

IBM DocLayNet Labeling Guide

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed Nassar, Peter W. J. Staar

IBM Research, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

{bpf, cau, dol, ahn, taa}@zurich.ibm.com

August 19, 2022

Table of Contents

| | |
|---|-----------|
| PART I. GENERAL RULES | 5 |
| 1 INTRODUCTION..... | 5 |
| 2 BASICS OF CLUSTER LABELLING | 7 |
| 2.1 COMMON PRACTICES | 7 |
| 2.2 REPORT ERROR OR CONFUSING PAGE | 9 |
| 2.3 EVERY ELEMENT ON A SUBMITTED PAGE NEEDS TO BE LABELLED..... | 10 |
| 3 MORE DETAILS FOR EACH LABEL – FOR ALL DOCUMENT CATEGORIES | 11 |
| 3.1 TEXT | 11 |
| 3.1.1 <i>Normal Paragraphs</i> | 11 |
| 3.1.2 <i>Sidebars</i> | 11 |
| 3.1.3 <i>Authors, Affiliations, and Abstracts</i> | 12 |
| 3.1.4 <i>Quotes</i> | 12 |
| 3.1.5 <i>Mathematical Theorems etc.</i> | 12 |
| 3.1.6 <i>Emphasized Paragraph Start: Text</i> | 13 |
| 3.1.7 <i>Multiple Short Lines without List Identifiers that Clearly Belong Together</i> | 13 |
| 3.1.8 <i>Text around a Centered Formula or Citation</i> | 13 |
| 3.1.9 <i>Code</i> | 14 |
| 3.2 PICTURE..... | 14 |
| 3.2.1 <i>Use Single-Page View</i> | 15 |
| 3.2.2 <i>Legends</i> | 15 |
| 3.2.3 <i>One Picture per Caption, if Captions Exist</i> | 15 |
| 3.2.4 <i>Multiple Small Pictures</i> | 16 |
| 3.2.5 <i>Forms</i> | 18 |
| 3.3 CAPTION | 19 |
| 3.3.1 <i>Normal Captions</i> | 19 |
| 3.4 SECTION-HEADER..... | 20 |
| 3.4.1 <i>Normal Headers</i> | 20 |
| 3.4.2 <i>Emphasized Text at Start of a Line: Normal Text</i> | 21 |
| 3.5 FOOTNOTE | 21 |
| 3.5.1 <i>Normal Footnotes</i> | 21 |
| 3.5.2 <i>Several Footnotes on One Page: Several Clusters</i> | 21 |
| 3.6 FORMULA | 22 |
| 3.6.1 <i>Normal Formulas</i> | 22 |
| 3.6.2 <i>Several Successive Formula Lines</i> | 22 |
| 3.7 TABLE..... | 23 |
| 3.7.1 <i>Normal Tables</i> | 23 |
| 3.8 TITLE | 25 |
| 3.8.1 <i>Title Repeated on Later Pages: "Text"</i> | 25 |
| 3.9 LIST-ITEM..... | 25 |
| 3.9.1 <i>Key Rule: Hanging Shape</i> | 25 |
| 3.10 PAGE-HEADER..... | 27 |
| 3.10.1 <i>Normal Page-headers</i> | 27 |

| | |
|--|-----------|
| 3.10.2 <i>Only Page Number: Page-header</i> | 27 |
| 3.10.3 <i>Sideways Page-header</i> | 27 |
| 3.11 PAGE-FOOTER..... | 28 |
| 3.12 NONE..... | 28 |
| 4 EXAMPLES OF PAGES WITH ERRORS..... | 30 |
| 4.1 TOO MANY TEXT BOXES: SUBMIT WITH “NONE” | 30 |
| 4.2 MISSING TEXT BOXES: DO NOT SUBMIT | 30 |
| 4.3 BROKEN-UP TEXT BOXES: DO NOT SUBMIT | 31 |
| PART II. SPECIAL CASES BY DOCUMENT CATEGORY | 33 |
| 5 LAWS AND REGULATIONS | 33 |
| 5.1 US LAWS | 33 |
| 5.1.1 <i>Text</i> | 33 |
| 5.1.2 <i>Page-header</i> | 34 |
| 5.1.3 <i>Section-header</i> | 34 |
| 5.1.4 <i>None</i> | 34 |
| 5.1.5 <i>US Law Title Pages</i> | 35 |
| 5.1.6 <i>US Law Table of Contents</i> | 35 |
| 5.1.7 <i>US Law Final Pages</i> | 36 |
| 5.2 BRITISH LAWS (COLLECTION “GB_LAWS”) | 36 |
| 5.2.1 <i>Text and List-items</i> | 36 |
| 5.2.2 <i>Section-header</i> | 36 |
| 5.2.3 <i>British Law Title Pages</i> | 37 |
| 5.2.4 <i>British Law Table of Contents</i> | 38 |
| 5.2.5 <i>British Law Final Pages</i> | 38 |
| 5.3 GERMAN LAWS | 39 |
| 5.3.1 <i>Section-header</i> | 39 |
| 5.3.2 <i>Table</i> | 39 |
| 5.3.3 <i>List-item</i> | 40 |
| 5.3.4 <i>List-like with Third Column: Table</i> | 41 |
| 5.3.5 <i>German Law Title Pages</i> | 41 |
| 5.3.6 <i>German Law Table of Contents</i> | 42 |
| 5.4 PHILIPPINE LAWS..... | 42 |
| 5.4.1 <i>Section-header</i> | 42 |
| 5.4.2 <i>List-item</i> | 42 |
| 5.4.3 <i>Philippine Law Title Pages</i> | 43 |
| 5.5 BOTSWANA LAWS..... | 44 |
| 5.5.1 <i>Botswana Law Title Pages</i> | 44 |
| 5.6 RUSSIAN LAWS..... | 44 |
| 5.6.1 <i>Tables where Gridlines End Early</i> | 45 |
| 5.6.2 <i>Russian Law Title Pages</i> | 46 |
| 5.7 CHINESE LAWS | 46 |
| 5.7.1 <i>Chinese Law Title Pages</i> | 47 |
| 5.7.2 <i>Chinese Law Table of Contents</i> | 48 |
| 5.8 JAPANESE LAWS | 48 |
| 5.8.1 <i>Japanese Law Title Pages</i> | 48 |
| 5.9 FAA REGULATIONS | 50 |
| 5.9.1 <i>FAA Regulation Title Pages</i> | 50 |
| 6 PATENTS | 51 |
| 6.1 US PATENTS..... | 51 |
| 6.1.1 <i>Caption-like Text Continues after Table Line: Part of Table</i> | 51 |
| 6.1.2 <i>Rich Chemical Formulas: Pictures</i> | 52 |
| 6.1.3 <i>Column header: Also a Page-header</i> | 54 |
| 6.1.4 <i>Line Numbering: None</i> | 54 |

| | | |
|----------|--|-----------|
| 6.1.5 | <i>US Patent Title Pages</i> | 55 |
| 6.2 | WIPO PATENTS | 57 |
| 6.2.1 | <i>WIPO Patent Title Pages</i> | 57 |
| 7 | SCIENTIFIC ARTICLES..... | 57 |
| 7.1 | SECTION-HEADER VS. LIST-ITEM..... | 58 |
| 7.2 | FORMULA | 58 |
| 7.2.1 | <i>Number on Separate Line.....</i> | 58 |
| 7.2.2 | <i>Several Numbers in Multi-Line Formula.....</i> | 59 |
| 7.3 | LIST-ITEM..... | 59 |
| 7.3.1 | <i>Lists with Missed Hanging Shape: Error.....</i> | 59 |
| 7.3.2 | <i>Citations: Treat with the Normal List Rules.....</i> | 59 |
| 7.3.3 | <i>Continuation of List-item at Top of Page: Text</i> | 61 |
| 7.4 | SCIENTIFIC TITLE PAGES | 61 |
| 8 | FREE-FORM DOCUMENT CATEGORIES: FINANCIAL REPORTS, MANUALS, AND GOVERNMENT TENDERS | 65 |
| 8.1 | TEXT | 65 |
| 8.1.1 | <i>Slightly Emphasized Text after Section-header: Text</i> | 65 |
| 8.1.2 | <i>Text Flowing around a Picture</i> | 66 |
| 8.1.3 | <i>Text in Boxes</i> | 67 |
| 8.1.4 | <i>Text Boxes as Pictures: Don't Submit</i> | 68 |
| 8.1.5 | <i>Not-hanging Multi-column Lists: Text.....</i> | 69 |
| 8.1.6 | <i>Code: Text, except if very Tabular</i> | 69 |
| 8.1.7 | <i>Diagonal Text over a Page</i> | 70 |
| 8.1.8 | <i>Wrong Text that Can Still be Labeled.....</i> | 70 |
| 8.2 | PICTURE..... | 71 |
| 8.2.1 | <i>Multiple Small Pictures</i> | 71 |
| 8.2.2 | <i>Background Images behind Text.....</i> | 72 |
| 8.2.3 | <i>Logos</i> | 73 |
| 8.2.4 | <i>Signatures</i> | 74 |
| 8.2.5 | <i>"Irrelevant Pictures": Still Pictures</i> | 75 |
| 8.2.6 | <i>Table in Picture</i> | 76 |
| 8.3 | CAPTION (OR NOT?) | 76 |
| 8.3.1 | <i>Indicators of Captions</i> | 76 |
| 8.3.2 | <i>Special-looking Paragraph near Picture or Table: Caption</i> | 77 |
| 8.3.3 | <i>Normal-looking Paragraph near Picture or Table: Text</i> | 77 |
| 8.3.4 | <i>Multi-paragraph Caption</i> | 79 |
| 8.3.5 | <i>Additional Context for Pictures or Tables: Not a Second Caption</i> | 79 |
| 8.4 | SECTION-HEADER..... | 80 |
| 8.5 | FORMULAS..... | 80 |
| 8.6 | TABLE..... | 81 |
| 8.6.1 | <i>One-column Paragraphs in Table Grid</i> | 81 |
| 8.6.2 | <i>Glossaries: Usually Table.....</i> | 82 |
| 8.7 | LIST-ITEM..... | 82 |
| 8.7.1 | <i>Multi-level Lists</i> | 83 |
| 8.7.2 | <i>Each "Hanging" Line Starts a List-item</i> | 83 |
| 8.7.3 | <i>Bullets or Numbers Not Needed.....</i> | 83 |
| 8.7.4 | <i>Normal Paragraphs Inside a List Item: Text</i> | 84 |
| 8.8 | PAGE-HEADER..... | 84 |
| 8.8.1 | <i>Very Large Page-header can become Section-header.....</i> | 84 |
| 8.9 | TITLE PAGES | 84 |
| 8.9.1 | <i>Title Pages with Mostly Text</i> | 85 |
| 8.9.2 | <i>Text in Unusual Orientation</i> | 86 |
| 8.9.3 | <i>Manuals</i> | 87 |
| 8.9.4 | <i>Government Tenders.....</i> | 89 |
| 8.9.5 | <i>Financial Report Title Pages.....</i> | 93 |

| | | |
|--------|--|-----|
| 8.10 | TABLES OF CONTENT | 97 |
| 8.10.1 | Neither Page Numbers nor Lines: List-items | 97 |
| 8.10.2 | With Gridlines between Items: Table | 97 |
| 8.10.3 | With a Page Number Column: Usually Table | 98 |
| 8.10.4 | Page Number Column, but Deeply Indented: List-items..... | 100 |
| 8.10.5 | Special Case Manuals / Redbooks | 101 |
| 8.10.6 | Special Case Financial Reports / SEC Filings | 101 |
| 8.11 | INDEX PAGES | 102 |
| 8.11.1 | Redbook Indices..... | 102 |
| 8.12 | FINAL PAGES..... | 103 |
| 8.12.1 | Redbook Final Pages | 104 |

PART I. General Rules

This labeling guide belongs to the *DocLayNet* document layout analysis dataset.

1 Introduction

The purpose of this document is to provide guidance for ground-truth annotation of documents with labelled clusters.

Cluster labelling means grouping all elements on the page into reasonable paragraphs, headings, tables, pictures, etc. Below you see an example with the original page on the left, a labeled version in the middle, and the labels on the right.



Figure 1: Model architecture of fastText for a sentence with N n-gram features x_1, \dots, x_N . The features are embedded and averaged to form the hidden variable.

This means that the probability of a node is always lower than the one of its parent. Exploring the tree with a depth first search and tracking the maximum probability among the children allows us to discard any node associated with a small probability. In practice, we observe a reduction of the complexity to $O(k \log_2(k))$ at test time. This approach is further extended to compute the T -top targets at the cost of $O(\log(T))$, using a binary heap.

2.2 N-gram features

Bag of words is invariant to word order but taking explicitly this order into account is often computationally very expensive. Instead, we use a bag of n-grams as additional features to capture some partial information about the local word order. This is very efficient in practice while achieving comparable performance to methods that are explicitly on the order (Wang and Manning, 2012).

We maintain a fast and memory efficient mapping of the n-grams by using the *hashing trick* (Weinberger et al., 2009) with the same hashing function as in Mikolov et al. (2011) and 10M bins if we only used bigrams, and 100M otherwise.

3 Experiments

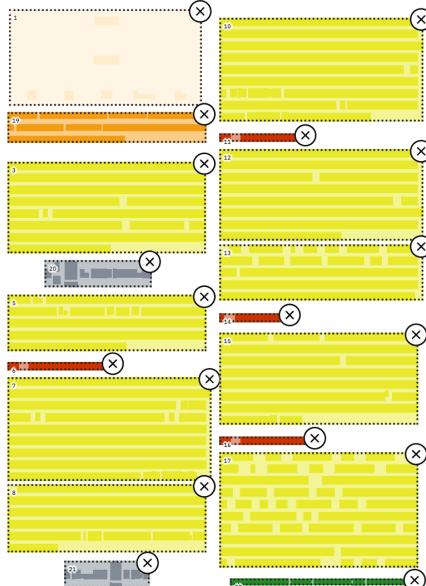
We evaluate fastText on two different tasks. First we compare it to existing classifiers on the problem of sentiment analysis. Then, we evaluate its capacity to scale to large output space on a tag prediction dataset. Note that our model could be implemented with the Vowpal Wabbit library² but we observe in practice, that our tailored implementation is at least 2.5× faster.

3.1 Sentiment analysis

Datasets and baselines. We employ the same 8 datasets and evaluation protocol of Zhang et al. (2015). We report the n-grams and TFIDF baselines from Zhang et al. (2015), as well as the character level convolutional model (char-CNN) of Kim (2014), the character level LSTM, character based convolutional recurrent network (char-CRNN) of Xiao and Cho (2016) and the very deep convolutional network (VDCNN) of Conneau et al. (2016). We also compare

$$P(n_{l+1}) = \prod_{i=1}^l P(n_i).$$

²Using the option --nn, --ngrams and --logmulti.



| Field labels |
|--|
| Identify document elements using the labels below. |
| <input checked="" type="checkbox"/> None |

| Clusters |
|---|
| ■ Text |
| ■ Picture |
| ■ Formula |
| ■ Section-header |
| ■ Page-footer |
| ■ Page-header |
| ■ Footnote |
| ■ Table |
| ■ Caption |
| ■ List-item |
| ■ Title |

The labels visible in the example above have the following meaning:

Text: Regular paragraphs.

Picture: A graphic or photograph.

Caption: Special text outside a picture or table that introduces this picture or table.

Section-header: Any kind of heading in the text, except overall document title.

Footnote: Typically small text at the bottom of a page, with a number or symbol that is referred to in the text above.

Formula: Mathematical equation on its own line.

Further labels not shown in the example above:

Table: Material arranged in a grid alignment with rows and columns, often with separator lines.

List-item: One element of a list, in a hanging shape, i.e., from the second line onwards the paragraph is indented more than the first line.

Page-header: Repeating elements like page number at the top, outside of the normal text flow.

Page-footer: Repeating elements like page number at the bottom, outside of the normal text flow.

Title:¹ Overall title of a document, (almost) exclusively on the first page and typically appearing in large font.

None: Initial state of each cell/element. Only keep this if the element is not a text or picture or anything else of value. For instance, a smear or an invisible/empty cell should remain “None”.

Clusters should be *non-overlapping rectangles* enclosing all elements.

Another example page with several Tables and Captions:

| Model | AG | Sogou | DBP | Yelp P. | Yelp F. | Yah. A. | Amz. F. | Amz. P. |
|-----------------------------------|------|-------|------|---------|---------|---------|---------|---------|
| BoW (Zhang et al., 2015) | 88.8 | 92.9 | 96.6 | 92.2 | 58.0 | 68.9 | 54.6 | 90.4 |
| ngrams (Zhang et al., 2015) | 92.0 | 97.1 | 98.6 | 95.6 | 56.3 | 68.5 | 54.3 | 92.0 |
| ngrams TFIDF (Zhang et al., 2015) | 92.4 | 97.2 | 98.7 | 95.4 | 54.8 | 68.5 | 52.4 | 91.5 |
| char-CNN (Zhang and LeCun, 2015) | 87.2 | 95.1 | 98.3 | 94.7 | 62.0 | 71.2 | 59.5 | 94.5 |
| char-CRNN (Xiao and Cho, 2016) | 91.4 | 95.2 | 98.6 | 94.5 | 61.8 | 71.7 | 59.2 | 94.1 |
| VDCNN (Conneau et al., 2016) | 91.3 | 96.8 | 98.7 | 95.7 | 64.7 | 73.4 | 63.0 | 95.7 |
| fastText, $h = 10$ | 91.5 | 93.9 | 98.1 | 93.8 | 60.4 | 72.0 | 55.8 | 91.2 |
| fastText, $h = 10$, bigram | 92.5 | 96.8 | 98.6 | 95.7 | 63.9 | 72.3 | 60.2 | 94.6 |

Table 1: Test accuracy (%) on sentiment datasets. fastText has been run with the same parameters for all the datasets. It has 10 hidden units and we evaluate it with and without bigrams. For char-CNN, we show the best reported numbers without data augmentation.

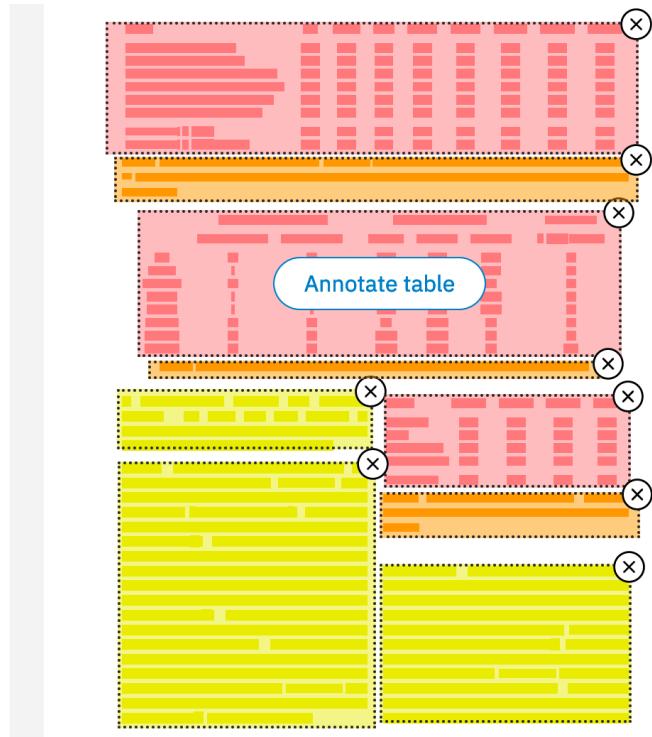
| | Zhang and LeCun (2015) | | Conneau et al. (2016) | | | fastText |
|---------|------------------------|--------------|-----------------------|----------|----------|-------------------|
| | small char-CNN | big char-CNN | depth=9 | depth=17 | depth=29 | $h = 10$, bigram |
| AG | 1h | 3h | 24m | 37m | 51m | 1s |
| Sogou | - | - | 25m | 41m | 56m | 7s |
| DBpedia | 2h | 5h | 27m | 44m | 1h | 2s |
| Yelp P. | - | - | 28m | 43m | 1h09 | 3s |
| Yelp F. | - | - | 29m | 45m | 1h12 | 4s |
| Yah. A. | 8h | 1d | 1h | 1h33 | 2h | 5s |
| Amz. F. | 2d | 5d | 2h45 | 4h20 | 7h | 9s |
| Amz. P. | 2d | 5d | 2h45 | 4h25 | 7h | 10s |

Table 2: Training time for a single epoch on sentiment analysis datasets compared to char-CNN and VDCNN.

to Tang et al. (2015) following their evaluation protocol. We report their main baselines as well as their two approaches based on recurrent networks (Conv-GRNN and LSTM-GRNN).

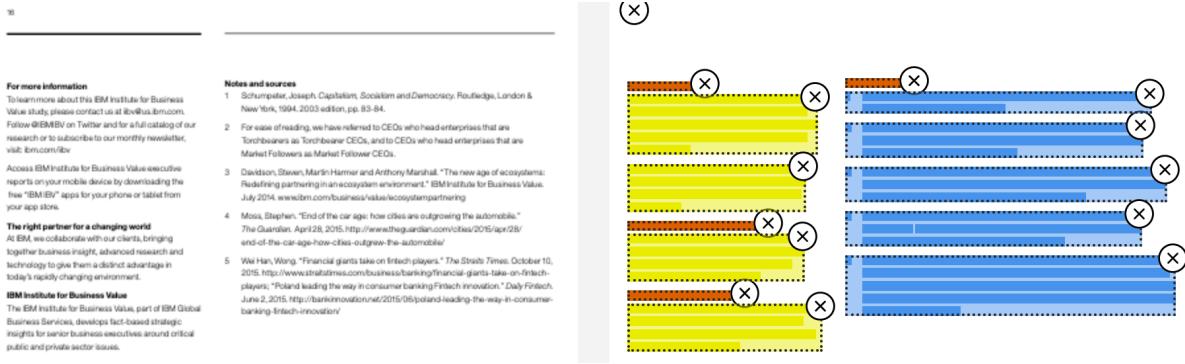
Results. We present the results in Figure 1. We use 10 hidden units and run fastText for 5 epochs with a learning rate selected on a validation set from {0.05, 0.1, 0.25, 0.5}. On this task, adding bigram information improves the performance by 1.4%. Overall our accuracy is slightly better than char-CNN and char-CRNN and, a bit worse than VDCNN. Note that we can increase the accuracy slightly by using more n-grams, for example with trigrams, the performance on Sogou goes up to 97.1%. Finally, Figure 3 shows that our method is competitive with the methods presented in Tang et al. (2015). We tune the hyperparameters on the validation set and observe that using n-grams up to 5 leads to the best performance. Unlike Tang et al. (2015), fastText does not use pre-trained word embeddings, which can be explained the 1% difference in accuracy.

Training time. Both char-CNN and VDCNN are trained on a NVIDIA Tesla K40 GPU, while our models are trained on a CPU using 20 threads. Table 2 shows that methods using convolutions are several orders of magnitude slower than fastText. While it is possible to have a 10x speed up for char-CNN by using more recent CUDA implementations of convolutions, fastText takes less than a minute to train on these datasets. The GRNNs method of Tang et al. (2015) takes around 12 hours per epoch on CPU with a single thread. Our speed-



The third example shows a list, some Section-headers, and a tiny Page-header containing the page number “16”:

¹ Because title pages are generally more difficult, we separate them in the annotation phase so that annotators know whether they are annotating a set of title pages or other pages.



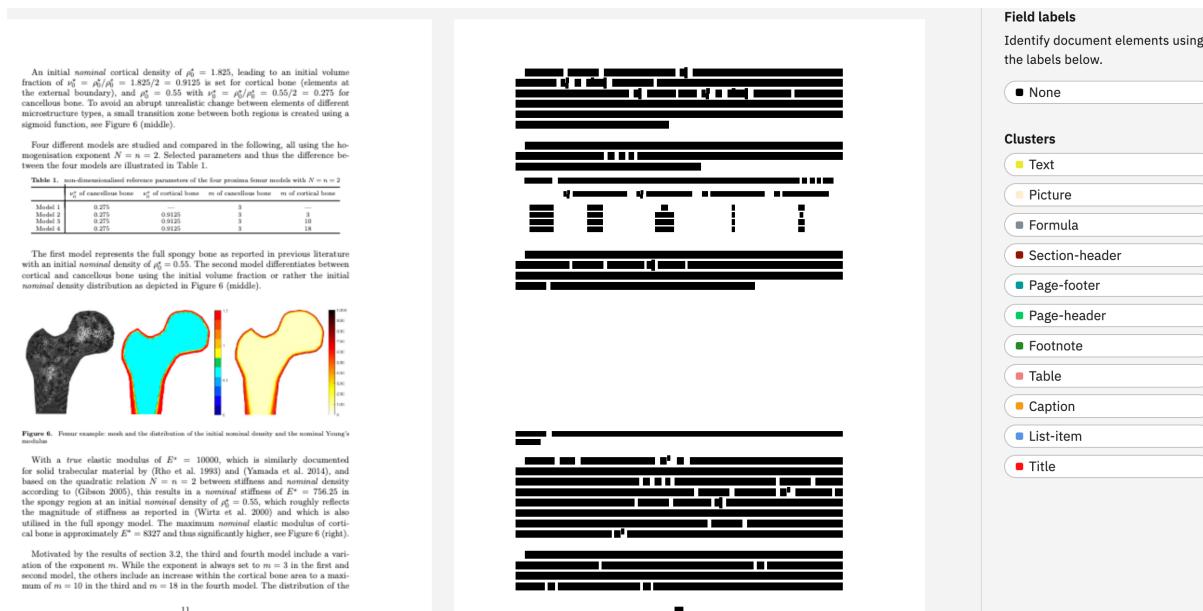
2 Basics of Cluster Labelling

Cluster labelling is performed on the IBM Corpus Conversion Service (CCS), which is part of the Deep Search platform. Annotation is performed page by page, without seeing neighbor pages, and is largely based on each element's geometrical appearance. One typically does not need to read the text content. Exceptions are mentioned below.

