  movement<lb/>
was largely a parochial expression of the American legal style." John Henry
  Merryman,<lb/>
Comparative Law and Social Change: On the Origins, Style, Decline and Revival of the<lb/>
Law and Development Movement, 25 Am. J. Comp. L. 457, 479 (1977).<lb/>
</listBibl>

<listBibl>
24 Elena Vargas, Law and Development Reconsidered 112-14 (2005) (arguing that the<lb/>
movement's failures reflected its parochial assumptions rather than any flaw in its<lb/>
underlying goals).<lb/>
</listBibl>
```

Stage 2: reference segmentation model

The reference segmenter splits the raw footnote text into `<label>` + `<bibl>` pairs — one `<bibl>` per footnote, or per individual citation for a footnote that (like note 19) bundles several. Nothing inside a `<bibl>` is parsed yet:

```
<listBibl>
  <bibl><label>8</label> Günther Kaiser, Kriminologie: Ein Lehrbuch (3d ed. 1996).
  </bibl>

  <bibl><label>19</label> See, e.g., Neil J. Smelser, Theory of Collective Behavior 175-
  76 (1962). </bibl>
  <bibl>On criminology, see Kaiser, supra note 8, at 89, as well as </bibl>
  <bibl>Blazicek &amp;amp; Janeksela, Some Comments on Comparative Methodologies in Criminal
  Justice, 6 Int'l J. Comp. &amp;amp; Applied Crim. Just. 233, 240 (1978). </bibl>
  <bibl>The uncritical transfer of such concepts to countries of the Third World has
  proven especially dangerous, as discussed further infra note 24. Commentators reached
  the following conclusion: "The U.S. law and development movement was largely a
  parochial expression of the American legal style." John Henry Merryman, Comparative Law
  and Social Change: On the Origins, Style, Decline and Revival of the Law and
  Development Movement, 25 Am. J. Comp. L. 457, 479 (1977).</bibl>

  <bibl><label>24</label> Elena Vargas, Law and Development Reconsidered 112-14 (2005)
  (arguing that the movement's failures reflected its parochial assumptions rather than
  any flaw in its underlying goals).</bibl>
</listBibl>
```

Stage 3: citation (`references`) model

Finally, the `references` model labels the internal structure of each `<bibl>` — authors, titles, extents, pinpoints, signal phrases, the quotation, and both anaphoric references:

```
<listBibl>
  <bibl><label>8</label> <author>Günther Kaiser</author>, <title level="m">Kriminologie:
  Ein Lehrbuch</title>
    <edition n="3">3d ed.</edition> (<date type="publication">1996</date>).
  </bibl>

  <bibl><label>19</label> <seg type="citationContext">See, e.g.,</seg>
    <author>Neil J. Smelser</author>, <title level="m">Theory of Collective
  Behavior</title>
    <biblScope unit="page" from="175" to="176">175-76</biblScope>
    (<date type="publication">1962</date>).
  </bibl>
  <bibl><seg type="citationContext">On criminology, see</seg>
    <author>Kaiser</author>, <ref type="precedingWork">supra</ref> <ref type="footnote"
  n="8">note 8</ref>,
    <biblScope unit="page" from="89" to="89">89</biblScope>, as well as
  </bibl>
  <bibl>
    <author>Blazicek</author> &amp;amp; <author>Janeksela</author>,
    <title level="a">Some Comments on Comparative Methodologies in Criminal
  Justice</title>,
    <biblScope unit="volume" from="6" to="6">6</biblScope>
    <title level="j">Int'l J. Comp. &amp;amp; Applied Crim. Just.</title>
    <biblScope unit="page" from="233">233</biblScope>,
    <citedRange unit="page" from="240" to="240">240</citedRange>
    (<date type="publication">1978</date>).
  </bibl>
  <bibl>
    <seg type="citationContext">The uncritical transfer of such concepts to countries of
  the Third World has proven especially dangerous, as discussed further</seg>
```

```
    <ref type="subsequentWork">infra</ref> <ref type="footnote" n="24">note 24</ref>.
    <seg type="citationContext">Commentators reached the following conclusion:</seg>
    <quote>The U.S. law and development movement was largely a parochial expression of
  the American legal style.</quote>
    <author>John Henry Merryman</author>,
    <title level="a">Comparative Law and Social Change: On the Origins, Style, Decline
  and Revival of the Law and Development Movement</title>,
    <biblScope unit="volume" from="25" to="25">25</biblScope>
    <title level="j">Am. J. Comp. L.</title>
    <biblScope unit="page" from="457">457</biblScope>
    <citedRange unit="page" from="479" to="479">479</citedRange>
    (<date type="publication">1977</date>).
  </bibl>

  <bibl><label>24</label> <author>Elena Vargas</author>, <title level="m">Law and
  Development Reconsidered</title>
    <biblScope unit="page" from="112" to="114">112-14</biblScope>
    (<date type="publication">2005</date>)
    (arguing that the movement's failures reflected its parochial assumptions rather than
  any flaw in its underlying goals).
  </bibl>
</listBibl>
```