The primary criterion for deciding the cluster label is: "How does it look?" (not: "What is the meaning?")

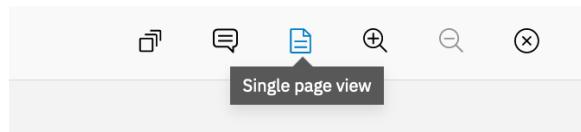
2.1 Common Practices

When you start to label a page in the annotation tool, it looks like this:

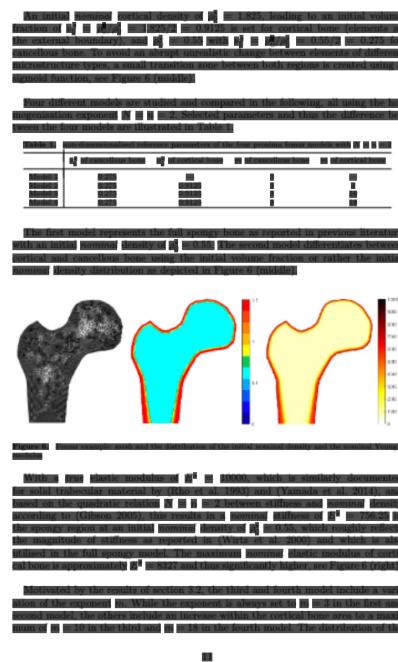


The PDF page is shown on the left. Cell representations to label are in the middle (showing only black cells), and the available labels are in the rightmost palette.

Depending on the type of annotation you make, you can switch the annotator to "single-page view", such that the PDF page and the cells are displayed overlapping. The switch is in the top row:



The result looks like this:



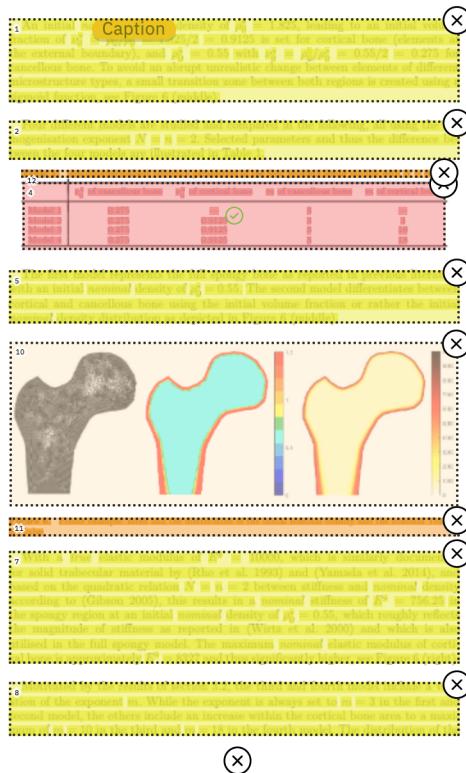
To annotate a paragraph, picture, etc., proceed as follows:

- First, click on the label you want to use, e.g., “Picture” for the picture in the middle.
- For “Picture” or “Table” labels, you now draw the rectangle accurately around the part of the page where this picture or table is. This should include the black cells and all other visual elements (like gridlines, arrows, symbols, legend, etc.) with as little whitespace around as possible.
- For all other labels, you only need to touch all the cells that should be in your cluster with your selection rectangle (sometimes a line or a click is enough) and the cluster will grow to completely include the touched cells. In other words, the cluster snaps to the cells.

The above page should contain the following clusters:

- 1 Picture,
- 1 Table,
- 2 Captions,
- 1 Page-footer,
- 5 “Text” paragraphs.

Here is the final state:



(X)

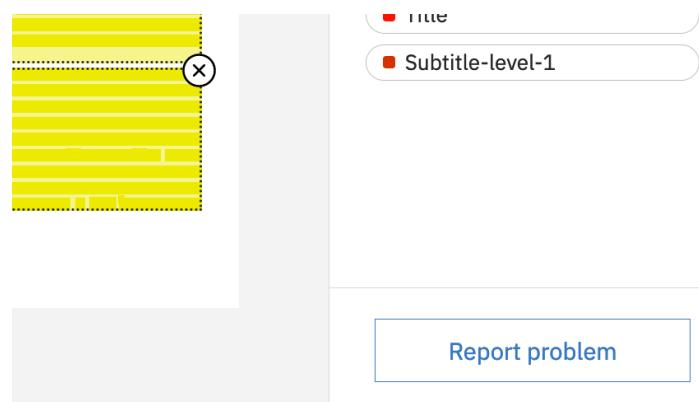
2.2 Report Error or Confusing Page

Note: It is highly advised to NOT submit a wrong or confusing page than trying your best effort on it, because even your best effort is likely to give confusing input to the future AI model training.

If you encounter a page that....

- contains parsing errors (i.e., the black cells don't correspond to the text elements),
- contains some elements that are hard to annotate properly,
- contains elements that you can't decide how to annotate correctly,

...then you should report the page to have problems, using the button on the lower right:



A modal form opens for more details:

Report problem ×

What kind of problem(s) would you like to report?

Parsing problems:

- Incorrect text cell
- Incorrect image

OCR problems:

- Incorrect bounding box
- Incorrect text

I am unsure how to label this page

Other (please describe problem below)

Please let us know what went wrong...

DISCLAIMER
By reporting a problem with this page you empower the administrators to investigate the issue and have access to your document.

Cancel Send

The choices you may use for this annotation campaign:

- “Incorrect text cell”: Either there is no black box where there is some visible text, or a black box where there is no visible text.
- “Incorrect bounding box”: Some text/element and the corresponding black box are not well aligned.
- “I am unsure how to label this page”: It is not clear how to correctly annotate the elements on this page, or you feel undecided between multiple labelling alternatives.

It is appreciated if you use the Description box to provide details about where the error or the confusion is.

2.3 Every Element on a Submitted Page needs to be Labelled

Unless the page is subject to a problem report, it is important to make sure that all elements on a page are labelled correctly and completely. This can include the “None” label, but only if the elements with “None” really are no proper text, picture, or table.

Section 3 provides details for each label, in the order as the labels were introduced above. For each label, we start with clear, common cases, and then describe how certain borderline cases should be labeled.

3 More Details for Each Label – for All Document Categories

3.1 Text

“Text” is the most common label. It is used for normal text paragraphs, and for any other text that fits none of the more special categories. For instance, author names, sidebars, and centered citations are also “Text”.

3.1.1 Normal Paragraphs

Each paragraph becomes a cluster. A paragraph often has a little white space above and below. Or, as in the first example in Section 1, it starts with an indented line and ends without reaching the right margin. A few documents have neither white space nor indents, and one has to guess from the line lengths, and that paragraphs tend to start with capital letters.

Some documents are “double-spaced”, i.e., with large distances between lines. Still, text belonging together and characterized with indents, even more space etc. is one paragraph, like here:

lags as small as 10. Conversely, correlations, for sample size N are extremely high leading to much less reliable inference.

In practice, if only the vector \mathbf{m} is of interest, far fewer samples are needed which would significantly speed up the algorithm. Physical timings for runs of these and other algorithms are given in Table 2. These show that as executed, this algorithm required about 11-12 hours to run. If only every 10th sample were collected, 50,000 total samples, the algorithm would have required approximately 1 hour.

In order to understand the behavior of the proportions, we compared various estimates of the m_i . These include posterior means from the Gibbs sampler, standard MLE's, corrected



3.1.2 Sidebars

Sidebars are often used for small additional explanations, reminders etc. All normal sidebars are Text.

In the first example, the sidebar is in a larger font, but it is not a heading for the text, rather an accompanying citation, so “Text” is the correct label.

Undiscovered country

CEOs say we are at a watershed moment. Technological advances are creating massive upheavals, with industries converging and new ecosystems emerging as never before. So how are the trailblazers guiding their organizations through this turmoil?

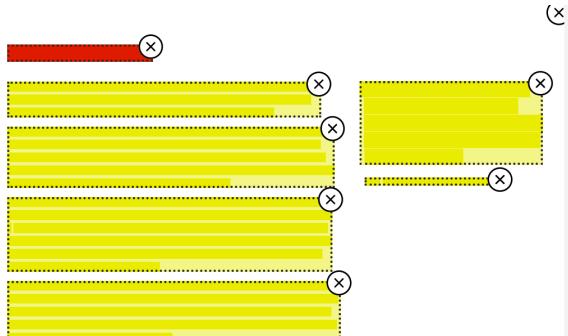
In the first installment of our latest C-suite Study, we interviewed 5,247 top executives to find out what they think the future will bring and how they're positioning their organizations to prosper in the “age of disruption.” This report probes more deeply into the perspectives of the 816 CEOs who contributed to our research. We've also focused on what the CEOs of the world's most successful enterprises in this study do differently.

We identified a small group of organizations that have both a strong reputation as leading innovators and a superior financial track record. The torchbearers, as we call them, comprise 4 percent of all enterprises tracked by CEO.com. In addition, we've identified a group of organizations that lag behind the rest. Market Followers have a much lower market profile in the opinion of the CEOs who head them, and almost all are far less financially successful. They account for 26 percent of our sample.

Comparing the two groups reveals significant variations in their strategic priorities, go-to-market plans and organizational approaches—all areas where the CEO has great sway. It also shows what CEOs in enterprises that are struggling can do to strengthen their positions. And, as Schumpeter's “gale of creative destruction” bows around the globe with unprecedented force, CEOs will need all the advantages they can get!

“New competitors, many of them from outside the industry, are entering the picture. What they're doing will disrupt the market and our customer base.”

CEO, Banking and Financial Markets, Canada



Another common example of sidebars is in US laws, e.g., here:

eriting “a covered”.
2 USC 2051 note.

Office of Personnel
to carry out this

romulgated under
Deadline.
an 180 days after

2 USC 2051 note.

the terms “con-
cess contract” have
l(a) of Public Law



3.1.3 Authors, Affiliations, and Abstracts

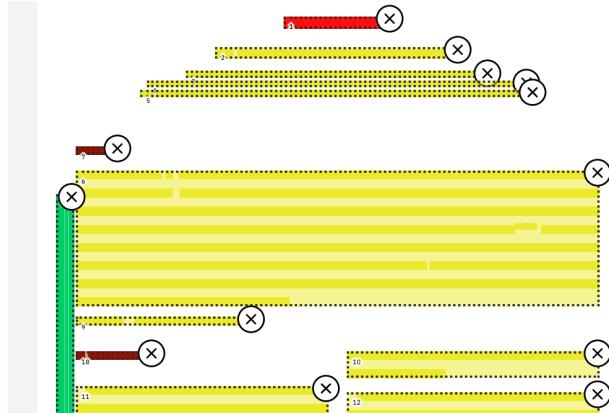
The next example shows that authors, affiliations (author’s employer and/or address), and abstracts are treated as Text:²

Minimal Inflation
Luis Álvarez-Gaumé^b, César Gómez^{b,a}, Raúl Jimenez^{a,c}
^aTheory Group, Physics Department, CERN, CH-1211, Geneva 23, Switzerland.
^bInstituto de Física Teórica UAM/CSIC, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.
^cICREA # Institute of Sciences of the Cosmos (ICC), University of Barcelona, 08028 Barcelona, Spain.

Abstract
Using the universal X superfield that measures in the UV the violation of conformal invariance we build up a model of multifield inflation. The underlying dynamics is the one controlling the natural flow of this field in the IR to the Goldstino superfield once SUSY is broken. We show that flat directions satisfying the slow roll condition exist only if R-symmetry is broken. Naturalness of our model leads to scales of SUSY breaking of the order of 10^{11-12} GeV, a nearly scale-invariant spectrum of the initial perturbations and negligible gravitational waves. We obtain that the inflaton field is lighter than the gravitino by an amount determined by the slow roll parameter η . The existence of slow-roll conditions is directly linked to the values of supersymmetry and R-symmetry breaking scales. We make cosmological predictions of our model and compare them to current data.
Key words: SUSY; cosmology; inflation

30 Dec 2009 [hep-th]

0v1 1. Introduction
In spite of the enormous success of inflationary cosmology [1, 2, 3, 4, 5, 6, 7, 8, 9] at describing the observed
of new problems appear [24], and we will discuss
some of them later.
Current observational constraints from CMB tempera-
ture and polarization experiments and large-scale struc-



Here you also see (as in several other examples) that paragraphs that continue at the top of a page or column become a cluster of their own, as we only use simple rectangular clusters.

3.1.4 Quotes

The following example shows quotes:

Plaut and Roughgarden [27] show two scenarios for which EFX allocations are guaranteed to exist: (i) All agents have identical valuations (i.e., $v_1 = v_2 = \dots = v_n$), and (ii) Two agents (i.e., $n = 2$). Unfortunately, starting from three agents, even for the well studied class of additive valuations, it is open whether EFX allocations exist. Plaut and Roughgarden [27] also remark that:

“The problem seems highly non-trivial even for three players with different additive valuations.”

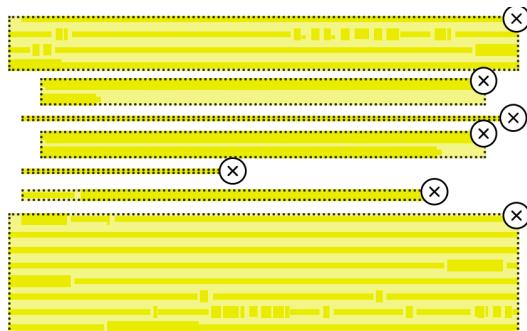
Furthermore, it is also suspected in [27] that EFX allocations may not exist in general settings:

“We suspect that at least for general valuations, there exist instances where no EFX allocation exists, and it may be easier to find a counterexample in that setting.”

Contrary to this suspicion, we show that

THEOREM *EFX allocations always exist for three agents with additive valuations.*

EFX with charity: Quite recently there have been studies [9, 13] that consider relaxations of EFX, called “EFX with charity”. Here we look for partial EFX allocations, where not all items need to be allocated (some of them remain unallocated). There is a trivial such allocation where no item is allocated to any agent. Therefore, the goal is to determine allocations with some qualitative or quantitative bound on the set of unallocated items. For instance, Chaudhury et al. [13] show how to determine a partial EFX allocation X and a pool of unallocated items P such that no agent envies the pool (i.e. for any agent i , we have $v_i(X_i) \geq v_i(P)$), and P has less than n items (i.e., $|P| < n$), even in the case of general valuations. In case of additive valuations, Caragiannis et al. [9] show the



3.1.5 Mathematical Theorems etc.

The previous example also contains a mathematical theorem. As the introductory word “Theorem” is in the same line with the normal text, this is also a “Text” cluster. The same holds for “Lemma”, “Definition”, “Proof” etc. Here is another, more visible theorem:

² Page 2f4cf34dd3344fbab47f75fb0a08b4f8cce64a6044c4d0fcedd08b16eac7c283

the posterior means of the m_i parameters. An exact calculation is possible based upon the marginalization properties of the Dirichlet (Bilodeau & Brenner, 1999, Corollary 3.4).

Theorem 4.1 Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p, \theta_{p+1}) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_p, \alpha_{p+1})$. Let $\alpha_0 = \sum_{i=1}^{p+1} \alpha_i$. Suppose that $S = \sum_{i=1}^p \theta_i$, and for $j < p$, let $\boldsymbol{\theta}_j = (\theta_1, \dots, \theta_j)$ be any subset of components of $\boldsymbol{\theta}$. It follows that $\delta_1 = \theta_1/S \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_j, \delta)$ with $\alpha_0 - \alpha_{p+1} = \delta + \sum_{d=1}^j \alpha_d$

This result demonstrates that, ignoring the missing categories, the distribution of the



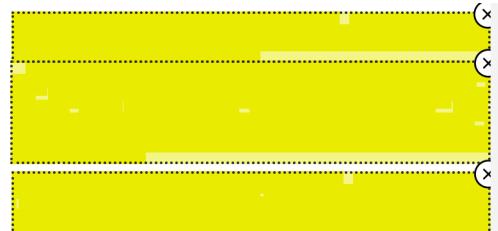
3.1.6 Emphasized Paragraph Start: Text

If the start of a paragraph is emphasized, e.g., bold or in italics, it still belongs into the “Text” cluster. It does not become a Section-header, because a Section-header must be on a separate line.

Sensitivity of the univariate analysis to the number of patients. For the four cohorts, the performance of each feature was determined using the area under the ROC curves and the Youden index given with its corresponding sensitivity (Se) and specificity (Sp) values. The optimal cut-off to separate the groups was defined as the cut-off value that maximized YI.

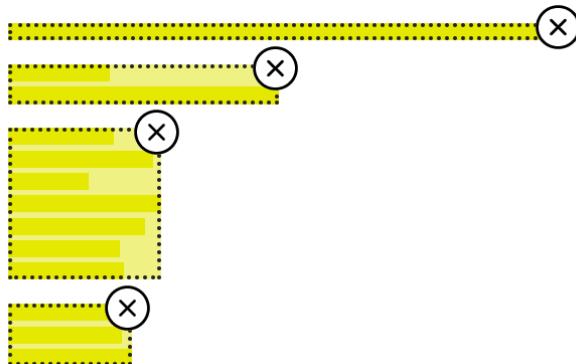
For each run (i) of each set up, we performed a ROC study for each feature using the training and validation set (TRS + VS) to define the optimal cut-off as the one maximizing the YI obtained on (TRS + VS), called $Y_{\text{TRS} + \text{VS}}^{\text{optimal}}$. $Y_{\text{TRS} + \text{VS}}^{\text{optimal}}$ was compared to the Youden index obtained on the test set (TES), $Y_{\text{TES}}^{\text{optimal}}$, using that cut-off. The mean of the fifty $Y_{\text{TRS} + \text{VS}}^{\text{optimal}}$ and $Y_{\text{TES}}^{\text{optimal}}$ values were defined as $Y_{\text{AVG}}^{\text{optimal}}$ and $Y_{\text{AVG}}^{\text{optimal}}$ respectively. The behavior of $Y_{\text{TRS} + \text{VS}}^{\text{optimal}}$ depending on the number of patients was studied as well as the difference between the predicted performance $Y_{\text{TRS} + \text{VS}}^{\text{optimal}}$ and the actual performance observed on test data $Y_{\text{AVG}}^{\text{optimal}}$, using the cut-off determined from the training data. The process was repeated independently for each radiomic feature and is illustrated in Fig. 1a. The AUC computed from TRS + VS were also given.

Sensitivity of the multivariate analysis to the number of patients. For the four cohorts, we studied three types of classifier implemented in scikit learn toolbox³: LR, SVM with a linear kernel, and LASSO. For these 3 classifiers, the following parameters were tuned: inverse of regularization strength for LR, penalty parameter for SVM and constant that multiplies the L1 term for LASSO. LR uses a logistic function to model a binary dependent variable, SVM identifies the hyper-plane that best differentiates the two classes and LASSO is a regression analysis



3.1.7 Multiple Short Lines without List Identifiers that Clearly Belong Together

If there are multiple short lines directly following one another and belonging together, as in an address or a verse of a poem, they are considered one Text cluster, as in the following example:



“Text” clusters are also a good choice for the following example:

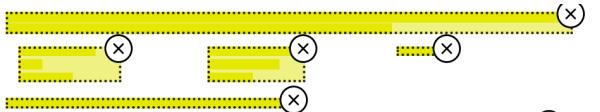
The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo) IBM®
IBM Cloud®

IBM Elastic Storage®
IBM Spectrum®
POWER®

Redbooks®

The following terms are trademarks of other companies:



3.1.8 Text around a Centered Formula or Citation

If text contains a centered formula, the texts before and after the formula become two different Text clusters, even if a sentence continues after the formula, as in the following example:

Evaluation of the quality of the prediction - Univariate models. To assess the accuracy of the predicted \bar{Y} when using univariate analysis, we tested 17 experimental conditions for each feature with an unbalanced number of patients (from 36 to 164) for the real data of cohort 1 and for synthetic cohort 1. N starts from 36 to have at least 3 points to perform the fits ($N=8$ patients). Similarly, we tested 19 experimental conditions for each feature with unbalanced number of patients (from $N=19$ to 73) for real cohort 2 and for synthetic cohort 2. A quality factor QF1 (Eq. 5):

$$QF1 = |\bar{Y}_{EUC}^{TBS+VS}(N) - \bar{Y}_{EUC}^{TBS+VS}(N_{av})| \quad (5)$$

was computed to determine how close the prediction $\bar{Y}_{EUC}^{TBS+VS}(N)$ was from $\bar{Y}_{EUC}^{TBS+VS}(N_{av})$ obtained with the largest number of patients N_{av} ($N_{av}=414$ for cohort 1 and 569 for cohort 2) when stability was reached. As a comparison, a second quality factor QF2 (Eq. 6):

$$QF2 = |\bar{Y}_{EUC}^{TBS+VS}(N) - \bar{Y}_{EUC}^{TBS+VS}(N_{av})| \quad (6)$$

was computed to determine how close the performance obtained with the N patients $\bar{Y}_{EUC}^{TBS+VS}(N)$ but without downscaling was from $\bar{Y}_{EUC}^{TBS+VS}(N_{av})$, as $\bar{Y}_{EUC}^{TBS+VS}(N)$ is usually used as an estimate of the expected performance. Having $QF1 < QF2$ shows the usefulness of the downscaling approach and the closer $QF1$ to 0, the more accurate the prediction.



3.1.9 Code

Some types of documents contain computer code or mathematical algorithms. We treat those as Text, as in the following example. Similar to Section 3.1.7, we keep multiple code lines together even if they are short, unless a significant gap occurs.

1. Copy the boot loader file (.img) and ONLY the boot loader file to the ASA through normal means, usually File Transfer Protocol (FTP), as in this example:

```
ciscoasa# copy ftp://</FTP_SERVER>/asasfr-5500x-boot-6.2.0-2.img  
disk0:/asasfr-5500x-boot-6.2.0-2.img
```

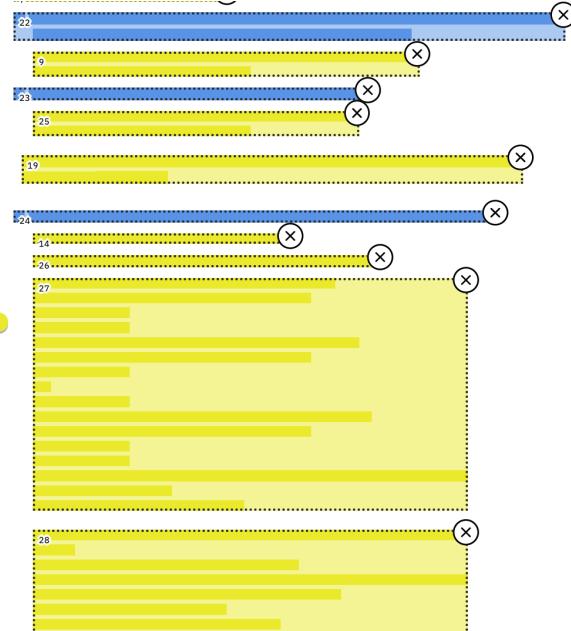
2. Specify the location of the boot loader image to the ASA.

```
ciscoasa# sw-module module sfr recover configure image  
disk0:/asasfr-5500x-boot-6.2.0-2.img
```

Note: If you want to see what happens during the next step, you can activate **debug module-boot** at this time.

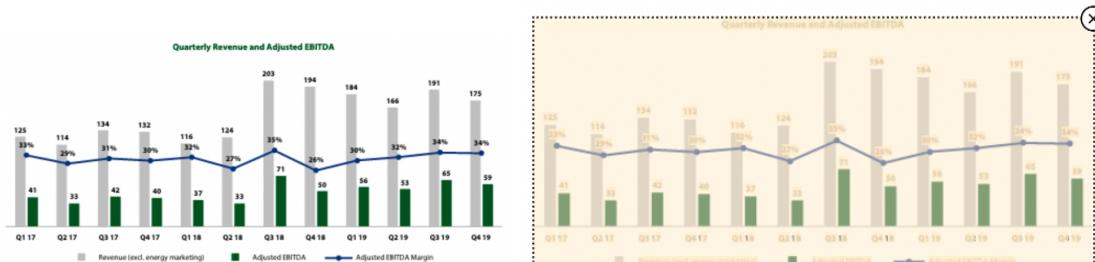
3. Load the boot image and create the environment where the SFR module runs.

```
ciscoasa# sw-module module sfr recover boot  
If you turned on debugging you'll see some output like this:  
Mod-sfr 788> *** EVENT: Creating the Disk Image...  
Mod-sfr 789> *** TIME: 05:50:26 UTC Jul 1 2014  
Mod-sfr 790> ***  
Mod-sfr 791> ***  
Mod-sfr 792> *** EVENT: The module is being recovered.  
Mod-sfr 793> *** TIME: 05:50:26 UTC Jul 1 2014  
Mod-sfr 794> ***  
...  
Mod-sfr 795> ***  
Mod-sfr 796> *** EVENT: Disk Image created successfully.  
Mod-sfr 797> *** TIME: 05:53:06 UTC Jul 1 2014  
Mod-sfr 798> ***  
Mod-sfr 799> ***  
Mod-sfr 800> *** EVENT: Start Parameters: Image: /mnt/disk0/vm/vm_3.img,  
ISO: -cdrom /mnt/disk0  
<DETAILS REMOVED TO CONSERVE SPACE>  
  
Mod-sfr 239> Starting Advanced Configuration and Power Interface daemon:  
acpid.  
Mod-sfr 240> acpid: starting up with proc fs  
Mod-sfr 241> acpid: opendir(/etc/acpi/events): No such file or directory  
Mod-sfr 242> starting Busybox inetd: inetd... done.  
Mod-sfr 243> Starting ntpd: done  
Mod-sfr 244> Starting syslogd/klogd: done  
Mod-sfr 245>
```



3.2 Picture

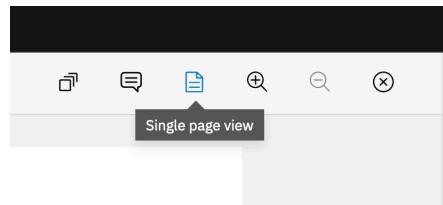
A “Picture” is a graphic, plot, photograph, or logo. It may also contain text, but the text would not mean much without its graphical surroundings. For instance, the following is one picture, because the bars and lines are key information, and the text only explains them.



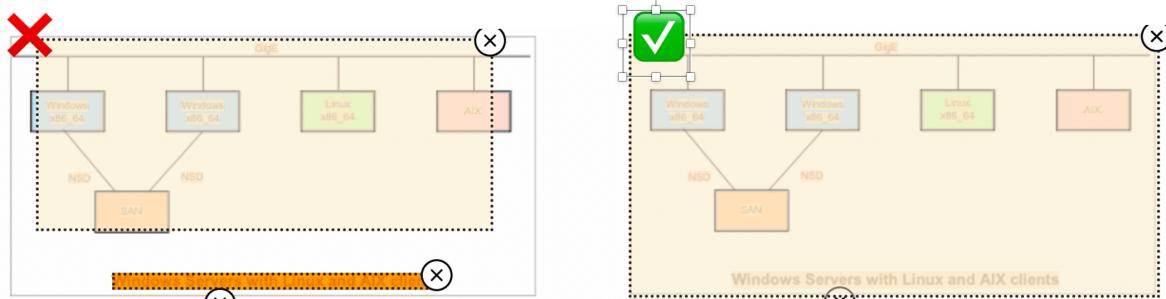
For graphics, draw the smallest boundary that comprises all elements (graphical and textual) that belong to the graphics. Rather include a few pixels too many than too few.

3.2.1 Use Single-Page View

To allow for accurate picture annotation, please switch to “single page view”.

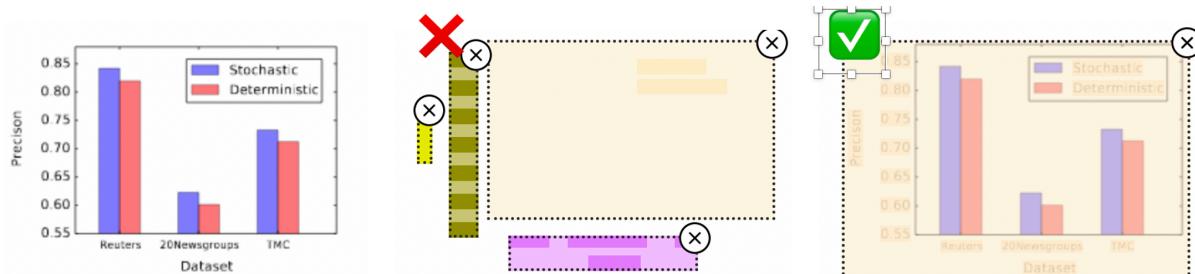


Otherwise it can easily happen that one does not capture the picture precisely:



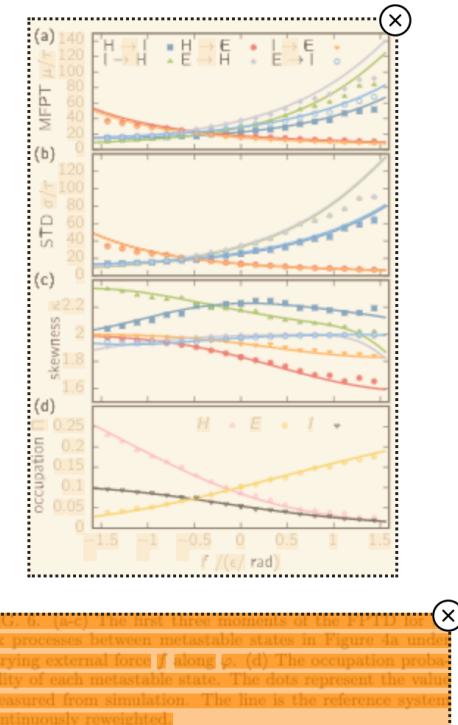
3.2.2 Legends

Legends of a diagram belong to the Picture, i.e., texts explaining the axes, colors, measurement units etc. The picture would not make sense without its legends, nor the legends without the picture.



3.2.3 One Picture per Caption, if Captions Exist

Below is a “Picture” cluster, together with its caption. The fact that there is only one caption (starting “FIG. 6.”) determines that there is only one picture, even if it has several sub-pictures.

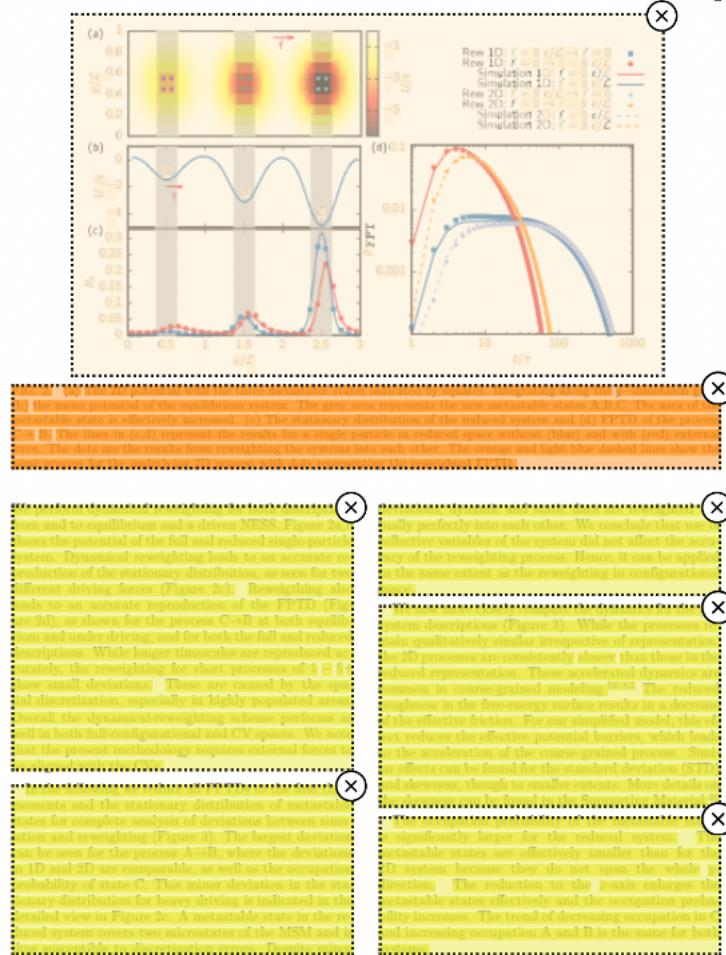


3.2.4 Multiple Small Pictures

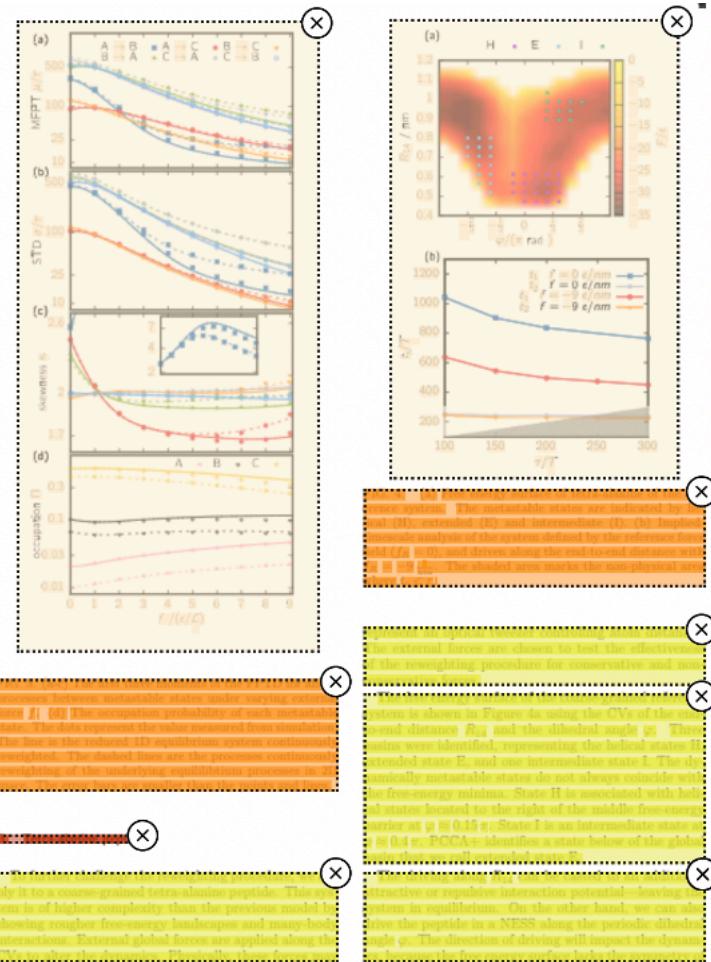
Sometimes several small pictures are close together. Even the figure just above can be seen as such a case, with 4 sub-pictures.

- If there is a single caption for them, they should become one cluster, as above.
- If each has its own caption, then each should be a separate picture.
- If there is nothing like captions around, it is a judgement call if they best belong together or not, e.g., by whether they touch, or whether they form a reasonable union rectangle.

Here is a whole-page example of how sub-pictures are grouped together along the single Caption:



The next example has two Captions, and thus two Pictures:



3.2.5 Forms

Some government forms, questionnaires, and signature pages contain irregular small text elements, mixed with checkboxes, small fields to fill, small pictures or hand-written signatures. We treat these as Pictures, unless there is a clear columns-and-rows structure as needed for a Table.

In the following example, we decided to make the upper part a “Picture” because of its empty lines and checkboxes. We then decided to take the frame as the picture boundary. However, with the lower frame on the page, we used “Text” and “Section-header” because there is no graphic in that frame. Such forms pages usually leave room for individual judgement.

4. Technische und berufliche Leistungsfähigkeit

Vorlage geeigneter Referenzen über die Ausführung von Bauleistungen in den letzten 5 Kalenderjahren**, die mit der zu vergiebenden Leistung vergleichbar sind.

Als vergleichbare Leistungen werden anerkannt:

.....
.....
.....

** Der Auftraggeber akzeptiert auch Referenzen, welche mehr als fünf Jahre zurückliegen.

1. Referenz: Bezeichnung der Leistung, des Auftragswertes des auf mein/unser Unternehmen entfallenden Anteils, des Ausführungszeitraums und des Auftraggebers:
.....
.....
.....

2. Referenz: Bezeichnung der Leistung, des Auftragswertes des auf mein/unser Unternehmen entfallenden Anteils, des Ausführungszeitraums und des Auftraggebers:
.....
.....
.....

3. Referenz: Bezeichnung der Leistung, des Auftragswertes des auf mein/unser Unternehmen entfallenden Anteils, des Ausführungszeitraums und des Auftraggebers:
.....
.....
.....

Es können auch mehr als drei Referenzen angegeben werden, diese sind dann auf gesonderte Anlage vorzunehmen.

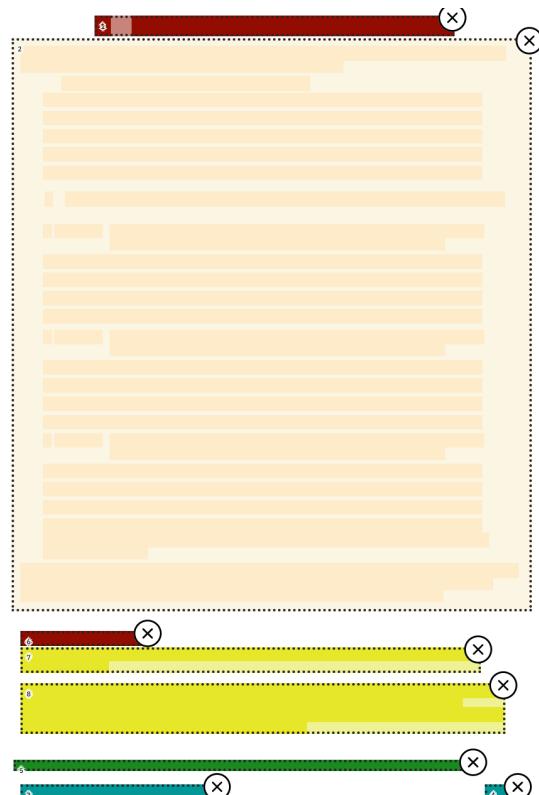
Falls mein(e)unser(e) Bewerbung/Angebot in die engere Wahl kommt, werde ich /werden wir für die oben genannten Leistungen Bescheinigungen über die ordnungsgemäße Ausführung und das Ergebnis in Anlehnung an beiliegendes Muster auf gesonderte Verlangen vorlegen.

Angabe zu Arbeitskräften
Ich/Wir erkläre(n), dass mir/uns die für die Ausführung der Leistung erforderlichen Arbeitskräfte zur Verfügung stehen.

Falls mein(e)unser(e) Bewerbung/Angebot in die engere Wahl kommt, werde ich /werden wir auf gesonderte Verlangen die Zahl der in den letzten drei abgeschlossenen Kalenderjahren jahresdurchschnittlich beschäftigten Arbeitskräfte, gegliedert nach Lohngruppen und gesondert ausgewiesenen technischen Leitungspersonal angeben.

** Vom Auftraggeber anzukreuzen, wenn ausnahmsweise Referenzen akzeptiert werden, die mehr als 5 Jahre zurückliegen.

HVA B-SIB Eigenerklärung Eignung 08-19
Seite 3



If a page mixes text and pictures that are worth looking at individually, then label them individually as Text, Picture, etc., as in the following example:



3.3 Caption

A Caption is text introducing a picture or table, outside the normal text flow, but not part of the picture or table itself.

3.3.1 Normal Captions

In scientific texts a caption is usually introduced by a label like “Figure 1” or “Fig. 4” or “Table 5”. In other material, it might be “Upper left: xxx” or a short text clearly referring to the picture, e.g., a separate line saying “CEO XYZ talking to children.”

- A caption is usually only one paragraph long.
- Examples were already shown together with the Pictures in the previous section.

Often, a Caption uses a slightly different font than normal text. Commonly, normal paragraphs flow around it. For instance, after a picture with caption at the top of a page, a normal paragraph from the previous page may be continued.

3.4 Section-header

The label “Section-header” is used for all types of headings and section titles, except for the overall title of a document.

3.4.1 Normal Headers

Section-headers are typically larger than normal text. In some cases, they are instead in a bold font, or written with capital letters. There should always be some visible emphasis.

Section-headers are commonly shorter than 1 line, but sometimes 2- or even 3-line headers exist. They are usually incomplete sentences.

In scientific or legal texts, Section-headers also tend to start with an identifier, like “1.3” or “A.”, but this is not always the case.

Sometimes there are several headers directly following after each other, in particular if they are two different hierarchy levels as in the following example. You should make each of them a separate “Section-header” cluster.

ne peptide
ach amino
the back-
e field for
air poten-
teractions,
nteraction
tions were

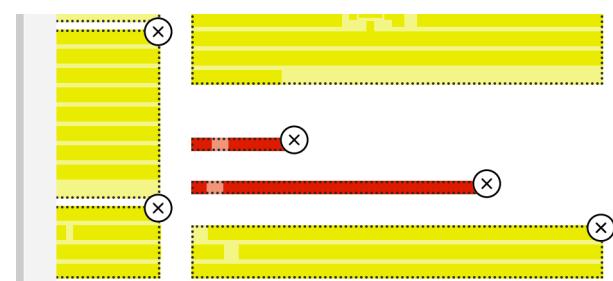
he end-to-
, (see Fig-
m is called
f constant

expectation value of $\langle \frac{e^2}{r} \rangle$. These moments are used
to compare FPTDs throughout the paper to capture the
main features and draw physical information from the
distribution.

IV. RESULTS

A. Particle in a two-dimensional potential

We first consider a toy model: a particle in a multi-
well. The system is originally in two dimensions, but
we also consider a reduced one-dimensional description.

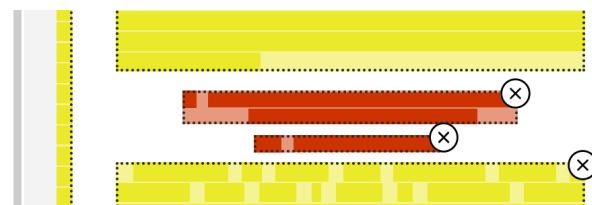


Usually it is quite clear where a Section-header cluster ends, also if a header spans over several lines like here:³

future research we also include seven rapid rotators in
Table 3 for which we have been unable to derive RVs
from our spectra.

3. EFFECTS OF STELLAR ROTATION ON
MEASUREMENT PRECISION
3.1. *Observed Rotation*

Because of its youth, M35 provides a sample of
late-type stars with a range of rotational periods



There is the start of Section 3, with a header over 2 lines, and inside that is the start of Section 3.1.

³ Page dca73946b2263c098f10b6b3ecde6aa05673b27126e0bd066c100d4d126dfec

3.4.2 Emphasized Text at Start of a Line: Normal Text

If there is bold or capitalized text (or even in larger font), but normal text starts in the same line and continues as a paragraph, do not consider it as a header. The bold part becomes part of the “Text” cluster. In the following picture, this is the case for the bold “Fully convolutional networks”:

2. Related work

Our approach draws on recent successes of deep nets for image classification [22, 34, 35] and transfer learning [5, 41]. Transfer was first demonstrated on various visual recognition tasks [5, 41], then on detection, and on both instance and semantic segmentation in hybrid proposal-classifier models [12, 17, 15]. We now re-architect and fine-tune classification nets to direct, dense prediction of semantic segmentation. We chart the space of FCNs and situate prior models, both historical and recent, in this framework.

Fully convolutional networks To our knowledge, the idea of extending a convnet to arbitrary-sized inputs first appeared in Matan *et al.* [28], which extended the classic LeNet [23] to recognize strings of digits. Because their net was limited to one-dimensional input strings, Matan *et al.*

- multi-scale p
- saturating tan
- ensembles [3,
- whereas our met
- we do study pa
- dense output [3,
- discuss in-netw
- nected predictio

Unlike these classification are p

pervised pre-tra

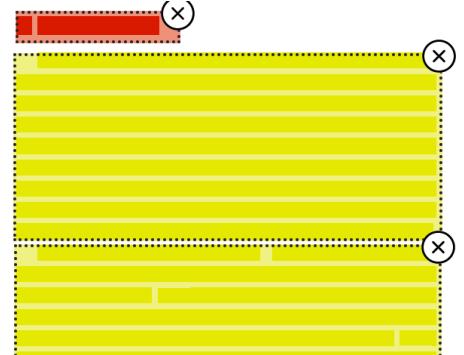
learn simply an

whole image gr

Hariharan *et* deep classificat

so in hybrid pr

fine-tune an R-



Here is a second example with both a Section-header and a bold non-header:

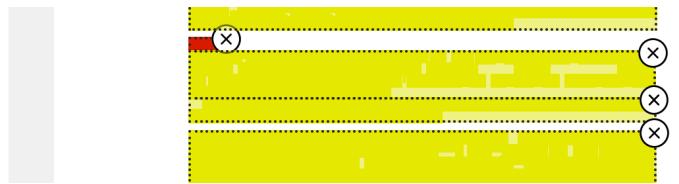
pared Y_{FCN} (84) and $Y_{\text{FCN}}^{\text{c}}(88)$ to $Y_{\text{FCN}}^{\text{c}}(414)$ obtained using a LASSO model built from the whole cohort of 414 patients. We repeated exactly the same analysis using AUC as a figure of merit.

Results

For cohort 1¹, the original study reported a YI of 0.73 ± 0.06 ($\text{Se} = 0.89 \pm 0.02$, $\text{Sp} = 0.84 \pm 0.04$) and AUC = 0.92 ± 0.02 for the direct LDA approach and 0.74 ± 0.06 ($\text{Se} = 0.89 \pm 0.02$, $\text{Sp} = 0.85 \pm 0.04$) and AUC = 0.91 ± 0.02 for the backward LDA approach. For cohort 2², a YI of 0.91 ± 0.01 ($\text{Se} = 0.94 \pm 0.00$ and $\text{Sp} = 0.98 \pm 0.00$) was reported with a logistic regression approach.

For the sake of conciseness, the results are detailed for cohort 1 and only summarized for cohort 2, with all detailed results for cohort 2 provided as “Supplemental Data”.

Sensitivity of the univariate analysis to the number of patients. The best performing feature that reached stability was SUVmin for cohort 1 with $Y_{\text{FCN}}^{\text{c},15} \pm SY_{\text{FCN}}^{\text{c},15} = 0.50 \pm 0.02$ ($\text{Se} = 0.71 \pm 0.11$, $\text{Sp} = 0.81 \pm 0.05$) and AUC = 0.79 ± 0.01 . The YI obtained on TRS + VS ($Y_{\text{FCN}}^{\text{c},15}$) and on TES ($Y_{\text{FCN}}^{\text{c}}$) as a function of the experimental conditions as well as the AUC are shown in Fig. 2 for cohort 1 and synthetic cohort 1 for



3.5 Footnote

3.5.1 Normal Footnotes

A footnote is almost always written in smaller text, at the bottom of a page, and below a horizontal line. It also usually has a superscript identifier by which it is referenced somewhere on the same page. It does not repeat on every page. Here are two examples:

Maximum Caliber (MaxCal) methods.¹⁷ A method similar to the Rosenbluth algorithm performs reweighting in NESS for minimal processes like birth-death processes.¹⁸ Another method based on iterative trajectory weighting is expected to scale to NESS systems,¹⁹ however these

tions made on a system entered on the system will affect the pathway. The microscopic change is described by the NESS system is able to preserve probabilities to be separated into two parts: dissipative contributions²⁰ is determined by the target

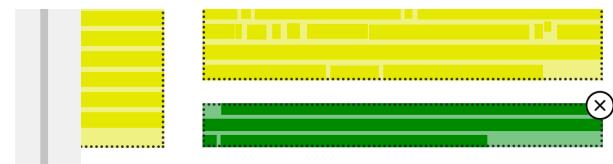
^a) Electronic mail: baume@mpip-mainz.mpg.de



sions fully with casts ut of forma-

input x pixels to the right and y pixels down, once for every (x, y) s.t. $0 \leq x, y < f$. Process each of these f^2 inputs, and interlace the outputs so that the predictions correspond to the pixels at the *centers* of their receptive fields.

¹Assuming efficient batching of single image inputs. The classification scores for a single image by itself take 5.4 ms to produce, which is nearly 25 times slower than the fully convolutional version.



3.5.2 Several Footnotes on One Page: Several Clusters

There can be several footnotes on one page, usually below only one horizontal line, but each with its own identifier. Then each one is a separate “Footnote” cluster. Here is an example with 3 footnotes:

achieved $\sim 75\%$ of state-of-the-art performance. The segmentation-equipped VGG net (FCN-VGG16) already

³Using the publicly available CaffeNet reference model.

⁴Since there is no publicly available version of GoogLeNet, we use our own reimplementation. Our version is trained with less extensive data augmentation, and gets 68.5% top-1 and 88.4% top-5 ILSVRC accuracy.

⁵Using the publicly available version from the Caffe model zoo.



Please keep in mind that a “Page-footer” is not a footnote; see more details on that below.

3.6 Formula

A formula is typically a sequence of mathematical symbols, centered on a line of its own.

3.6.1 Normal Formulas

The first example is a text with two formulas:

ing layer, these functions compute outputs y_{ij} by

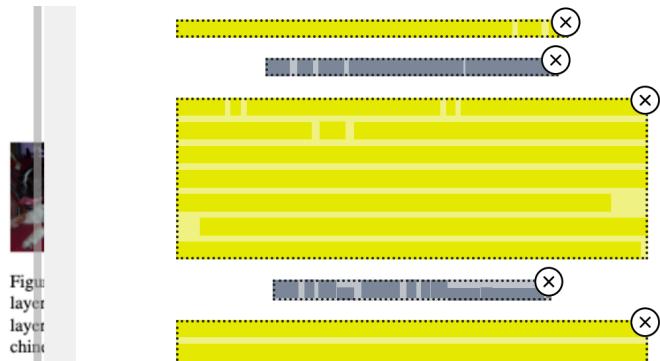
$$y_{ij} = f_{ks}(\{\mathbf{x}_{si+\delta i, sj+\delta j}\}_{0 \leq \delta i, \delta j \leq k})$$

where k is called the kernel size, s is the stride or subsampling factor, and f_{ks} determines the layer type: a matrix multiplication for convolution or average pooling, a spatial max for max pooling, or an elementwise nonlinearity for an activation function, and so on for other types of layers.

This functional form is maintained under composition, with kernel size and stride obeying the transformation rule

$$f_{ks} \circ g_{k's'} = (f \circ g)_{k'+(k-1)s', ss'}$$

While a general deep net computes a general nonlinear function, a net with only layers of this form computes a



Sometimes there is a formula number on the right; this belongs to the same formula cluster, like below:

ast 3 points to perform the fits ($S = 8$ patients). Similarly, we tested 19 experimental condi-
with unbalanced number of patients (from $N=19$ to 73) for real cohort 2 and for synthetic
ctor QF1 (Eq. 5):

$$QF1 = |Y_{ROC}(N) - Y_{ROC}^{TRS+VS}(N_{tot})| \quad (5)$$

etermine how close the prediction $Y_{ROC}(N)$ was from $Y_{ROC}^{TRS+VS}(N_{tot})$ obtained with the largest
 N_{tot} ($N_{tot} = 414$ for cohort 1 and 569 for cohort 2) when stability was reached. As a compari-
factor QF2 (Eq. 6):

$$QF2 = |Y_{ROC}^{TRS+VS}(N) - Y_{ROC}^{TRS+VS}(N_{tot})| \quad (6)$$

etermine how close the performance obtained with the N patients $Y_{ROC}^{TRS+VS}(N)$ but without
om $Y_{ROC}^{TRS+VS}(N_{tot})$, as $Y_{ROC}^{TRS+VS}(N)$ is usually used as an estimate of the expected performance.
ows the usefulness of the downsampling approach and the closer QF1 to 0, the more accu-



Note that mathematical symbols in-line with normal text are not Formula clusters, e.g., the equation “... = 414” above.

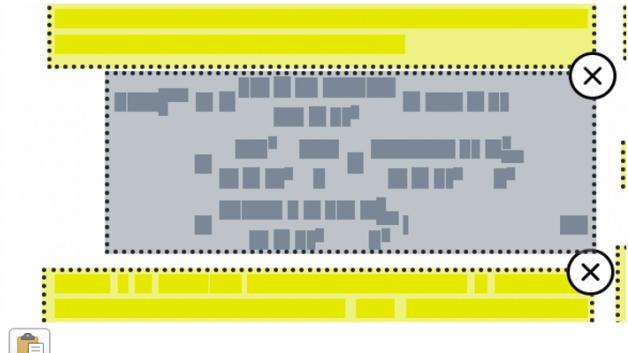
3.6.2 Several Successive Formula Lines

Sometimes there are several successive formula lines in a block. Here, one has to decide whether it is one long formula, or several individual formulas. The following example is only one formula and thus one cluster:

An analytic expression for this solvation free energy was derived by Chen and Weeks (CW)⁷⁰,

$$\begin{aligned}\beta\Delta\mu_v^{\text{CW}} = & -\frac{\eta(2-7\eta+11\eta^2)}{2(1-\eta)^3} - \ln(1-\eta) \\ & + \frac{18\eta^3}{(1-\eta)^3} \frac{R_{\text{HS}}}{\sigma} - \frac{18\eta^2(1+\eta)}{(1-\eta)^3} \frac{R_{\text{HS}}^2}{\sigma^2} \\ & + \frac{8\eta(1+\eta+\eta^2)}{(1-\eta)^3} \frac{R_{\text{HS}}^3}{\sigma^3},\end{aligned}\quad (9)$$

where $\eta = \pi\rho_B\sigma^3/6$ is the packing fraction, σ is the solvent hard core diameter, and R_{HS} is the hard sphere



One can recognize this by the single formula number (9) and by the layout, with the + symbols at the beginning of the second and third line aligned with the “=” in the first line.

An opposite example, with two separate formulas, is shown next:

$C_n(N)$ was assessed using metrics similar to the ones defined for the univariate analysis
!F2' (Eq. 8):

$$QF1' = |Y_{C_n}(N) - Y_{C_n}^{\text{VS}}(N_{\text{tot}})| \quad (7)$$

$$QF2' = |Y_{C_n}^{\text{VS}}(N) - Y_{C_n}^{\text{VS}}(N_{\text{tot}})| \quad (8)$$



There is no symbol like “=” or “+” linking the lines, and they have different formula numbers.

3.7 Table

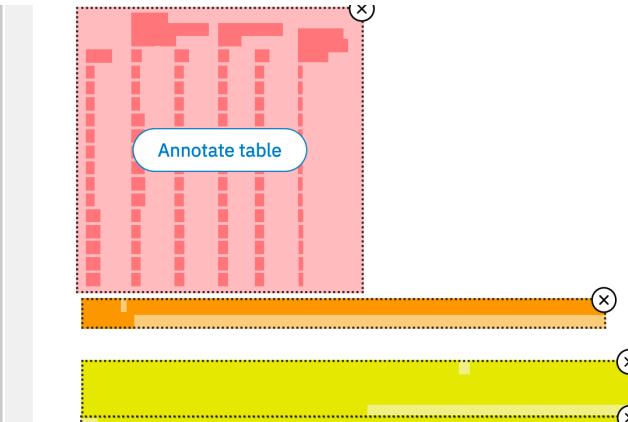
A table is text arranged in a grid with columns and rows, where this arrangement is important for understanding. Often there are gridlines separating the columns or rows or both, but there exist many tables without lines.

3.7.1 Normal Tables

The first example shows a table with gridlines in both directions.

| Set-up | Training set + Validation set (TRS + VS) | | External testing (ETS) | | Number of folds for the SKFCV |
|--------|--|-----------|------------------------|-----------|-------------------------------|
| | PL | MT | PL | MT | |
| S1 | 30 | 10 | 30 | 10 | 2 |
| S2 | 60 | 20 | 30 | 10 | 2 |
| S3 | 90 | 30 | 30 | 10 | 3 |
| S4 | 120 | 40 | 30 | 10 | 3 |
| S5 | 150 | 50 | 30 | 10 | 3 |
| S6 | 180 | 60 | 30 | 10 | 3 |
| S7 | 210 | 70 | 30 | 10 | 3 |
| S8 | 240 | 80 | 30 | 10 | 3 |
| S9 | 270 | 90 | 30 | 10 | 3 |
| S10 | 30 | 30 | 30 | 10 | 3 |
| S11 | 45 | 45 | 30 | 10 | 4 |
| S12 | 60 | 60 | 30 | 10 | 6 |
| S13 | 75 | 75 | 30 | 10 | 7 |
| S14 | 90 | 90 | 30 | 10 | 9 |

Table 1. Number of patients for each set-up of cohort 1 and synthetic cohort 1. Balanced situations in bold characters.



The table caption (here starting “Table 1”) must not be part of the table cluster.

As with Pictures, the Table cluster should be drawn tightly around the table, but include the complete gridlines and text. This is best done in the single-page view, as shown here:

Table 2: Comparison of skip FCNs on a subset⁴ of PASCAL VOC 2011 segval. Learning is end-to-end, except for FCN-32s-fixed. Where only the last layer is fine-tuned. Note that FCN-32s is FCN-16, renamed to highlight stride.

| | pixel acc. | mean acc. | mean IU | I.w. IU |
|-----------|---------------|--------------|------------|------------|
| FCN-32s-f | | | 54.4 | 72.0 |
| FCN-16s | | | 60.4 | 81.4 |
| FCN-16s | 90.0 | 75.7 | 62.4 | 83.0 |
| FCN-8s | 90.3 | 75.9 | 62.7 | 83.2 |

Maintain its receptive field size. In addition to their computational cost, we had difficulty learning such large filters. We attempted to re-architect the layers above pool15 with

The following tables have only horizontal gridlines, but one can clearly see the structure of rows and columns from the arrangement.

| Model | AG | Sogou | DBP | Yelp P. | Yelp F. | Yah. A. | Amz. F. | Amz. P. |
|-----------------------------------|------|-------|------|---------|---------|---------|---------|---------|
| BOW (Zhang et al., 2015) | 83.8 | 92.2 | 96.6 | 92.2 | 54.0 | 68.9 | 54.6 | 90.4 |
| agnews (Zhang et al., 2015) | 92.0 | 97.1 | 99.6 | 99.6 | 56.0 | 68.5 | 54.3 | 92.0 |
| agnews TFIDF (Zhang et al., 2015) | 92.4 | 97.2 | 98.7 | 95.4 | 54.8 | 68.5 | 52.4 | 91.5 |
| char-CNN (Zhang and LeCun, 2015) | 87.2 | 95.1 | 98.3 | 94.7 | 62.0 | 71.2 | 59.5 | 94.5 |
| char-CRNN (Xiao and Cho, 2016) | 91.4 | 95.2 | 98.6 | 94.5 | 61.8 | 71.7 | 59.2 | 94.1 |
| VDCNN (Conneau et al., 2016) | 91.3 | 96.8 | 98.7 | 95.7 | 64.7 | 73.4 | 63.0 | 95.7 |
| fastText, $h = 10$ | 91.5 | 93.9 | 98.1 | 93.8 | 60.4 | 72.0 | 55.8 | 91.2 |
| fastText, $h = 10$, bigram | 92.5 | 96.8 | 98.6 | 95.7 | 63.9 | 72.3 | 60.2 | 94.6 |

Table 1: Test accuracy [%] on sentiment datasets. fastText has been run with the same parameters for all the datasets. It has 10 hidden units and we evaluate it with and without bigrams. For char-CNN, we show the best reported numbers without data augmentation.

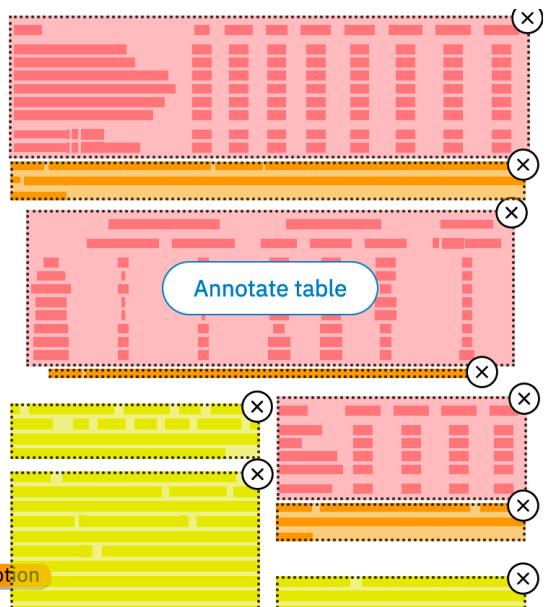
| | Zhang and LeCun (2015) | | Conneau et al. (2016) | | | fastText | |
|---------|------------------------|--------------|-----------------------|----------|----------|-------------------|--|
| | small char-CNN | big char-CNN | depth=9 | depth=17 | depth=29 | $h = 10$, bigram | |
| AG | 1h | 3h | 24m | 37m | 51m | 1s | |
| Sogou | - | - | 25m | 41m | 56m | 7s | |
| DBpedia | 2h | 5h | 27m | 44m | 1h | 2s | |
| Yelp P. | - | - | 28m | 43m | 1h09 | 3s | |
| Yelp F. | - | - | 29m | 45m | 1h12 | 4s | |
| Yah. A. | 8h | 1d | 1h | 1h33 | 2h | 5s | |
| Amz. F. | 2d | 5d | 2h45 | 4h20 | 7h | 9s | |
| Amz. P. | 2d | 5d | 2h45 | 4h25 | 7h | 10s | |

Table 2: Training time for a single epoch on sentiment analysis datasets compared to char-CNN and VDCNN.

to Tang et al. (2015) following their evaluation protocol. We report their main baselines as well as their two approaches based on recurrent networks (Conv-GRNN and LSTM-GRNN).

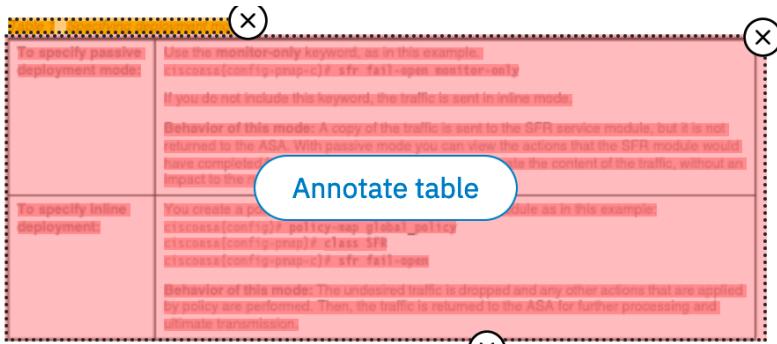
Results. We present the results in Figure 1. We use 10 hidden units and run fastText for 5 epochs with a learning rate selected on a validation set from {0.05, 0.1, 0.25, 0.5}. On this task, adding bigram information improves the performance by 1-4%. Overall our accuracy is slightly better than char-CNN and char-CRNN and, a bit worse than VDCNN. Note that we can increase the accuracy slightly by using more n-grams, for

Training time. Both char-CNN and VDCNN are trained on a NVIDIA Tesla K40 GPU, while our



Two of the examples above show that column headers sometimes stretch over several columns; this can also happen with row headers. Labeling such an arrangement as a Table is correct.

In the following table, the gridlines determine that this is a Table (with 2 rows and 2 columns), again a single cluster:



Some page parts may or may not have a table-like appearance. The key decision for or against using “Table” clusters should be: Do the text elements reasonably align into a grid structure or not? This means separable columns and rows.

Note that names or sums at the top or bottom of individual columns are part of the table, even if they look a little different:

| Grant | Year of grant | End of vesting period | Beginning of the year Number | Lapsed Number | Exercised Number | End of the year Number | Exercisable at the end of the year Number |
|--|---------------|-----------------------|------------------------------|----------------------|------------------------|------------------------|---|
| 2015 Performance Vested in prior years | 2008 | 31 Dec 2012 | 131,976 (131,976) | (10,299) (10,000) | = 4,142,738 | = 4,142,738 | = 4,142,738 |
| Total | | | 4,162,997 (42,235) | (10,000) | 4,142,738 | 4,142,738 | |
| Weighted average exercise price (\$) | | | 12.49 | 17.20 | 10.48 | 12.31 | 12.31 |
| 2012 Performance Vested in prior years | 2008 | 31 Dec 2018 | 176,025 (176,025) | = 131,976 | = 131,976 | = 131,976 | = 131,976 |
| F5 | 2008 | 31 Dec 2012 | 131,976 (131,976) | = 131,976 | = 131,976 | = 131,976 | = 131,976 |
| F4 | 2008 | 31 Dec 2018 | 131,976 (131,976) | = 131,976 | = 131,976 | = 131,976 | = 131,976 |
| G2 | 2009 | 1 Mar 2012 | 54,621 4306,576 | = 4306,576 | 54,621 4306,576 | 54,621 4306,576 | 54,621 4306,576 |
| Total | | | 490,574 (308,037) | = 4294,913 | 4,162,997 4,142,738 | | |
| Weighted average exercise price (\$) | | | 12.89 | 15.90 | N/A | 12.48 | 12.31 |

3.8 Title

The label “Title” is used for the overall title of a document only, as seen on the front page. Because front pages are quite difficult to annotate at times, they are treated separately for the different document categories in Part II of this document.

3.8.1 Title Repeated on Later Pages: “Text”

If the title is repeated on a later page, then “Title” is not used again; typically just “Text”.

However, if individual pages are given for annotation without further information, and a page looks like a title page, then annotate it as such.⁴ For this, you can check that no paragraph continues from another page, it does not start with header or list numbers greater than 1 or “a”), and other plausibility checks.

3.9 List-item

3.9.1 Key Rule: Hanging Shape

After some initial confusion about what to consider as List-item and what as Text, we aligned on **only 1 key rule**: List items should be “hanging”, i.e., this shape:

⁴ In the DocLayNet annotation campaign, title pages are presented separately.



I.e., the first line of each List-item hangs out to the left compared with the following lines, if any.

Each List-item is one cluster, including a potential list identifier (bullet, number etc.).

If there is a list-like structure, and some items have only 1 line, but the longer items in the same list are hanging, then make them all “List-item”, e.g., as here:

1. die vor dem 1. Januar 1991 ihren gewöhnlichen Aufenthalt im Gebiet der Bundesrepublik Deutschland genommen haben,
2. deren Rente nach dem 30. September 1996 beginnt und
3. über deren Rentenantrag oder über deren bis 31. Dezember 2004 gestellten Antrag auf Rücknahme des Rentenbescheides am 30. Juni 2006 noch nicht rechtskräftig entschieden worden ist.



By “1., 2., 3.” it is clear that these items belong together, and two of them are hanging, so we also declare item “2.” a List-item.

There are only 2 exceptions:

- **Lists with only 1-line items** (so that one cannot see if hanging or not) and clearly separated bullets, numbers, or such a scheme, must be still labelled as List-items.
E.g., this is also a list with three List-items:

- a) Apples
- b) Pears
- c) Oranges

- **Multi-line Section-header.** A numbered section header can also have a hanging shape, like below, but is **not** a list:

5. Discussion of Fruit Diseases and Countermeasures

We now come to ...

Any other shapes should not be lists, even if there are numbers or bullets.

Each list item (paragraph) should be a separate cluster, e.g., like this:

- Level of risk to operations: How exposed the business is to COVID-19 risks (e.g., facilities located in COVID hotspots, nature of work requires close personal contact)
- Expected time frame of recovery: How fast the business can resume normal operations once the country lockdowns begin being lifted
- Interdependencies between businesses and functions: How critical a particular business/function's operations are to other businesses/functions
- Change in business's strategic priorities: How the business' strategic priorities have evolved as a result of COVID-19



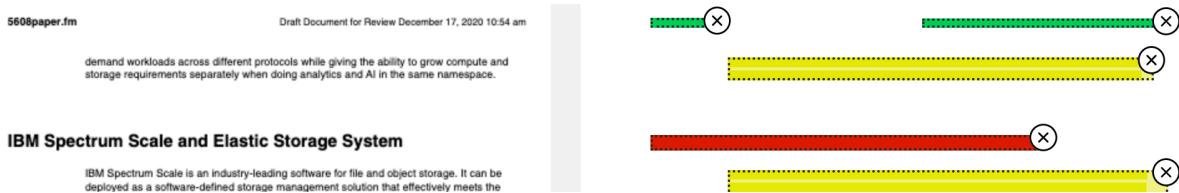
We provide more examples further below, but they all fit this rule.

3.10 Page-header

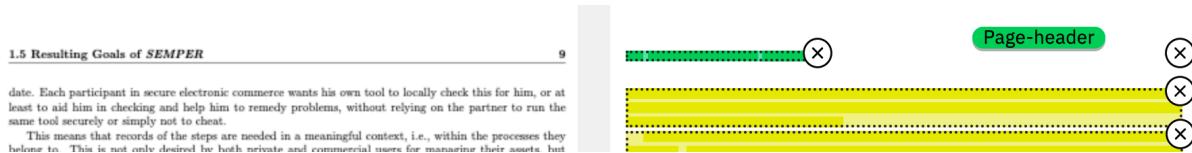
A page header is text at the top of a page that does not belong to the main text flow. Typically it repeats with small variation on all the pages of a document. Examples are page number, date, repetitions of document title or section header, or a confidentiality note.

3.10.1 Normal Page-headers

Often a Page-header is in a smaller font like here:



Or it is separated from the text by a horizontal line as in the next example:



In both examples, you saw that clearly separated parts of a page header should become separate clusters.

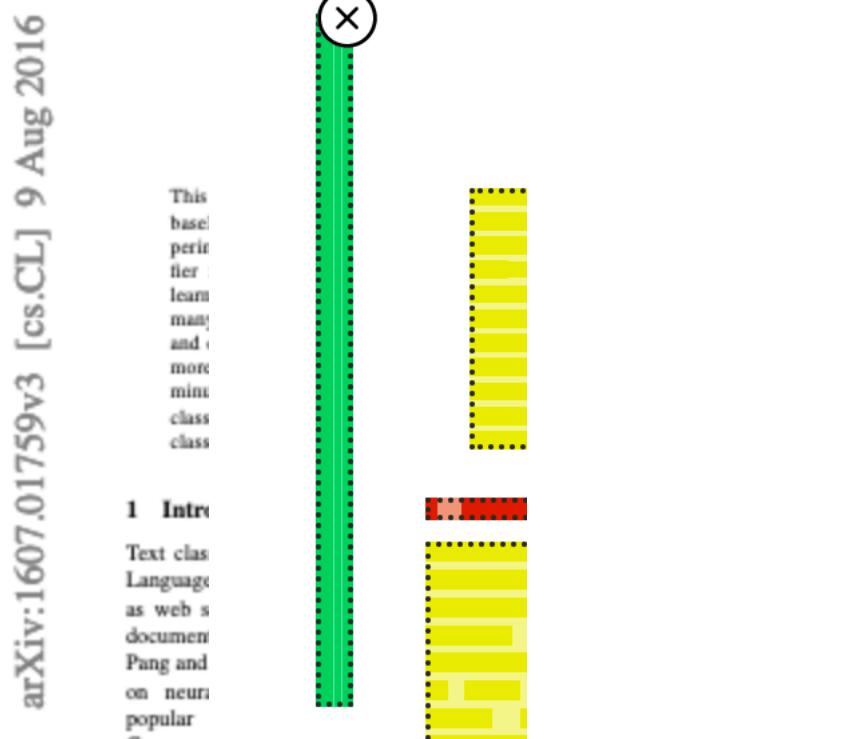
3.10.2 Only Page Number: Page-header

Please also annotate a Page header if it is only a page number as in the next screenshot. The fact that the "2" is labeled can only be seen by the cross, because of the small size.



3.10.3 Sideways Page-header

If Page-header-like text is sideways, then count it as Page-header, e.g., here:

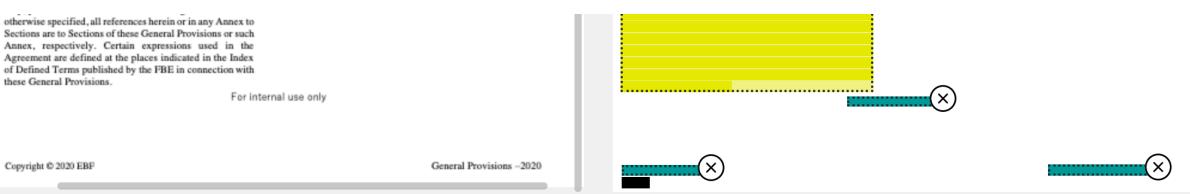


3.11 Page-footer

A Page-footer is like a Page-header, only at the bottom of the page. Some documents use only one of page header and page footer, some both. E.g., the following footer with page number and document title is from the same page as one of the headers above:



Again, parts with wide gaps become separate clusters. The following example is rather exotic because there are page-footers at different depths. While “For internal use only” is geometrically closer to the main text, it is not part of that text flow and repeats on every page, so it is also a Page-footer.



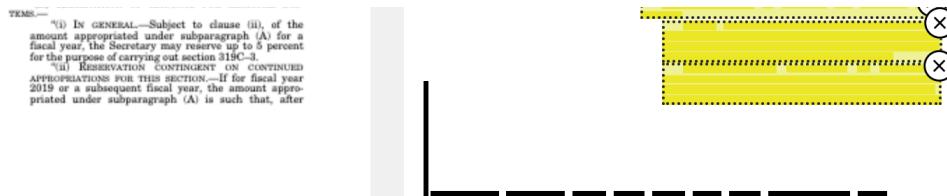
3.12 None

“None” will be interpreted in the machine-learning algorithms just like space where there really is no content. Initially, all elements are marked as “None”, so you typically only change them from “None” to other labels. However, there is an explicit label that you can use if you accidentally labeled something with another label and want to turn it back to “None”.

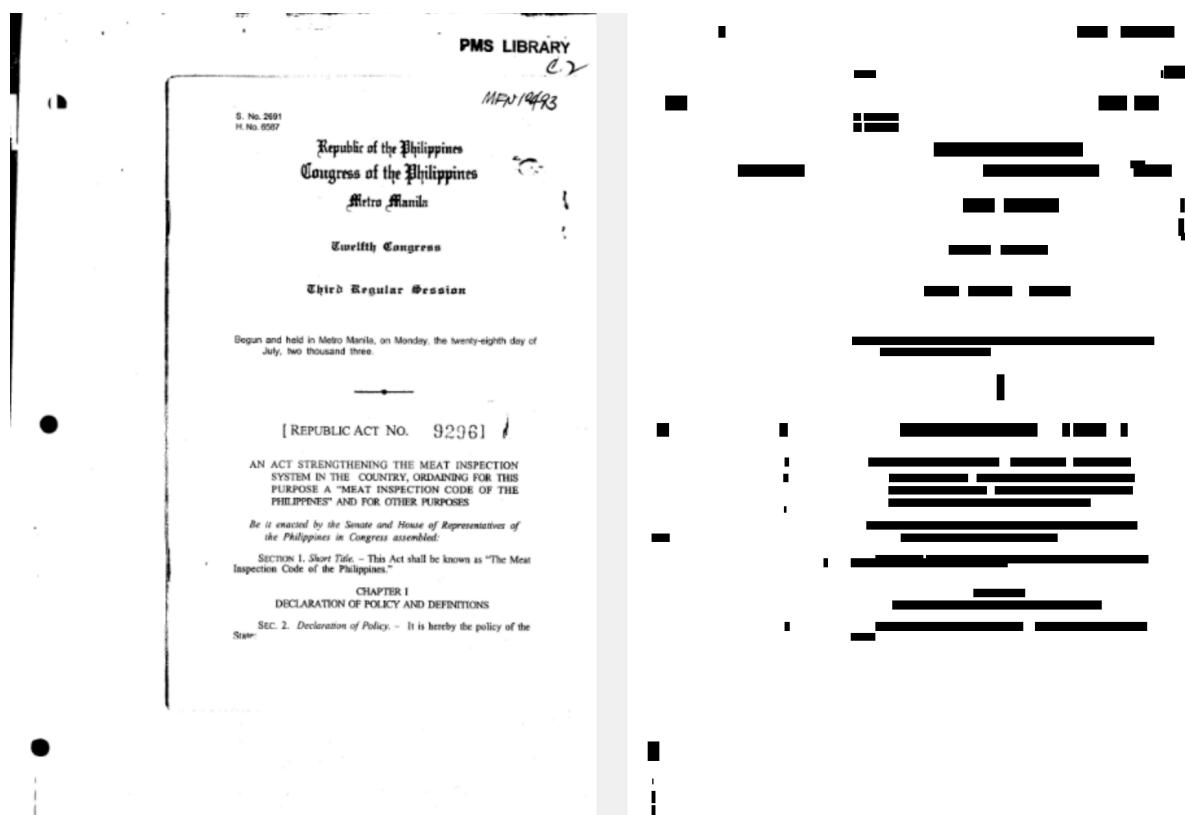
Recall that “None” should be used sparingly, only when there really is nothing visible or relevant. E.g., sometimes there are black boxes where you don’t see any text. Those should remain “None”. Also smears on a scanned page might be detected by the PDF parser, but form neither Text nor Tables nor Pictures.

In contrast, logos, page headers, page footers, well-visible pictures etc. should all be properly labeled.

An example exists on most US law pages: There is nothing to see on the left, but spurious boxes on the right around the page margin. Those remain “None”.



An extreme example is this scanned Philippine law title page. If we could annotate this at all⁵, we would keep the punch holes and the boxes that come from the page border as “None”.

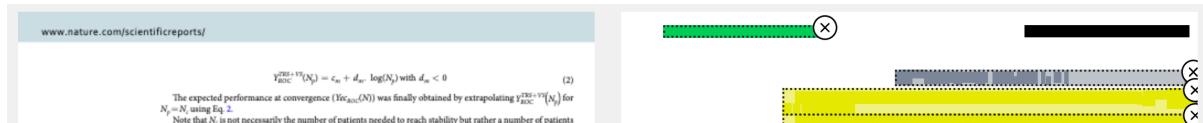


⁵ This example is actually not acceptable to annotate, since the word “Congress” is missing, and a few other boxes are misaligned.

4 Examples of Pages with Errors

4.1 Too many Text Boxes: Submit with “None”

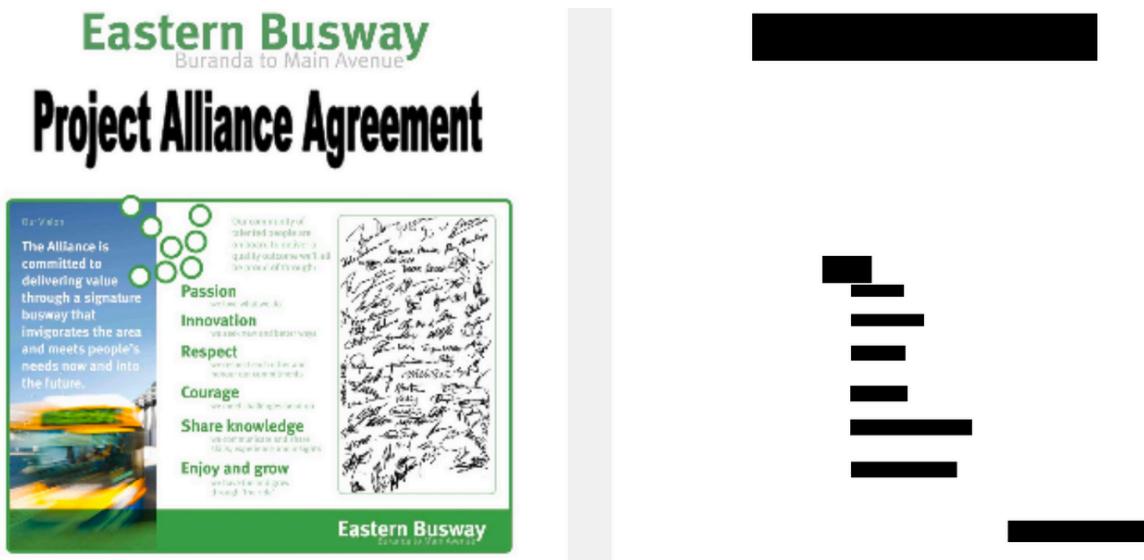
In the following example, the black box on the upper right is an “Incorrect text cell” because it corresponds to no visible text on the real page on the left. The label “None” is therefore correct. In such a case you can both submit the labeled page (because you can label it correctly) and report the error.



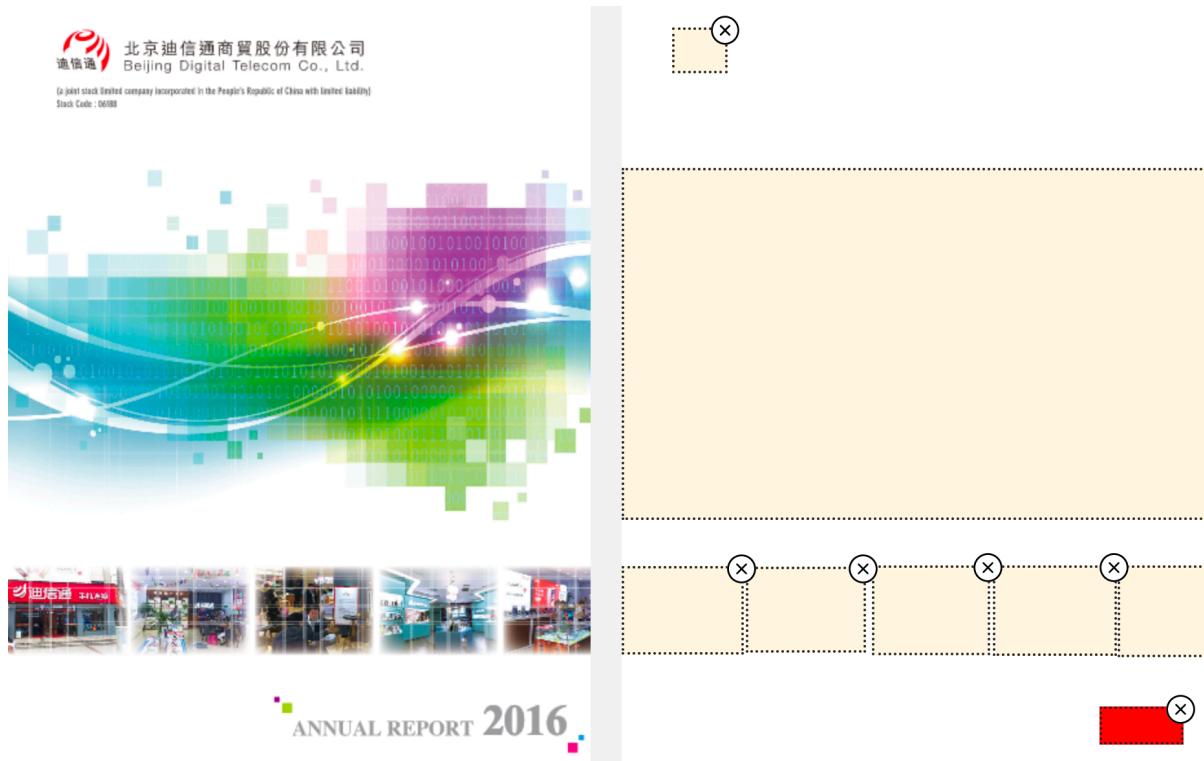
More examples were shown in Section 3.12 on “None”.

4.2 Missing Text Boxes: Do Not Submit

In the following example, you *only* report an error, but *do not* label the page, because “Project Alliance Agreement” is text and looks like the main title of a document, but there is no black box that you could annotate as “Title”:



Another example of a title page follows that you should NOT submit, because the company name “Beijing..., Ltd.” and “Annual Report” are important. While they are missing as text cells, you cannot annotate correctly.



In the next example, the real page contains a list with bullets, but the bullets are missing on the right. As the “List-item” cluster will only snap to the existing boxes, you cannot put “List-item” clusters where they should be on the left side. Hence you only report an error and do not label this page:

For the marketplace to be a strong transactional platform (e.g., e-commerce), the participants in a marketplace are acting as

- * **users**, (of the marketplace) i.e., buyers and sellers, also called customers and providers; and
- * **third parties**, who enable the session between the users, e.g., network and payment system providers.

Note that these are just convenient names of **roles** (i.e., we use “player” as a synonym of “role”, the words “participant”, “party”, and “person” are left for less restricted



Making such decisions, it is important to understand that the primary AI model we are planning to train with these annotations is vision-based. It will see exactly what is on the left side, and know nothing about the black boxes on the right. Hence even if you do your best with the provided boxes in the last two examples, you cannot provide what the AI model should learn.

4.3 Broken-up Text Boxes: Do Not Submit

The following is the way an old, scanned law was parsed:⁶

⁶ Page 4d095b1f30aad2bbf8fef8b791884e41789fa8b0521248a6d3d8ec10bfbc1a2c

2

- a) to promulgate specific policies and procedures governing the flow of food animals, meat and meat products through the various stages of marketing and the proper preservation and inspection of such products;
- b) to ensure food security and provide safety and quality standards for consumer products related to agriculture to assure the protection of the public against unreasonable risks of injury and hazards to health;
- c) to support local government units in their endeavor to be self-reliant and to continue exercising the powers and discharging the duties and functions currently vested upon them;
- d) to strengthen and harmonize various issuances on meat establishment operations and meat inspection and at the same time create a favorable climate of investment to encourage the meat and poultry industry to put up world-class meat establishments;
- e) to promote the application of risk analysis based on accepted scientific methodology on applied food safety standards so

Page-header

Interestingly, although the text boxes don't correspond to the letters or words of the text, it was possible to annotate this almost correctly with Text clusters. But there is a risk that some of the strange boxes reach out into the next paragraph or otherwise make the clusters imprecise. So it is not worth trying such annotations.

PART II. Special Cases By Document Category

There are many borderline cases between two possible labels, and we need consistent labeling for them. Fortunately, many of these cases only occur in one document category, such as “Patents” or “Laws and Regulations”, or even in a single collection, such as “Korean patents” or “US laws”. Thus we sort these cases by document category, and within a category, sometimes by collections. We start with those document categories that have the most standardized formats, and thus the fewest decision rules to remember.

Within each document category or collection, we sort subsections as follows:

- We first handle labels that may cause difficulties essentially in the same order as in Section 3, i.e., Text, Picture, Caption, etc. However, when you are hesitating between two labels, such as List vs. Table, you may need to look in both places.
- Next we consider Title pages, which are often particularly difficult.
- If the collection has other special pages, such as Tables of Content, Indices or Final pages, we consider those at the end.

5 Laws and Regulations

Laws of a particular country typically have a predefined format, i.e., they all look more or less the same. Thus we split this section by document collection. A collection usually corresponds to a country, but may also be a subgroup of regulations, here the FAA regulations separately from primary US laws.

5.1 US Laws

5.1.1 Text

US-laws have mostly numbered paragraphs, but they are indented instead of hanging, i.e., the first line starts further right than the following lines. Thus they must be labeled as Text. Most of the following examples come from one page.⁷

SEC. 302. Division A of the Consolidated Appropriations Act, 2019 (Public Law 116–6) is amended by adding after section 540 *Ante*, the following:

“SEC. 541. (a) Section 831 of the Homeland Security Act of 2002 (6 U.S.C. 391) shall be applied—
“(1) In subsection (a), by substituting ‘September 30, 2019,’ for ‘September 30, 2017,’; and
“(2) In subsection (c)(1), by substituting ‘September 30, 2019,’ for ‘September 30, 2017.’
“(b) The Secretary of Homeland Security, under the authority of section 831 of the Homeland Security Act of 2002 (6 U.S.C. 391(a)), may carry out prototype projects under section 2371b of title 10, United States Code, and the Secretary shall perform the functions of the Secretary of Defense as prescribed.
“(c) The Secretary of Homeland Security under section 831 of the Homeland Security Act of 2002 (6 U.S.C. 391(d)) may use



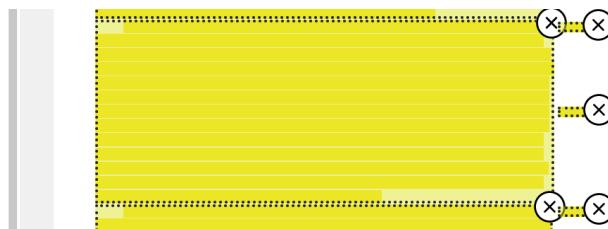
Splitting the text into paragraph clusters can be a little tricky when one comes back from a deeper-nested level to a less deep level. In the example above, see the return from the paragraph starting “(2) ...” to the paragraph starting “(b) ...”: The lines starting “2019” and

⁷ 6ae6f4978a1650f6d2c5f034f0da93ed399b471b913103311a64acb60a3a8e89

"(b)" start at the same indent. One can see from the lower indent in the line "of section ..." and the shortness of line "2019 ..." that a new paragraph started with "(b)".

There are also many short side-bars, which count as Text, e.g., here with 1 line each:

in section 231(d)(6) of title 10, UNITED STATES CODE.
 SEC. 303. None of the funds provided in this Act under "U.S. Customs and Border Protection—Operations and Support" for facilities shall be available until U.S. Customs and Border Protection establishes policies (via directive, procedures, guidance, and/or memorandum) and training programs to ensure that such facilities adhere to the National Standards on Transport, Escort, Detention, and Search, published in October of 2015: *Provided*. That not later than 90 days after the date of enactment of this Act, U.S. Customs and Border Protection shall provide a detailed report to the Committees on Appropriations of the Senate and the House of Representatives, the Committee on the Judiciary of the Senate, and the House Judiciary Committee regarding the establishment and implementation of such policies and training programs.
 SEC. 304. No later than 30 days after the date of enactment of this Act, the Secretary of Homeland Security shall provide a



Or here with a 2-line cluster:

tions serving communities that have experienced a significant influx of such aliens: *Provided further*, That such funds may be used to reimburse such jurisdictions or local recipient organizations for costs incurred in providing services to such aliens on or after January 1, 2019: *Provided further*, That such amount is designated by the Congress as being for an emergency requirement pursuant



5.1.2 Page-header

US laws have relatively large page headers. If one sees only 1 page, one may think they are section headers, but they repeat on every page and contain a page number (here "1021").

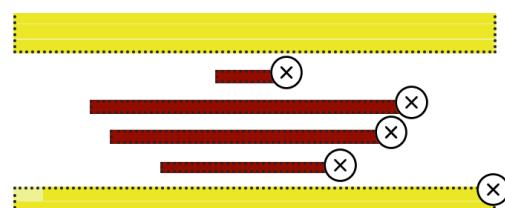
PUBLIC LAW 116-26—JULY 1, 2019 133 STAT. 1021
 Emergency Food and Shelter Program National Board shall distribute such funds only to jurisdictions or local recipient organizations serving communities that have experienced a significant influx of such aliens: *Provided further*, That such funds may be used



5.1.3 Section-header

Section-headers are easy to see, only it is sometimes not clear how many clusters there should be:⁸

That such amount is designated by the Congress as being for an emergency requirement pursuant to section 251(b)(2)(A)(ii) of the Balanced Budget and Emergency Deficit Control Act of 1985.
 TITLE III
 DEPARTMENT OF HOMELAND SECURITY
 U.S. CUSTOMS AND BORDER PROTECTION
 OPERATIONS AND SUPPORT
For an additional amount for "Operations and Support" for necessary expenses to respond to the significant rise in aliens



These are probably 3 levels of sections starting, but as the distance between the first and second line is as large as between the second and third line, we make 1 cluster per line.

5.1.4 None

Most pages in US laws have spurious cells at the bottom, which should remain "None":

*a return to northern border staffing levels that are no less than the number committed in the June 12, 2018 Department of Homeland Security Northern Border Strategy: *Provided further*, That the report shall include the number of officers temporarily assigned*



⁸ Page 744172015c62e664ec889072cd71bf225b1579eacfb0fda17bd72a71e28c9e12

5.1.5 US Law Title Pages

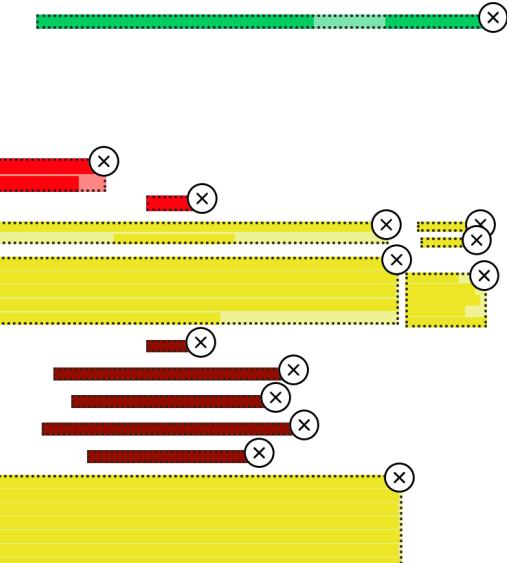
US law title pages are special in the way that what looks like a title is “An Act” and always the same. Hence we make both this and the law number above it part of the “Title” cluster.⁹

The most interesting short description of the document is the very small text below “An Act”, and as humans we might call that a title, but given that it is so small, we cannot train our AI to recognize it as a title, hence we make it “Text”.

PUBLIC LAW 116–20—JUNE 6, 2019

133 STAT. 871

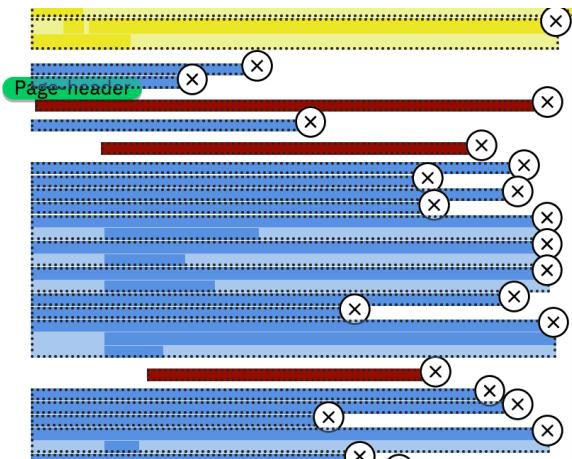
Public Law 116–20
116th Congress
An Act
Making supplemental appropriations for the fiscal year ending September 30, 2019, and for other purposes. June 6, 2019
[H.R. 2157]
Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, That the following sums in this Act are appropriated, out of any money in the Treasury not otherwise appropriated, for the fiscal year ending September 30, 2019, and for other purposes, namely:
Additional Supplemental Appropriations for Disaster Relief Act, 2019.
TITLE I
DEPARTMENT OF AGRICULTURE
AGRICULTURAL PROGRAMS
PROCESSING, RESEARCH AND MARKETING
OFFICE OF THE SECRETARY
For an additional amount for the “Office of the Secretary”, \$3,005,442,000, which shall remain available until December 31, 2020, for necessary expenses related to losses of crops (including milk, on-farm stored commodities, crops prevented from planting in 2019, and harvested adulterated wine grapes), trees, bushes, and vines, as a consequence of Hurricanes Michael and Florence, other hurricanes, floods, tornadoes, tsunamis, volcanic activity,



5.1.6 US Law Table of Contents

Some of the longer laws have a table of contents. It starts directly on the title page, and looks as in the following example.¹⁰ As the 2-line paragraphs are hanging, we also make the 1-line paragraphs List-items.

(b) TABLE OF CONTENTS.—The table of contents for this Act is as follows:
Sec. 1. Short title; table of contents.
Sec. 2. References in Act.
TITLE I—STRENGTHENING THE NATIONAL HEALTH SECURITY STRATEGY
Sec. 101. National Health Security Strategy.
TITLE II—IMPROVING PREPAREDNESS AND RESPONSE
Sec. 201. Improving benchmarks and standards for preparedness and response.
Sec. 202. Amendments to preparedness and response programs.
Sec. 203. Regional health care emergency preparedness and response systems.
Sec. 204. Military and civilian partnership for trauma readiness.
Sec. 205. Public health and health care system situational awareness and surveillance capabilities.
Sec. 206. Strengthening and supporting the public health emergency rapid response fund.
Sec. 207. Improving all-hazards preparedness and response by public health emergency volunteers.
Sec. 208. Clarifying State liability law for volunteer health care professionals.
Sec. 209. Report on adequate national blood supply.
Sec. 210. Report on the public health preparedness and response capabilities and capacities of hospitals, long-term care facilities, and other health care facilities.
TITLE III—REACHING ALL COMMUNITIES
Sec. 301. Strengthening and assessing the emergency response workforce.
Sec. 302. Health system infrastructure to improve preparedness and response.
Sec. 303. Considerations for at-risk individuals.
Sec. 304. Improving emergency preparedness and response considerations for children.
Sec. 305. National advisory committees on disasters.



⁹ Page 2e30cae45be5c9d68de36dd8a2408e141416d8ff0c9fef4c7f2adcc9c3f4e673

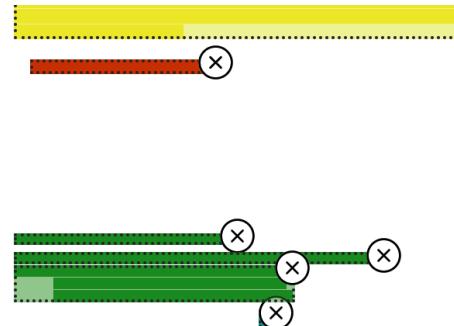
¹⁰ Page 7cc098303bad59616e8c42807b0240c21d29fcf4844451ccd9434e1d56c04f5

5.1.7 US Law Final Pages

US laws do not have a complete final page, but a specific footnote-like structure at the end. We indeed label that as several “Footnote” clusters. The circle under them is in a page-number position and becomes “Page-footer”.¹¹

mittee, provided that such statement has been submitted prior to the vote on passage.

Approved June 25, 2019.



5.2 British Laws (Collection “gb_laws”)

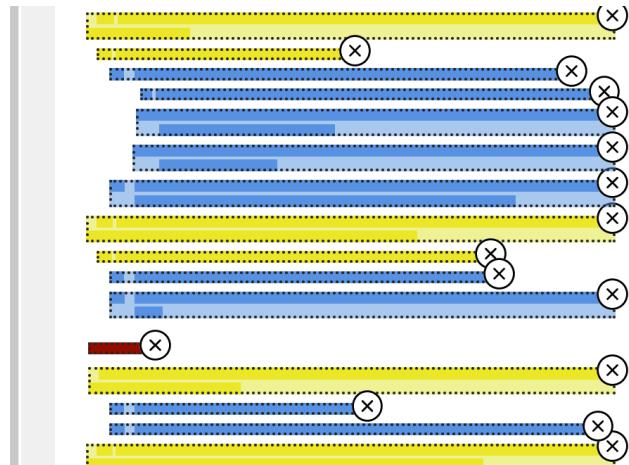
5.2.1 Text and List-items

British laws have numbered paragraphs throughout. Some levels are hanging, others not, as you see in the following example:¹²

(3) Paragraph (1)(a) does not apply to a person during an interim period specified under regulation 14(c)(i).
(4) For the purpose of paragraph (1)(c)(ii)(dd)—
(a) a person is the close relative of someone in service of the Crown if the person is—
(i) the spouse, civil partner or cohabiting partner of the person in service of the Crown;
(ii) a child under the age of 18 of the person in service of the Crown or of a person to whom sub-paragraph (i) applies;
(iii) a child of the person in service of the Crown who is dependent on that person as a result of disablement;
(b) a person was an accompanying close relative of a person in service of the Crown if they were a close relative and living with that person in the same household.
(5) Paragraph (1)(d) does not apply in any case where a panel considers that the application of that paragraph would undermine the purposes of this Scheme.
(6) For the purpose of paragraph (5), the purposes of this Scheme are to—
(a) acknowledge the harm suffered by those injured in the Troubles, and
(b) promote reconciliation between people in connection with Northern Ireland's troubled past.

Convictions

6.—(1) A person is not entitled to victims' payments in relation to a particular Troubles-related incident where the person—
(a) has a conviction (whether spent or not), and
(b) that conviction was in respect of conduct which caused, wholly or in part, that incident.
(2) A person is not entitled to victims' payments where the Board considers that the person's relevant conviction makes entitlement to victims' payments inappropriate.



The outermost two levels (one paragraph starting “6.—” and those starting “(3)” etc. are indented and thus “Text”, the inner two levels hanging and thus “List-item”.

5.2.2 Section-header

Certain Section-headers are broken into two widely separated lines with different formats, so we make them 2 clusters, like “PART 4” and its text here:¹³

¹¹ Page 7704a0e93c941c3732ef2d62d78cc06c49acd05b5aa08548473a60194324b482

¹² Page a46ffd3e5349b8848a789a351bdc524bffcb80f07c6c362b371232f02fcacc62

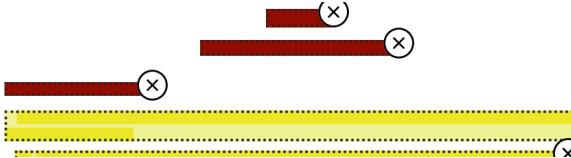
¹³ Page 58c0b15b3f381588ce25d3537aec5ffd1e95b0439e0930386e8c5a7e45b6aa1a

PART 4
Determination of entitlement

Priority of applications

11.—(1) The Board may decide the order of priority in which applications for victims' payment are to be determined.

(2) In making a decision under paragraph (1), the Board must, in particular, have regard to—



5.2.3 British Law Title Pages

British law front pages have several elements where the decision is difficult. We decided as in the following page:¹⁴

- The title is all the lines in the same large font, including the law number. Because of the large gaps, we make it four Title clusters.
- The top part between two horizontal lines is considered a Page-header. (In particular as it repeats across all laws, which you can't see from just one page.)
- The dates with the large gaps are considered a table, but in this law there is a third line not fitting that, so that has to become Text.
- “Part I” and “Introductory” might belong together, but on different lines with different fonts, two separate clusters are better.

STATUTORY INSTRUMENTS

2020 No. 108

EXITING THE EUROPEAN UNION

AUDITORS

INSURANCE

The Statutory Auditors and Third Country Auditors (Amendment) (EU Exit) Regulations 2020

Made - - - - - 31st January 2020
Laid before Parliament 3rd February 2020
Coming into force in accordance with regulation 1(2)

The Secretary of State makes the following Regulations in exercise of the powers conferred by section 2(2) of the European Communities Act 1972(a), sections 484(1), 1240A, 1240B and 1292(1) of the Companies Act 2006(b) and sections 15 and 17(3) of the Limited Liability Partnerships Act 2000(c).

The Secretary of State is a Minister designated(d) for the purposes of section 2(2) of the European Communities Act 1972 in relation to auditors and the audit of accounts.

PART 1
Introductory

Citation, commencement and application

1.—(1) These Regulations may be cited as the Statutory Auditors and Third Country Auditors (Amendment) (EU Exit) Regulations 2020.

(2) These Regulations come into force—

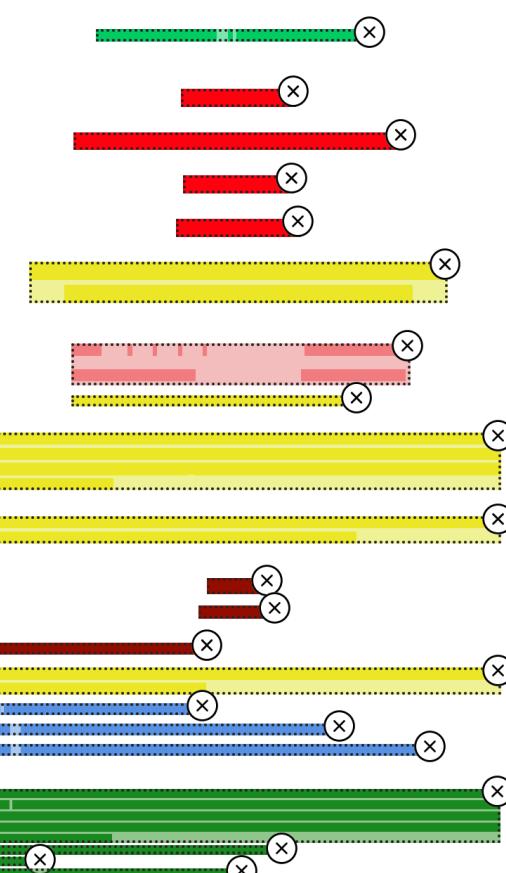
- in the case of Part 2, immediately before IP completion day;
- otherwise on the 21st day after the day on which these Regulations are made.

(a) 1972 c. 68. Section 2(2) was relevantly amended by section 27(1) of the Legislative and Regulatory Reform Act 2006 (c. 51) and by section 3 of, and Part 1 of the Schedule to, the European Union (Amendment) Act 2008 (c. 7). The European Communities Act 1972 is repealed on exit day by the European Union (Withdrawal) Act 2018 (c. 16), but continues to have effect until IP completion day pursuant to section 1A of that Act, inserted by section 1 of the European Union (Withdrawal) Act 2020 (c. 11).

(b) 2006 c. 46. Sections 1240A and 1240B were inserted by S.I. 2019/177.

(c) 2000 c. 12.

(d) S.I. 2007/1679. See S.I. 2018/1011 for relevant amendments.

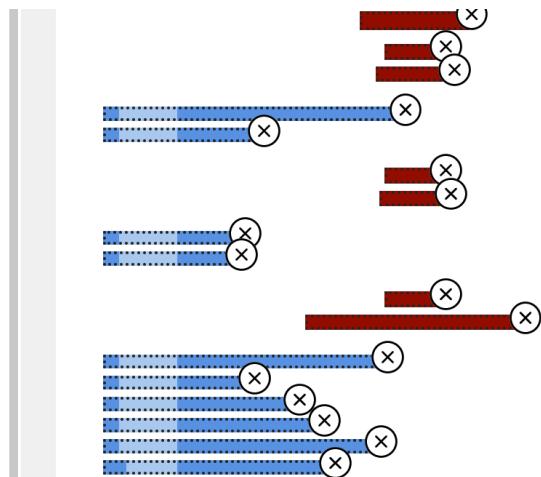


¹⁴ Page 363e0115b10d4aa2f7b4e4c71ae54bf99233ff91bbc6d28228ec590b739595df

5.2.4 British Law Table of Contents

Some of the longer laws have tables of contents, starting directly on the title page, e.g., here:¹⁵

| CONTENTS | |
|--|-----------------------------------|
| PART 1 Preliminary | |
| 1. | Citation, commencement and extent |
| 2. | Interpretation |
| PART 2 The Board | |
| 3. | The Board |
| 4. | Principles |
| PART 3 Entitlement to victims' payments | |
| 5. | Entitlement to victims' payments |
| 6. | Convictions |
| 7. | Causation of injury |
| 8. | Making of applications |
| 9. | Transfer of entitlement on death |
| 10. | Posthumous applications |

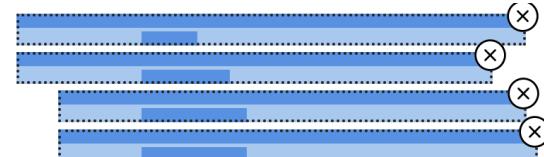


All paragraphs numbered 1., 2., ... are 1-liners, also on the following page. However, with the large distance between the number and the text, we assume that they would be hanging if longer (by the first exception from Section 3.9.1), and make them “List-item”.

We handled the 2-line Section-headers as above.

There is a second kind of table of content with clear List-items like this:¹⁶

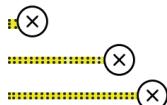
| |
|--|
| SCHEDULE 4 — LAND OF WHICH TEMPORARY POSSESSION MAY BE TAKEN |
| SCHEDULE 5 — REPLACEMENT AND CLOSURE OF ROAD LEVEL CROSSINGS |
| PART 1 — LEVEL CROSSINGS FOR WHICH A SUBSTITUTE IS TO BE PROVIDED |
| PART 2 — LEVEL CROSSINGS FOR WHICH NO SUBSTITUTE IS TO BE PROVIDED |



5.2.5 British Law Final Pages

British laws typically have an almost empty final page with this lower part:¹⁷

£4.99
UK20200311007 03/2020 39585
<http://www.legislation.gov.uk/ukds/id/20200301>



¹⁵ Page 77b1ee64a8471505ce0fb901e0fb146fdb1d7f76c5b8e0eda260a3bbd940bb12

¹⁶ Page 5f42e6272ae84dbad0cba7e220f679fe333321ceaca4bac6f4a6f0809f2298c2

¹⁷ Page 72a24f978cee30eb3fcf2d0349463eb4889f1aa97f8ce110f3c9f00d462fae58

5.3 German Laws

5.3.1 Section-header

In German laws, certain Section-headers are broken into 2 lines although the first line is not completely filled. You can recognize this in the next example even if you don't speak German: There is a Section 5 with its name on the next line (same font, both bold, close together), and within that, there is a paragraph 23. Thus we make 2 Section-headers:¹⁸

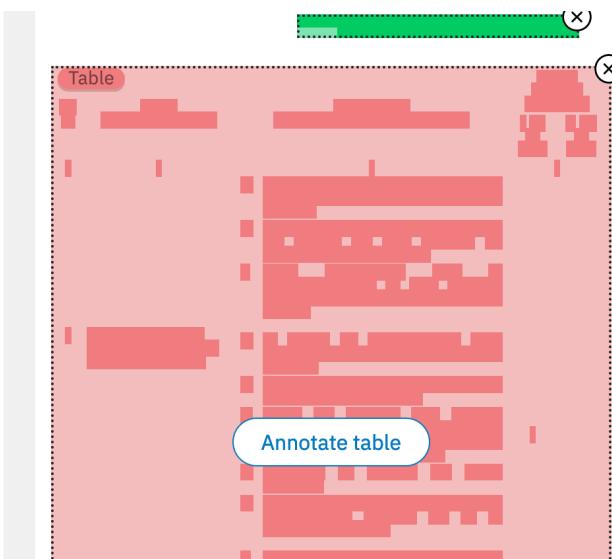
Zulassungsschein an die ausstellende Stelle zurückgesandt. Hat der Eigentümer oder Halter des Fahrzeugs keinen Wohn- oder Aufenthaltsort im Inland, ist für Maßnahmen nach Satz 1 jede Verwaltungsbehörde nach § 46 Absatz 1 zuständig.
Abschnitt 5
Überwachung des Versicherungsschutzes der Fahrzeuge
§ 23 Versicherungsnachweis
(1) Der Nachweis nach § 3 Absatz 1 Satz 2, § 16 Absatz 1 Satz 1 oder § 16a Absatz 1 Satz 1 Nummer 3, dass ein dem Pflichtversicherungsgesetz entsprechende Kraftfahrzeug-Haftpflichtversicherung besteht, ist bei



5.3.2 Table

There are some large tables in German laws, recognizable by clear gridlines.¹⁹

| Lfd. Nr. | Teil des Ausbildungsberufsbildes | Zu vermittelnde Fertigkeiten, Kenntnisse und Fähigkeiten | Zeitliche Richtwerte in Wochen im | |
|----------|--|--|-----------------------------------|-------------------|
| | | | 1. bis 18. Monat | 19. bis 36. Monat |
| 1 | 2 | 3 | 4 | |
| | | g) Aufgaben im Team sowie mit internen und externen Kunden und Kundinnen planen und abstimmen h) betriebswirtschaftlich relevante Daten erheben und bewerten und dabei Geschäfts- und Leistungsprozesse berücksichtigen i) eigene Vorgehensweise sowie die Aufgabendurchführung im Team reflektieren und bei der Verbesserung der Arbeitsprozesse mitwirken | | |
| 2 | Informieren und Beraten von Kunden und Kundinnen (§ 4 Absatz 2 Nummer 2) | a) im Rahmen der Marktbeobachtung Preise, Leistungen und Bedingungen von Wettbewerbern vergleichen b) Bedarfe von Kunden und Kundinnen feststellen sowie Zielgruppen unterscheiden c) Kunden und Kundinnen unter Beachtung von Kommunikationsregeln informieren sowie Sachverhalte präsentieren und dabei deutsche und englische Fachbegriffe anwenden d) Maßnahmen für Marketing und Vertrieb unterstützen e) Informationsquellen auch in englischer Sprache aufgabenbezogen auswerten und für die Kundeninformation nutzen f) Gezielte situationsgerechte führen und Kunden | 3 | |



We are not aware of Captions; the tables we saw were under normal Section-headers:²⁰

| § 43 Inkrafttreten, Außerkrafttreten | | | | |
|---|---|--|-----------------------------------|-------------------|
| Diese Verordnung tritt am 1. August 2020 in Kraft. Gleichzeitig tritt die Verordnung über die Berufsausbildung im Bereich der Informations- und Telekommunikationstechnik vom 10. Juli 1997 (BGBl. I S. 1741), die durch Artikel 1 der Verordnung vom 28. Mai 2018 (BGBl. I S. 654) geändert worden ist, außer Kraft. | | | | |
| Anlage (zu § 3 Absatz 1) Ausbildungsrahmenplan für die Berufsausbildung zum Fachinformatiker und zur Fachinformatikerin | | | | |
| (Fundstelle: BGBl. I 2020, 260 - 267) | | | | |
| Abschnitt A: fachrichtungsübergreifende berufsprofilgebende Fertigkeiten, Kenntnisse und Fähigkeiten | | | | |
| Lfd. Nr. | Teil des Ausbildungsberufsbildes | Zu vermittelnde Fertigkeiten, Kenntnisse und Fähigkeiten | Zeitliche Richtwerte in Wochen im | |
| 1 | 2 | 3 | 1. bis 18. Monat | 19. bis 36. Monat |
| 1 | Planen, Vorbereiten und Durchführen von Arbeitsergebnissen in | a) Grundsätze und Methoden des Projektmanagements anwenden | | |



¹⁸ Page 1c01afeef57bc78652a1791da7ae51c4b3c33c9a50db255fcfb0657fb369468c

¹⁹ Page c8d39b3d6b79ae0cc176c747d68ca9e2c636b7d06cc9c736f870fc3b27931137

²⁰ Page 93ad52537c360b58167932153c42715fda4af7d01e64ea596870f2e349c4f764

5.3.3 List-item

There are many list-like paragraphs in German laws. They are treated by the rules from Section 3.9, but you will quickly recognize repeating major types.

5.3.3.1 Single-Line Paragraphs

Recall that for single-line paragraphs that might be Text or List-item, you can compare with neighboring paragraphs. E.g., you see this:²¹

(6) Bei Berechtigten nach dem Fremdrentengesetz, die

Or this:

2. deren Rente nach dem 30. September 1996 beginnt und

You may guess by the white space that the second may be hanging and the first not. But it becomes really clear when you look at the entire lists, and see the non-hanging items (4a) and (5) in the same list as (6), and the hanging items 1. ... and 3. ... in the same list as 2.

(4a) Ist eine bereits vorher geleistete Rente neu festzustellen und sind dabei die persönlichen Entgeltpunkte neu zu ermitteln, sind die Vorschriften des Fremdrentengesetzes maßgebend, die bei erstmaliger Feststellung der Rentenzuflüsse waren, soweit § 317 Abs. 2 des Sechsten Buches Sozialgesetzbuch nichts anderes bestimmt.

(5) § 22 Abs. 3 des Fremdrentengesetzes in der bis zum 31. Dezember 1991 geltenden Fassung und § 22 Abs. 4 des Fremdrentengesetzes in der ab 1. Januar 1992 sowie in der vom 7. Mai 1996 an geltenden Fassung finden keine Anwendung auf Berechtigte, die nach Maßgabe des Abkommens vom 8. Dezember 1990 zwischen der Bundesrepublik Deutschland und der Republik Polen über Soziale Sicherheit Ansprüche und Answartschaften auf der Grundlage des Abkommens vom 9. Oktober 1975 zwischen der Bundesrepublik Deutschland und der Volksrepublik Polen über Renten und Unfallversicherung haben.

(6) Bei Berechtigten nach dem Fremdrentengesetz, die

a) ihren gewöhnlichen Aufenthalt im Beitragsgebiet haben und dort nach dem 31. Dezember 1991 einen Anspruch auf Zahlung einer Rente nach dem Fremdrentengesetz erwerben,

b) nach dem 31. Dezember 1990 ihren gewöhnlichen Aufenthalt aus dem Beitragsgebiet in das Gebiet der Bundesrepublik Deutschland ohne das Beitragsgebiet verlegen und dort nach dem 31. Dezember 1991 eine Anspruch auf Zahlung einer Rente nach dem Fremdrentengesetz erwerben oder

c) nach dem 31. Dezember 1991 ihren gewöhnlichen Aufenthalt aus dem Gebiet der Bundesrepublik Deutschland ohne das Beitragsgebiet in das Beitragsgebiet verlegen und bereits vor Verlegung des gewöhnlichen Aufenthalts einen Anspruch auf Zahlung einer Rente nach dem Fremdrentengesetz haben, werden für nach dem Fremdrentengesetz anerkannte Zeiten Entgeltpunkte (Ost) ermittelt; im Falle von Bürgern, die am 1. Januar 1992 einen Aufenthalt im Beitragsgebiet haben, wenn das Beitragsgebiet in dem Fremdrentengesetz nicht bestand. Dies gilt auch für die Zeiten eines weiteren Rentenbezuges aufgrund neuer Rentenfeststellungen, wenn sich die Rentenbezugszeiten ununterbrochen aneinander anschließen. Bei Berechtigten nach Satz 1 Buchstabe a und c, die ihren gewöhnlichen Aufenthalt aus dem Beitragsgebiet in das Gebiet der Bundesrepublik Deutschland ohne das Beitragsgebiet verlegen, verbleibt es für Zeiten nach dem Fremdrentengesetz bei den ermittelten Entgeltpunkten (Ost).

(7) (weggefallen)

§ 4a

§ 22a des Fremdrentengesetzes gilt nicht für Personen nach § 4 Abs. 5.

§ 4b

§ 22b des Fremdrentengesetzes ist nicht für Berechtigte anzuwenden, die vor dem 7. Mai 1996 ihren gewöhnlichen Aufenthalt in der Bundesrepublik Deutschland genommen haben.

§ 4c

(1) Für Berechtigte, die vor dem 7. Mai 1996 ihren gewöhnlichen Aufenthalt im Gebiet der Bundesrepublik Deutschland genommen haben und deren Rente vor dem 1. Oktober 1996 beginnt, sind für die Berechnung dieser Rente das § 22 Abs. 3 des Fremdrentengesetzes in der bis zum 31. Dezember 1991 geltenden Fassung und § 22 Abs. 4 des Fremdrentengesetzes in der ab dem 1. Januar 1992 geltenden Fassung sowie § 4 Abs. 5 und 7 in der am 6. Mai 1996 geltenden Fassung anzuwenden.

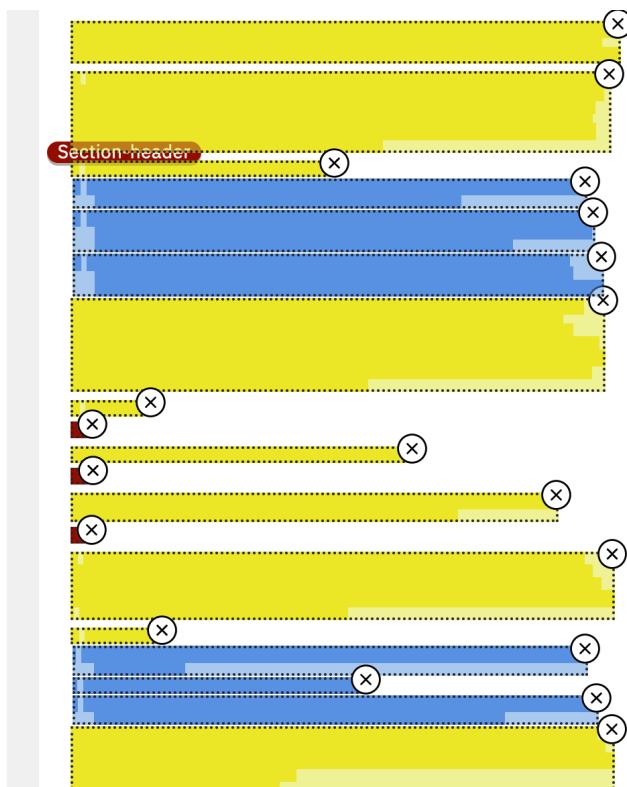
(2) Für Berechtigte,

1. die vor dem 1. Januar 1991 ihren gewöhnlichen Aufenthalt im Gebiet der Bundesrepublik Deutschland genommen haben;

2. deren Rente nach dem 30. September 1996 beginnt und

3. Über deren Rentenantrag oder über deren bis 31. Dezember 2004 gestellten Antrag auf Rücknahme des Rentenbescheides am 30. Juni 2006 noch nicht rechtskräftig entschieden worden ist,

wird für diese Rente einmalig zum Rentenbeginn ein Zuschlag an persönlichen Entgeltpunkten ermittelt. Der Zuschlag an persönlichen Entgeltpunkten ergibt sich aus der Differenz zwischen der mit und ohne Anwendung von § 22 Abs. 4 des Fremdrentengesetzes ermittelten Summe aller persönlichen Entgeltpunkte. Dieser Zuschlag wird monatlich für die Zeiten des Rentenbezuges vom 1. Oktober 1996 bis 30. Juni 1997 voll.



²¹ Page cfa3052ff79548db033e28ff2b9596d075673d4fc4c23dd027ec9ac42c7f28a2

5.3.4 List-like with Third Column: Table

German laws contain several mixtures of List and Table like this:²²

3. für den Prüfungsbereich Wirtschafts- und Sozialkunde kommen Aufgaben, die sich auf praxisbezogene Fälle beziehen sollen, insbesondere aus folgenden Gebieten in Betracht:
allgemeine wirtschaftliche und gesellschaftliche Zusammenhänge der Berufs- und Arbeitswelt.
- (4) Für den Teil B der Prüfung ist von folgenden zeitlichen Höchstwerten auszugehen:
- | | |
|---|--------------|
| 1. im Prüfungsbereich Beschichtungstechnik und Gestaltung | 180 Minuten, |
| 2. im Prüfungsbereich Instandsetzung und Instandhaltung | 120 Minuten, |
| 3. im Prüfungsbereich Wirtschafts- und Sozialkunde | 60 Minuten. |
- (5) Der Teil B der Prüfung ist auf Antrag des Prüflings oder nach Erlassen des Prüfungsausschusses in einzelnen Prüfungsteilen eine mündliche Prüfung zu ergänzen, wenn diese für das Bestehen der Prüfung den Ausschlag geben kann. Bei der Ermittlung der Ergebnisse für die mündlich geprüften Prüfungsbereiche sind die jeweiligen bisherigen Ergebnisse und die entsprechenden Ergebnisse der mündlichen Ergänzungsprüfung im Verhältnis 2:1 zu gewichten.
- (6) Innerhalb des schriftlichen Teils der Prüfung sind die Prüfungsbereiche wie folgt zu gewichten:
- | | |
|--|-------------|
| 1. Prüfungsbereich Beschichtungstechnik und Gestaltung | 55 Prozent, |
| 2. Prüfungsbereich Instandsetzung und Instandhaltung | 25 Prozent, |
| 3. Prüfungsbereich Wirtschafts- und Sozialkunde | 20 Prozent. |
- (7) Die Prüfung ist bestanden, wenn jeweils in den Prüfungsteilen A und B mindestens ausreichende Leistungen erbracht sind. Weiterhin sind in zwei der Prüfungsbereiche mindestens ausreichende Leistungen zu erbringen.

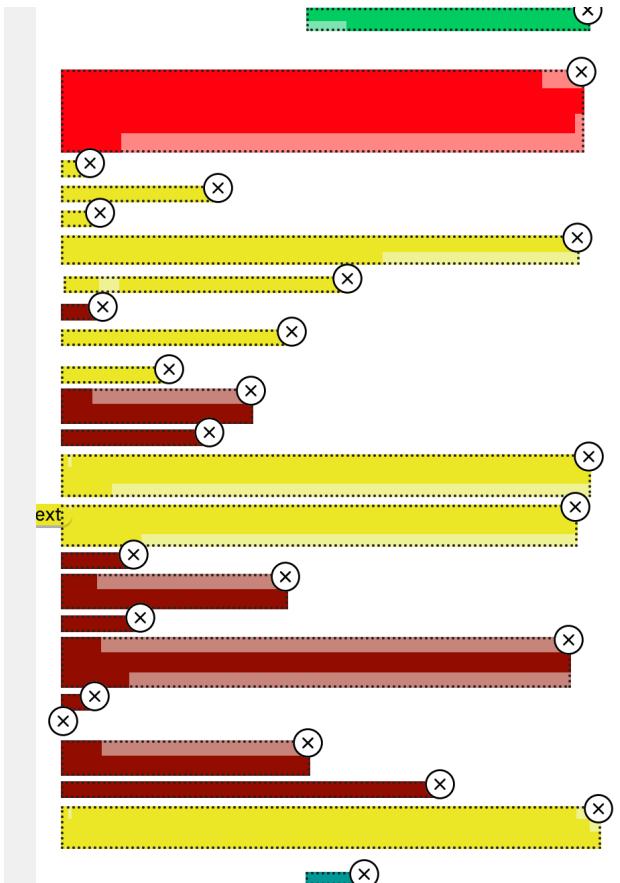


There are no gridlines, and the start of each line is as for list items, but there is a third column, here with minutes in the first list, and with percentages in the second list. We decided to use “Table” In these cases.

5.3.5 German Law Title Pages

In German laws, the Title is quite clear. Then some standard front-matter follows that is only “Text”, until “Fußnote” as a first small Section-header. From then on, various material can follow, here normal Section-headers and Texts.²³

- Ein Service des Bundesministeriums der Justiz und für Verbraucherschutz sowie des Bundesamts für Justiz – www.gesetze-im-internet.de
- Gesetz zur Anpassung verschiedener Vorschriften über die Finanzbeziehungen zwischen dem Bund und den Ländern an die Neuregelung der Finanzverfassung (Finanzanpassungsgesetz - FAnpG)**
- FAnpG
Ausfertigungsdatum: 30.08.1971
Vollzitat:
"Finanzanpassungsgesetz vom 30. August 1971 (BGBl. I S. 1426), das durch Artikel der Verordnung § 6 des Gesetzes vom 23. Mai 1975 (BGBl. I S. 1173) geändert worden ist"
Stand: Geändert durch Art. V § 6 G v. 23.5.1975 I 1173
Fußnote
(+++ Textnachweis Geltung ab: 1.7.1975 +++)
Art. 5: FVG 1971 600-1
**Teil I
Kosten des Aufgabenvollzugs**
Art 1 Verwaltungsausgaben
(1) Der Bund und die Länder tragen gesondert die Verwaltungsausgaben, die sich aus der Wahrnehmung der ihnen obliegenden Verwaltungsaufgaben ergeben. Die Erstattung von Verwaltungskosten bei Amtshilfe bleibt unberührt.
(2) Erfüllen die Länder oder der Bund auf Grund von Verwaltungsvereinbarungen Verwaltungsaufgaben, die dem anderen Teil obliegen, richtet sich die Erstattung von Verwaltungsausgaben nach den getroffenen Vereinbarungen.
Art 2 bis 4 ----
**Teil II
Neuordnung der Finanzverwaltung**
Art 5 bis 11 ----
**Teil III
Anpassung des Gesetzes zur Förderung der Stabilität und des Wachstums der Wirtschaft**
Art 12
-
**Teil IV
Übergangs- und Schlußbestimmungen**
Art 13 Überleitung bestimmter Beamter und Versorgungsberechtigter
(1) Bleibt das nach § 9 Abs. 2 Satz 3 des Gesetzes über die Finanzverwaltung in der Fassung des Artikels 5 einem Oberfinanzpräsidenten zustehende Grundgehalt hinter dem Grundgehalt zurück, das ihm am Tage vor dem Inkrafttreten dieser Vorschrift zustand, so erhält er eine Ausgleichszulage entsprechend Artikel IX § 11 des



²² Page e59e26d609c2ababb27898bc8a24d659d69120c9ebcd1d0bff21d8ed370e35bc

²³ Page f6701f4148263a549b115a787ee40a082e5034b5898687535a78d544aa462fd

5.3.6 German Law Table of Contents

Some longer laws have a Table of Contents starting on the title page, e.g., like this:²⁴

| | |
|--|--|
| Inhaltsübersicht Abschnitt 1 Gegenstand, Dauer und Gliederung der Berufsausbildung § 1 Staatliche Anerkennung des Ausbildungsbereichs § 2 Dauer der Berufsausbildung § 3 Gegenstand der Berufsausbildung und Ausbildungsrahmenplan § 4 Struktur der Berufsausbildung, Ausbildungsberufsbild § 5 Einsatzgebiet § 6 Ausbildungsplan | |
| Abschnitt 2 Abschlussprüfung Unterabschnitt 1 | |

This is an example of a list with only 1-line items, but with the wide space after “§ 1” etc. it looks like it would wrap to a hanging shape if any item were longer.

5.4 Philippine Laws

5.4.1 Section-header

Philippine laws often have Section-headers as in the following example:²⁵

| | |
|--|--|
| (P10,000.00) or an imprisonment exceeding six (6) years but not more than twelve (12) years and fine and imprisonment, shall prescribe in ten (10) years. CHAPTER IX Transitory and Final Provisions | |
| ARTICLE 95. Within two (2) years from the promulgation of this Code, all claims for arrears existing on or before December 31, 1974 shall be registered with the Council which shall cor | |

Presumably “Transitory and Final Provisions” is the header text of “CHAPTER IX”, but it is smaller and at significant distance, so by the rule “what it looks like”, we make 2 separate clusters.

5.4.2 List-item

If some items in the middle of a list are 1-liners and therefore don’t get the hanging shape, but they clearly belong to the same list, we also make them List-items.²⁶

| | |
|---|--|
| The Third Judicial District, of the Province of Ilocos Sur, including the subprovinces of Abra, and the Province of Ilocos Norte. The Fourth Judicial District, of the Province of La Union and the Mountain Province. The Fifth Judicial District, of the Province of Pangasinan. The Sixth Judicial District, of the Province of Tarlac and the Province of Nueva Ecija. | |
|---|--|

²⁴ Page adf02361b033556cf2ff8a25c0f5bc99efd6333eb7ffa09628447cec6131f28f

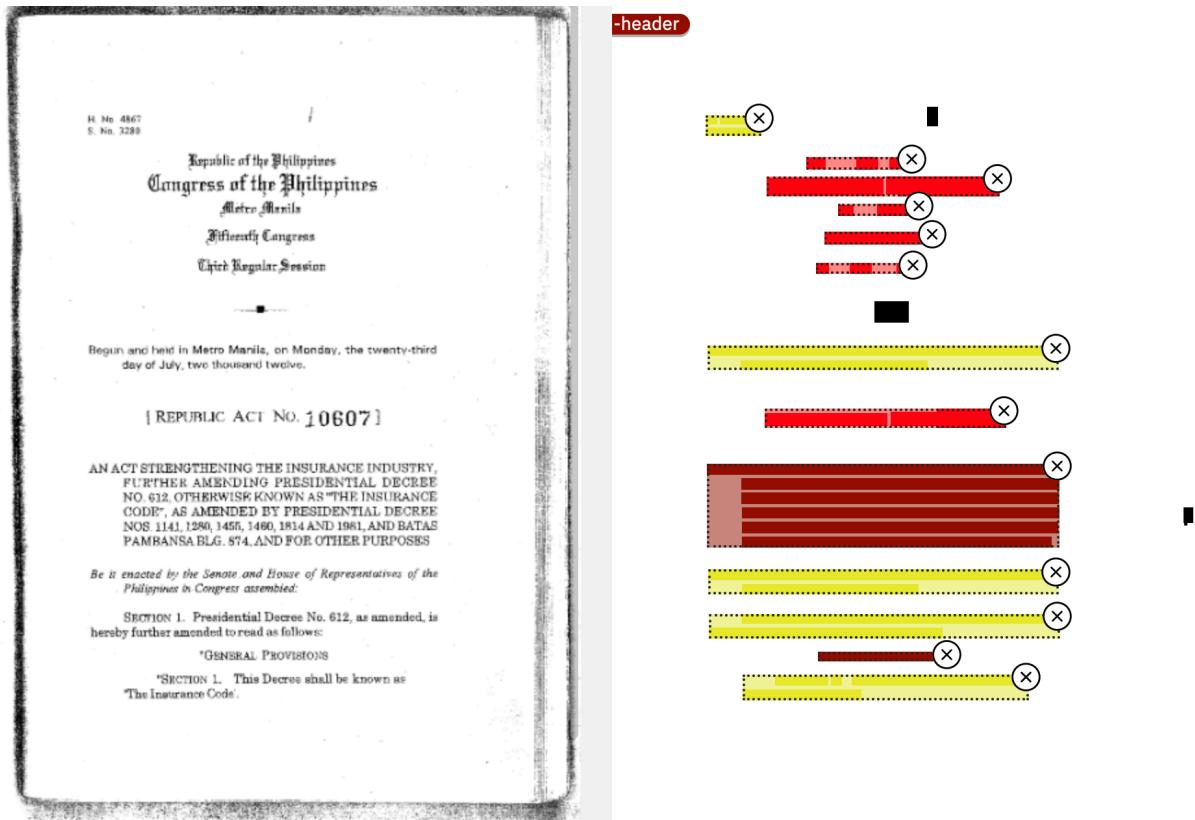
²⁵ Page b42a6bdacea139084d89a376c9bd8bea55b150f26a221ae056d13d1065217bbb

²⁶ Page f6e37bb6073dcf988060349176af2adf7a7b08b741e494ba43c35d35a1f493a

5.4.3 Philippine Law Title Pages

For the Philippine laws, we mostly have scanned pages. In the following page, the text boxes have been recognized quite well though.²⁷

We struggled a bit with this. “Congress of the Philippines” is largest, but it’s really the authoring body of all Philippine laws and not a characteristic title. Hence we added the next-largest elements, which help to identify this law uniquely. With “AN ACT STRENGTHENING ...” we decided on Section-header, as it is still larger than the rest and an important introduction.

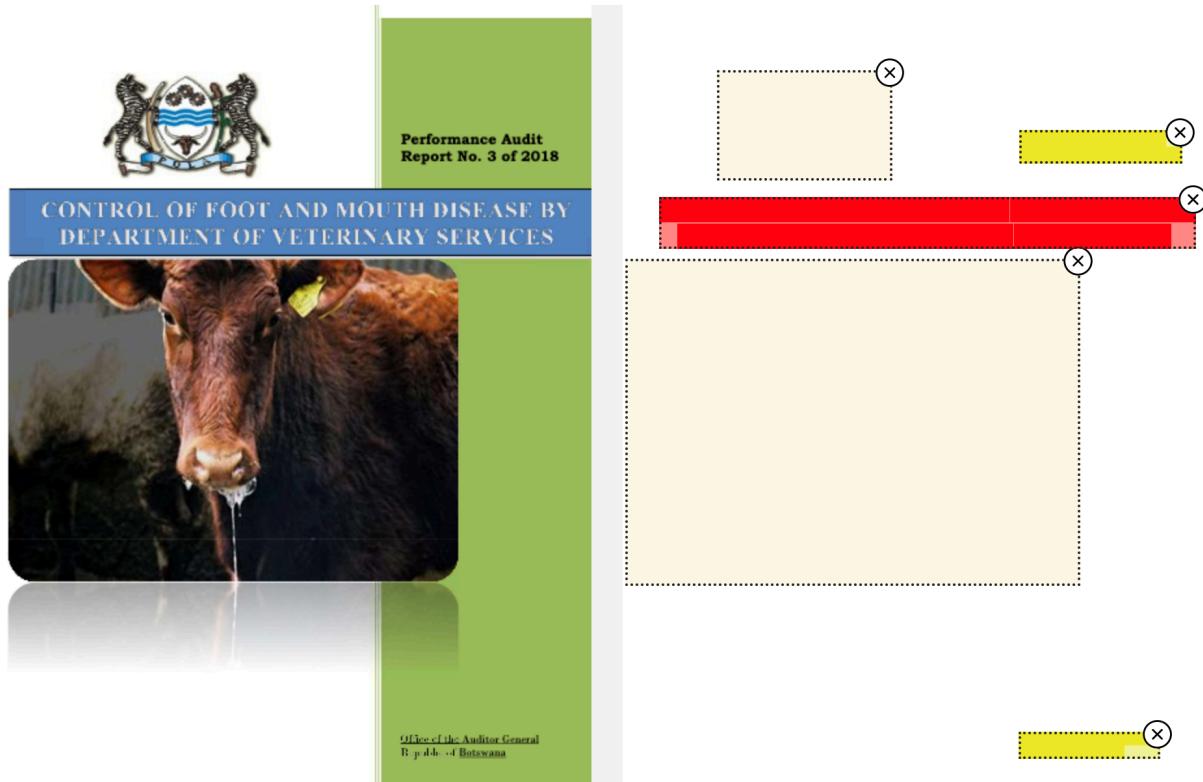


²⁷ Page ccc9a9470208ac0088f4316f84244942b28318707cfb067780a0bc7183c7ce58

5.5 Botswana Laws

5.5.1 Botswana Law Title Pages

We did not find a uniform collection of laws for Botswana, but only the constitution and certain guidelines and audit reports, of which we selected a few. We show the title page of an audit report.²⁸



5.6 Russian Laws

Russian laws are largely textual and do not expose specific annotation problems.

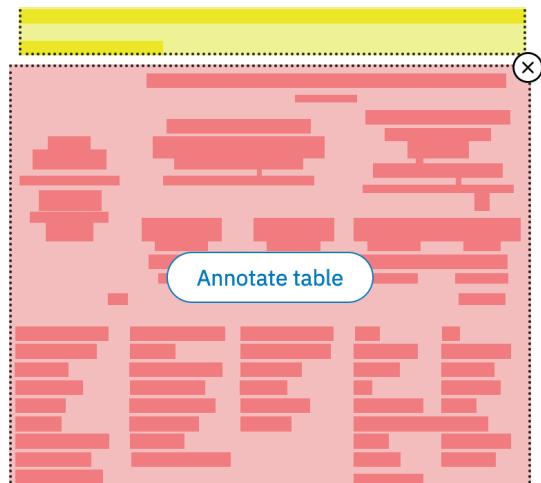
²⁸ Page 420184eeeb37513e7f35bcf26d9a0e883b4d382f948da29493461829d90e6fb5

5.6.1 Tables where Gridlines End Early

In Russian laws, there are several tables where the gridlines end early, but the text continues in the same columns. We make such text part of the Table anyway, as in the following example.²⁹

(0,0 или 6,0 процента на финансирование накопительной пенсии) по следующим тарифам:

| Тариф страхового взноса для лиц 1966 года рождения и старше | Тариф страхового взноса для лиц 1967 года рождения и моложе | | | |
|---|--|---|---|---|
| | Вариант пенсионного обеспечения 0,0 процента на финансирование накопительной пенсии | Вариант пенсионного обеспечения 6,0 процента на финансирование накопительной пенсии | на финансирование страховой пенсии | на финансирование накопительной пенсии |
| 20,0 процента на финансирование страховой пенсии, из них: 4,0 процента - солидарная часть тарифа страховых взносов; 16,0 процента - солидарная часть тарифа | 20,0 процента, из них: 4,0 процента - солидарная часть тарифа страховых взносов; 16,0 процента - | 0,0 процента - индивидуальная часть тарифа страховых взносов | 14,0 процента, из них: 4,0 процента - солидарная часть тарифа страховых | 6,0 процента - индивидуальная часть тарифа страховых взносов. |



²⁹ Page af7fe892efca8b2ad2f0ebfd061cc359f075ec1c43d197d57ac2fa707c6a5316

5.6.2 Russian Law Title Pages

The documents in our Russian laws collection seems to be at different levels, with distinct title pages. The following is a regulation, not a law.³⁰ As in other countries' laws, a problem is that the largest two lines are standard, not about the specific regulation. The more characteristic parts are the date and number in normal Text format, and the bold-faced text starting "О подписании". We compromised by making the latter a Section-header.



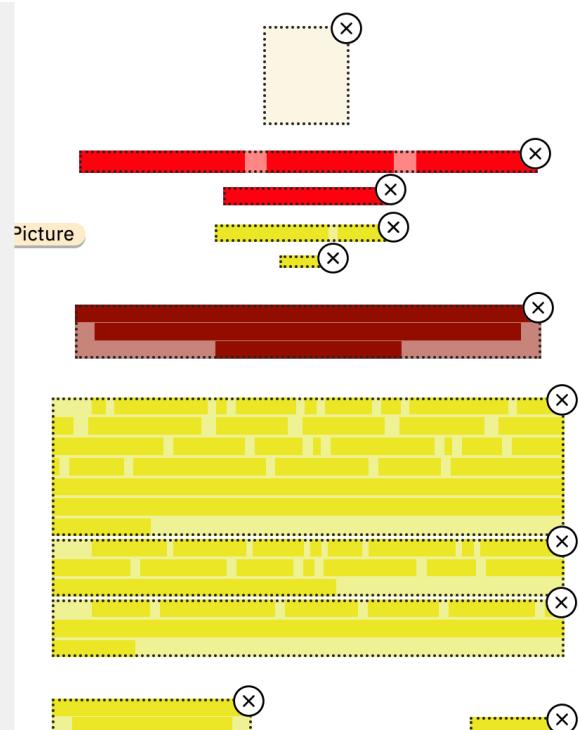
В соответствии с пунктом 1 статьи 11 Федерального закона "О международных договорах Российской Федерации" одобрить представленный Минтрудом России и согласованный с МИДом России и другими заинтересованными федеральными органами исполнительной власти проект Соглашения о сотрудничестве в сфере содействия занятости населения государств - участников Содружества Независимых Государств (прилагается).

Разрешить Минтруду России в ходе переговоров о подписании указанного Соглашения вносить в прилагаемый проект изменения, не имеющие принципиального характера.

Считать целесообразным подписать указанное Соглашение на очередном заседании Совета глав правительств Содружества Независимых Государств.

Председатель Правительства
Российской Федерации

М.Мишустин



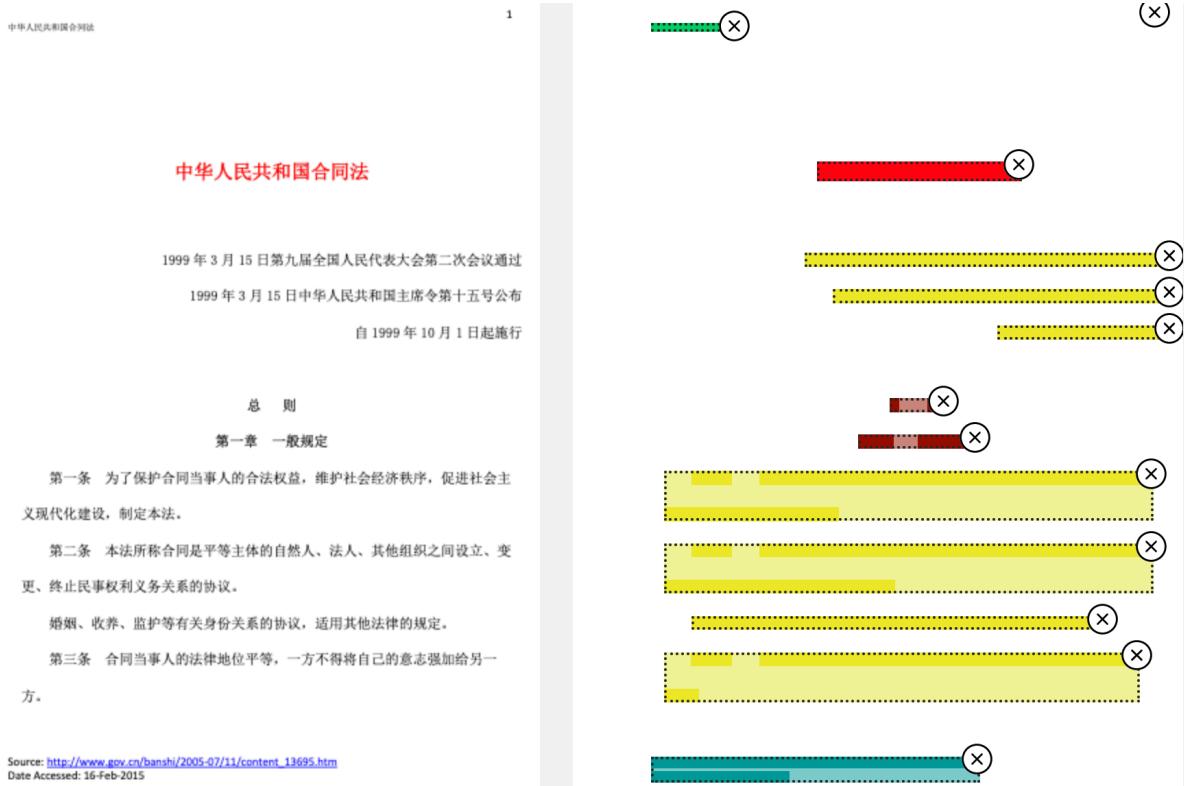
5.7 Chinese Laws

Normal pages of Chinese laws do not pose any specific challenges.

³⁰ Page c5a84ac02931ca16fcf165f81eae4ce7e80c97bdff2509636983862ebf4bf7b1

5.7.1 Chinese Law Title Pages

For the following Chinese law title page, we had to rely on what it looks like.³¹



³¹ Page b716a453086387f1b4df40ba1d68d0b91432d4a96f9c12ce271afde52536f290

5.7.2 Chinese Law Table of Contents

The following Chinese law title page contain a Table of Content.³² It looked like the apple-and-orange list example with only 1-line items, but significant distances from an identifier to main text, so we decided on List-items.

中华人民共和国网络安全法
(2016年11月7日第十二届全国人民代表大会常务委员会第二十四次会议通过)
浏览字号：大 中 小 来源：中国人大网 2016年11月7日 17:31:34

目 录

- 第一章 总 则
- 第二章 网络安全支持与促进
- 第三章 网络运行安全
 - 第一节 一般规定
 - 第二节 关键信息基础设施的运行安全
- 第四章 网络信息安全
- 第五章 监测预警与应急处置
- 第六章 法律责任
- 第七章 附 则

第一章 总 则

第一条 为了保障网络安全，维护网络空间主权和国家安全、社会公共利益，保护公民、法人和其他组织的合法权益，促进经济社会信息化健康发展，制定本法。

第二条 在中华人民共和国境内建设、运营、维护和使用网络，以及网络安全的监督管理，适用本法。

第三条 国家坚持网络安全与信息化发展并重，遵循积极利用、科学发展、依法管理、确保安全的方针，推进网络基础设施建设和互联互通，鼓励网络技术创新和应用，支持培养网络安全人才，建立健全网络安全保障体系，提高网络安全保护能力。

第四条 国家制定并不断完善网络安全战略，明确保障网络安全的基本要求和主要目标，提出重点领域网络安全政策、工作任务和措施。

Source: http://www.npc.gov.cn/npc/xinwen/2016-11/07/content_2001605.htm
Accessed: 28-March_2018

5.8 Japanese Laws

5.8.1 Japanese Law Title Pages

In the collection “Japanese laws”, we have original versions, English translations, and inline translations. We show the first half-page of one law in all 3 versions.³³

農薬を使用する者が遵守すべき基準を定める省令（暫定版）
(平成十五年三月七日農林水産省・環境省令第五号)

農業取締法（昭和二十三年法律第八十二号）第十二条第一項の規定に基づき、農薬を使用する者が遵守すべき基準を定める省令を次のように定める。

（農薬使用者の責務）

第一条 農薬を使用する者（以下「農薬使用者」という。）は、農薬の使用に関し、次に掲げる責務を有する。

一 農作物等に害を及ぼさないようすること。
二 人畜に被害が生じないようすること。
三 農作物等又は当該農作物等を家畜の飼料の用に供して生産される畜産物の利用が原因となって人に被害が生じないようすること。
四 農地等において栽培される農作物等又は当該農作物等を家畜の飼料の用に供して生産される畜産物の利用が原因となって人に被害が生じないようすること。

³² Page 44173ee2d452b91862f8e5efcd09b5379094fa6ae171208a59b362aafc022aa5
³³ Page 7670ad7ef97e848a3b134cf19d11a405b365f3e41dce8af8efcfb4346897027e

**Ministerial Order to Provide Standards Which Users of Agricultural Chemicals Must Comply With
(Tentative translation)**

(Order of the Ministry of Agriculture, Forestry and Fisheries and Ministry of the Environment No. 5 of March 7, 2003)

In accordance with the provisions of Article 12, paragraph (1) of the Agricultural Chemicals Regulation Act (Act No. 82 of 1948), a Ministerial Order to Provide Standards Which Users of Agricultural Chemicals Must Comply With is provided as follows.

(Responsibilities of Agricultural Chemical Users)

Article 1 Persons that use agricultural chemicals (hereinafter referred to as "agricultural chemical users") have the following responsibilities for the use of the agricultural chemicals:

- (i) to prevent harm to crops, etc.;
- (ii) to prevent causing damage to humans and animals;
- (iii) to prevent causing damage to humans through the use of crops, etc., or through the use of livestock products produced with those crops, etc. used for feed;
- (iv) to prevent causing damage to humans through the use of the crops, etc. cultivated in farmland, etc., or through the use of livestock products produced with those crops, etc. used for feed;

農薬を使用する者が遵守すべき基準を定める省令（暫定版）

**Ministerial Order to Provide Standards Which Users of Agricultural Chemicals Must Comply With
(Tentative translation)**

（平成十五年三月七日農林水産省・環境省令第五号）
(Order of the Ministry of Agriculture, Forestry and Fisheries and Ministry of the Environment No. 5 of March 7, 2003)

農薬取締法（昭和二十三年法律第八十二号）第十二条第一項の規定に基づき、農薬を使用する者が遵守すべき基準を定める省令を次のように定める。

In accordance with the provisions of Article 12, paragraph (1) of the Agricultural Chemicals Regulation Act (Act No. 82 of 1948), a Ministerial Order to Provide Standards Which Users of Agricultural Chemicals Must Comply With is provided as follows.

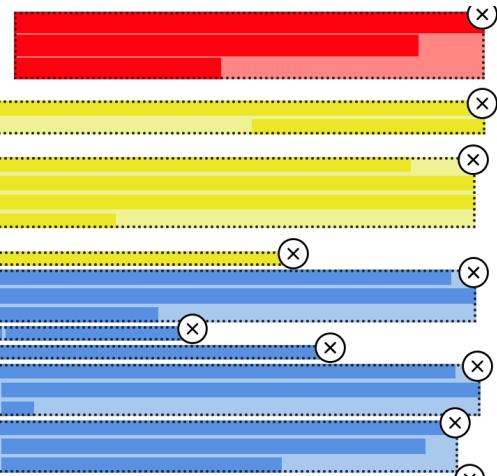
（農薬使用者の責務）

（Responsibilities of Agricultural Chemical Users）

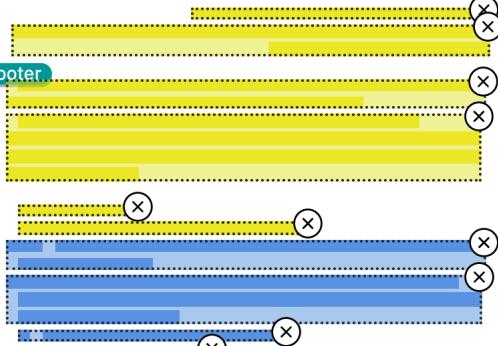
第一条 農薬を使用する者（以下「農薬使用者」という。）は、農薬の使用に關し、次に掲げる責務を有する。

Article 1 Persons that use agricultural chemicals (hereinafter referred to as "agricultural chemical users") have the following responsibilities for the use of the agricultural chemicals:

- 一 農作物等に害を及ぼさないようにすること。



Page-footer

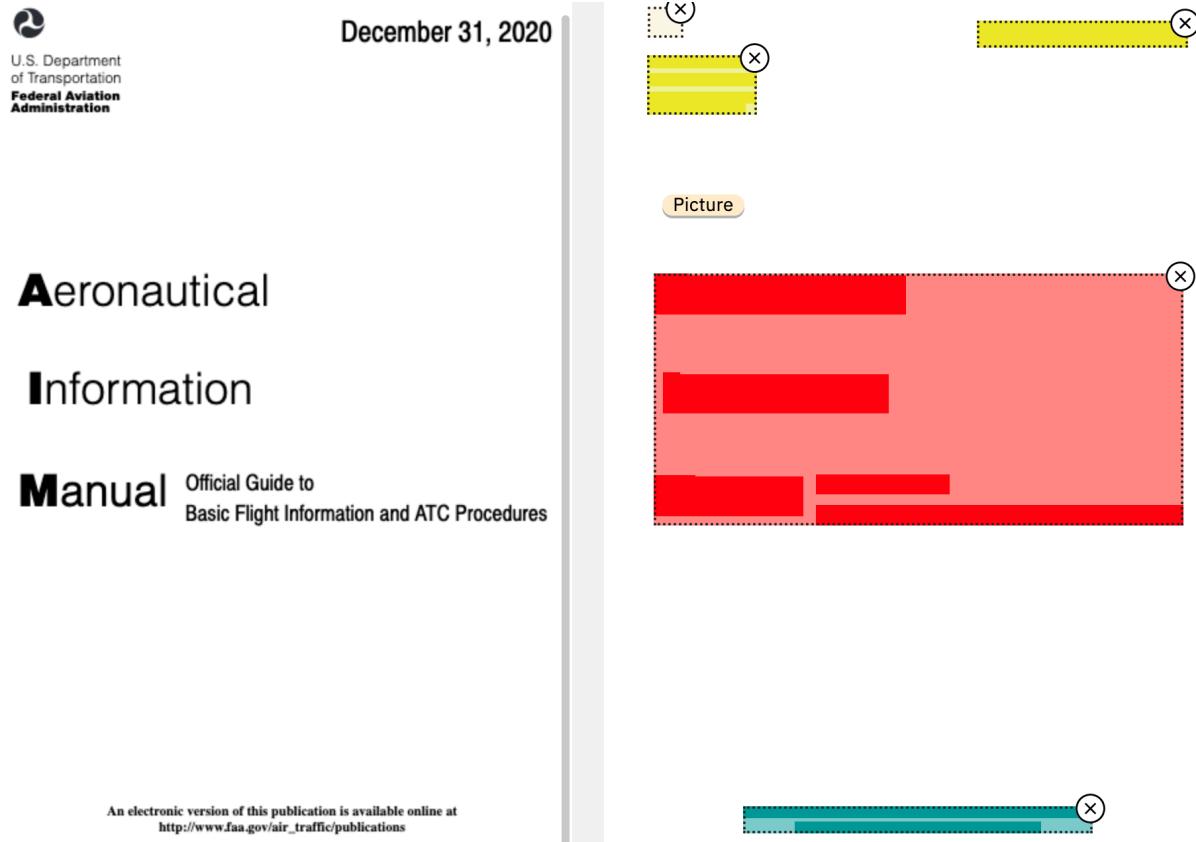


5.9 FAA Regulations

The FAA regulations are US aviation regulations. Their formats differ much more than those of higher-level laws, so for some borderline cases you may also look in Section 8 on free-form documents.

5.9.1 FAA Regulation Title Pages

In the first example, the main title is clear, but we had to decide what to do with the subtitle “Official Guide ...”. As we only use rectangles, it seemed best to include it in the title.³⁴

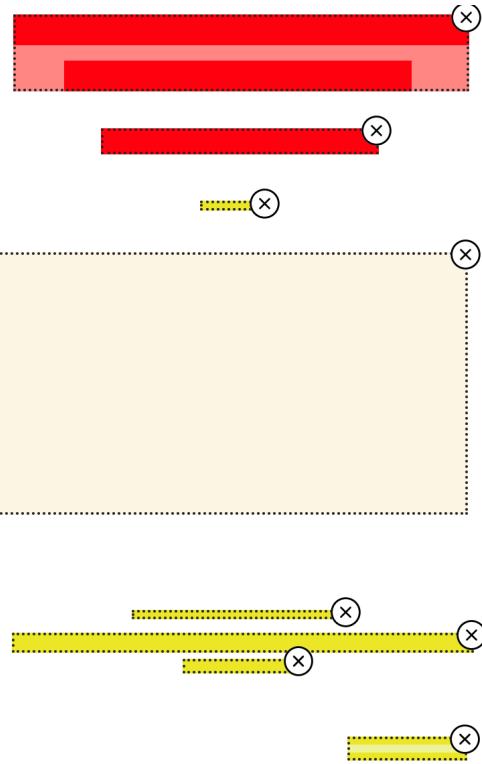


³⁴ Page 1847bb014d5e26b0b609e2fe8dbf14f43f239722de605b4fe34dfd94ef7e5acc

In the second example, a decision was whether to still consider “Pilot’s Manual” as a part of the title, as it is a little smaller. We decided on yes as it is only a little smaller, but on 2 clusters because of the lines in between, while the even smaller “Goodyear ...” below is similar to an authors line and thus Text.³⁵



Picture



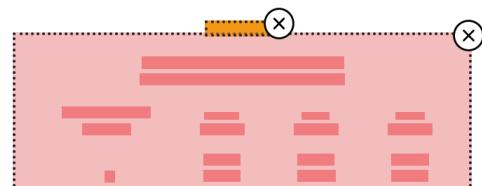
6 Patents

6.1 US Patents

6.1.1 Caption-like Text Continues after Table Line: Part of Table

Multiple patents contain tables starting like this:³⁶

| TABLE 3 (Individual results for electrostatic field measurements for Example 2) | | | |
|--|--------------------------|-------------------------|-------------------------|
| Individual disc number | F _{mean} (kV/m) | F _{min} (kV/m) | F _{max} (kV/m) |
| 1 | -43.7 | -59.5 | -21.2 |
| 2 | -43.7 | -60.4 | -24.5 |



Semantically, the two text lines after the first horizontal bar are caption text. However, “what it looks like” is that the table starts at this first horizontal bar, and only “TABLE 3” is the Caption, so that’s how it should be labeled.

³⁵ Page 96e7b57ddebc5fb55960335e6771f4d948532830c02dcf445e27a2864333a189

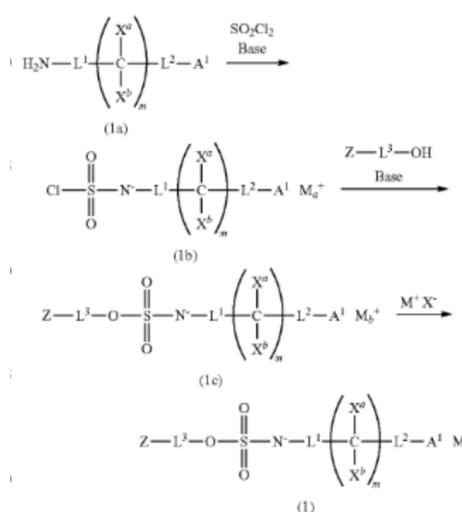
³⁶ Page 3fa4629199eaa5384f20c778af522b758721f68d3cb17f1e34e4bfb6345d0781

6.1.2 Rich Chemical Formulas: Pictures

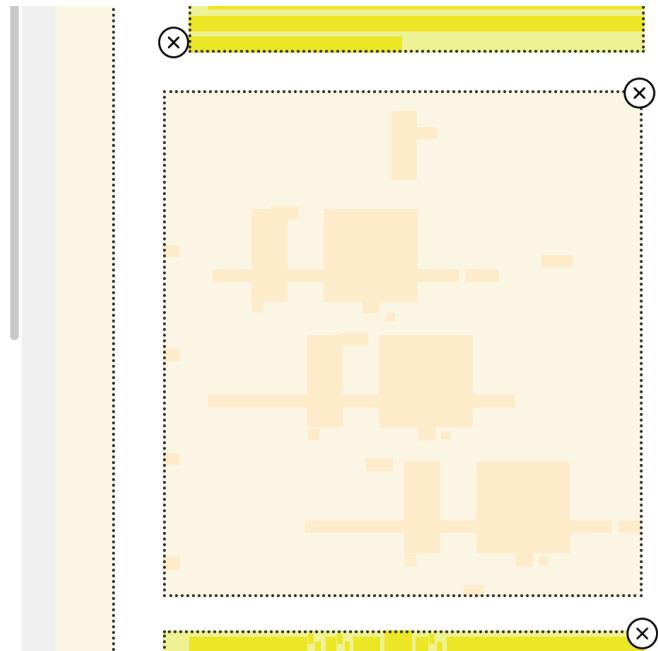
In particular our patent collections contain a lot of what's termed as a "formula" in chemistry, but they represent molecules or functional groups. We need to treat them as "Picture" because of their graphical appearance.

The case is clear when the graphics are large like this:³⁷

the following scheme, for example, although the synthesis route is not limited thereto.

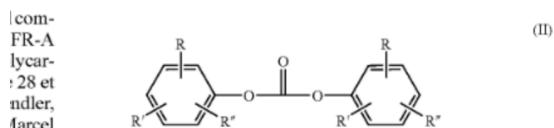


Herein Z, L¹, L², L³, A¹, X^a, X^b, m, and M⁺ are as defined

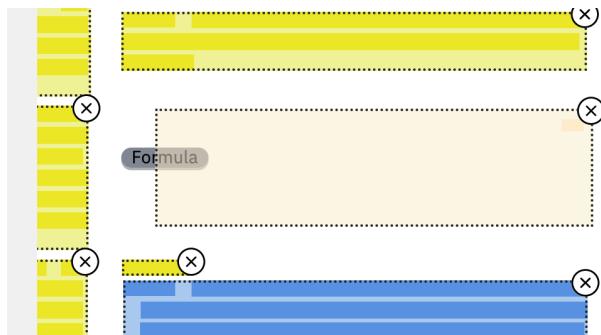


It is also rather clear in the following medium-sized example – it would be impossible to read this aloud as a text.³⁸ Note how we included the "Formula number" "(II)" with the Picture, like we do with formulas.

[0072] The diaryl carbonates suitable for the reaction with the dihydroxyaryl compounds are those of the general formula (II)



in which
[0073] R, R' and R'', independently of one another, are identical or different and represent hydrogen, linear or branched C₁-C₃₄-alkyl, C₇-C₃₄-alkylaryl or C₆-C₃₄-aryl,



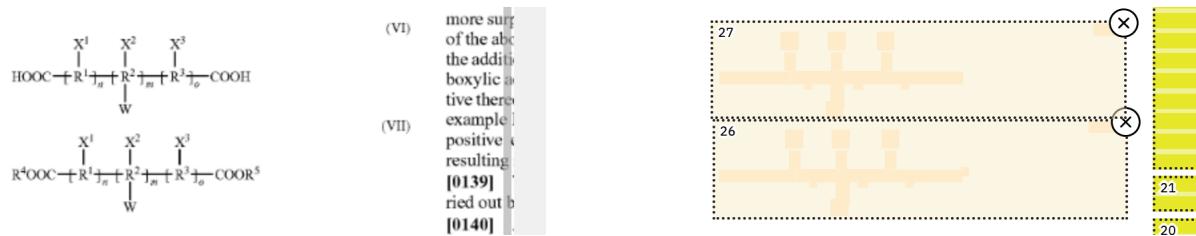
Multiple such molecule drawings may occur directly after each other. If they

- have different numbers, like "(VI)" and "(VII)" below,
- are not connected with any lines,
- and one can draw separate rectangles around them,

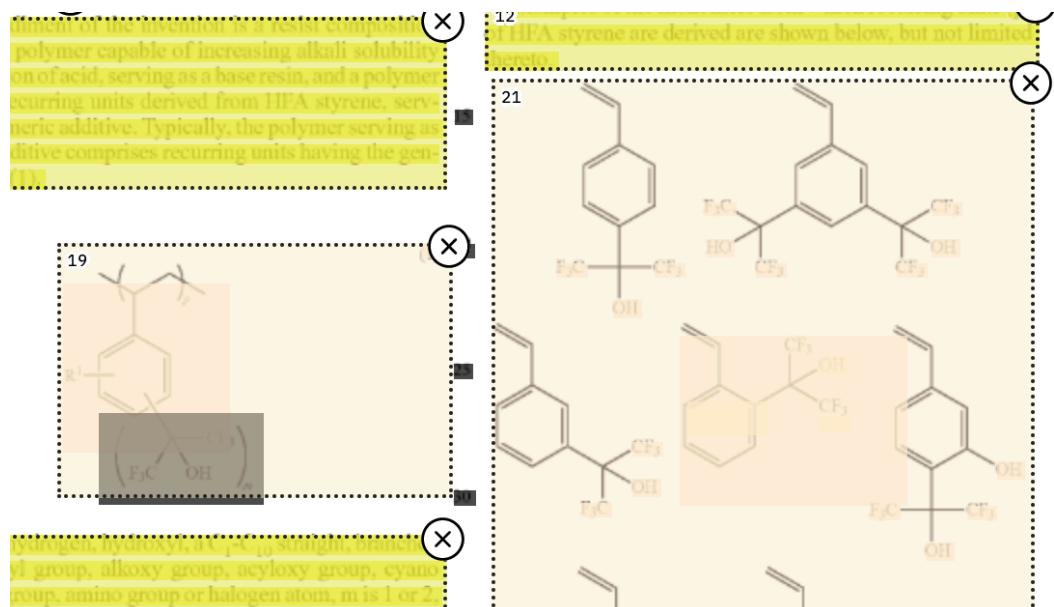
³⁷ Page b94fd02b3b81630b1b5cd1da2261717a8b48a6cd424ea78c23f4af857a6c75b

³⁸ Page 3800b8220f648e029bb45d9b5c904fc56ab4ad57209b042f31f99ed2d36d957

then they should be different Picture clusters, like here:³⁹



If there are no numbers or other text in between, and the drawings are close together, then make them one large picture, like here:⁴⁰



Tiny chemical formulas without branches, circles etc. should instead be labelled as a regular “Formula”, as there is nothing in it that “looks like” a molecule drawing. Example here:⁴¹

[0064] Dihydroxyaryl compounds suitable for the process according to the invention are those of the general formula (I)

HO-Z-OH

[0071] with
droxy
[0072] with
from

in which Z is an aromatic radical which has 6 to 34 C atoms and may contain one or more optionally substituted aromatic



³⁹ Page 8505d24049fa6dac0e4c8f2ba3d651077eb38258d96f5b8c0f6d71af11651e19

⁴⁰ Page fc05b99d801b32c85054052b1942beca175dd49539ffb9909c03857144e6ff79

⁴¹ Page 3800b8220f648e029bb45d9b5c904fc56ab4ad57209b042f31f99ed2d36d957

6.1.3 Column header: Also a Page-header

In many patents, the columns also have numbers, similar to page numbers. We also label these as Page-headers.⁴²



6.1.4 Line Numbering: None

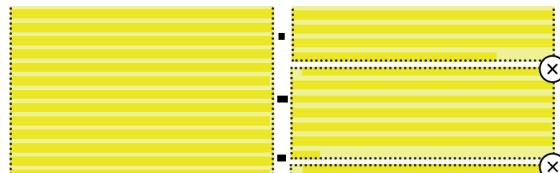
Many patents contain line numbers. As they don't really belong to the text, nor make sense for any other label, we keep those labelled as "None", e.g., here:⁴³

free carboxylic acids derived from a fermentation broth. An advantage of the present inventive is that one can use free carboxylic acids directly from a fermentation broth and generate corresponding esters therefrom without the need to isolate or purify the acids from the fermentation broth, as is necessary in conventional extractions from broth. In comparison to certain fermentation processes that neutralize or convert the carboxylic acids to their salts, the present process provides an easier way to isolate and extract carboxylic acids from a fermentation broth. The present process eliminates a need for titration and neutralization of the fermentation broth that can precipitate metal salts, and

in de minimis or trace amount of less than catalytic efficacy. In other words, no other acid catalyst is present, or is present at a level less than 10%, 5%, 3%, or 1% weight/weight relative to the carboxylic acid in the reaction.

An advantageous feature of the inventive process is that activation of the free carboxylic acid as an acyl halide (e.g., fluoride, chloride, bromide) or by using strong mineral acids is unnecessary. Acyl halides are inconvenient to use because these species are inherently reactive, have issues with stability, waste treatment, and can be cumbersome and costly to make.

In the present process, carbon dioxide functioning as a



If the text cells are wrongly parsed, so that a line number cannot be separated from a line, then the best solution would be NOT to submit, but report an error, e.g., here for the 3rd of the numbers:



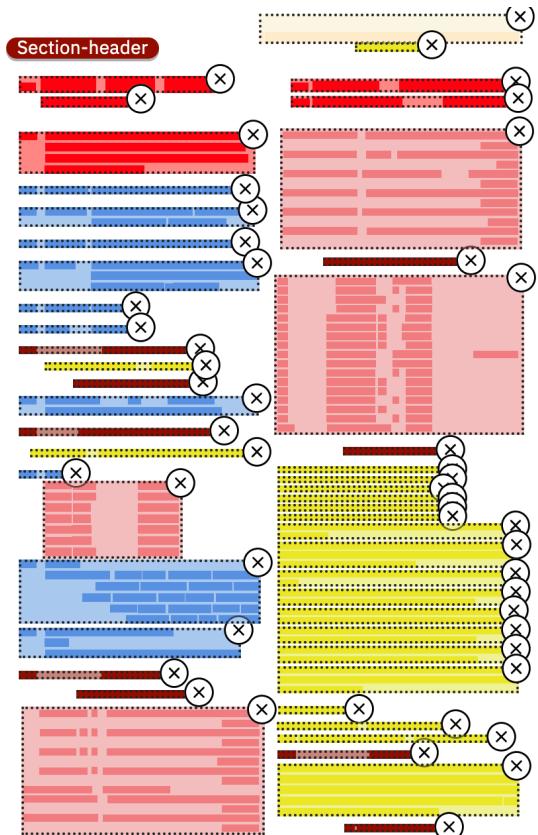
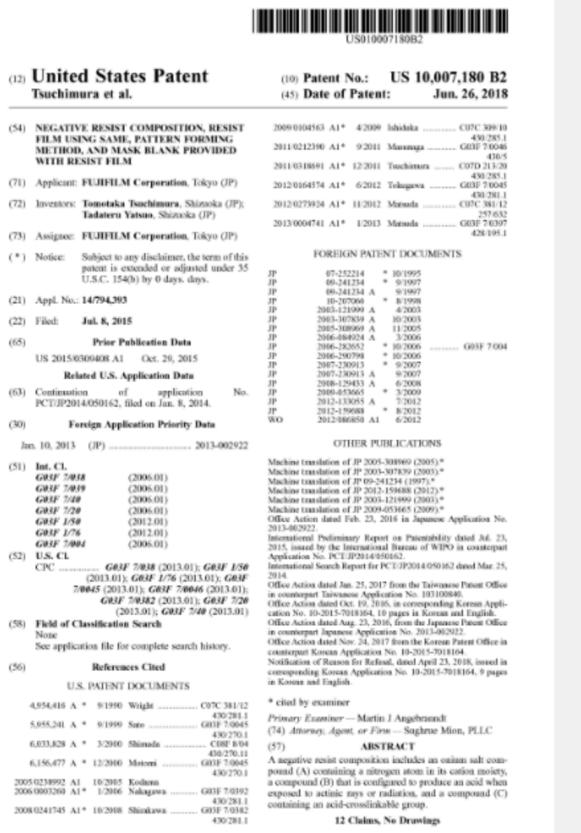
As a compromise: If a page has **at most 2** paragraphs that merge in the line numbers, then we **submit the page and** make an error report, with a note "line numbers joined".

⁴² Page c9fa1f9ba53d0401f478d10a05d6778184bac5de7df64534379a2a30a7695bfd

⁴³ Page c9fa1f9ba53d0401f478d10a05d6778184bac5de7df64534379a2a30a7695bfd

6.1.5 US Patent Title Pages

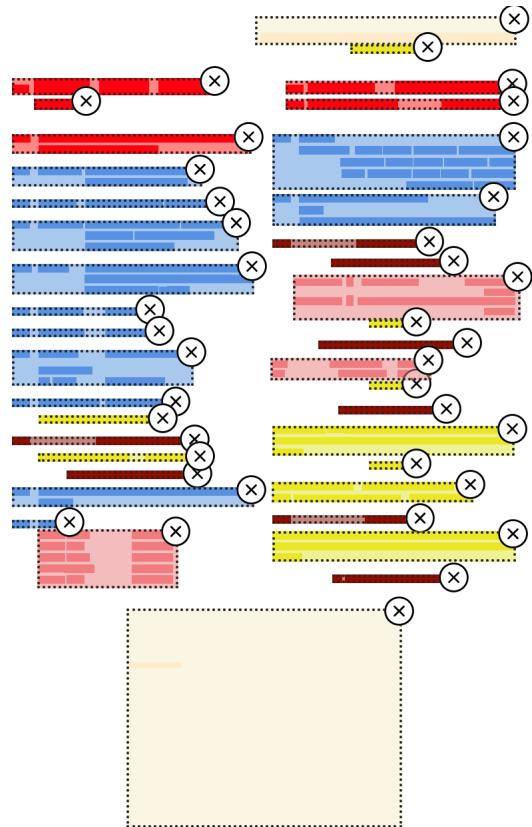
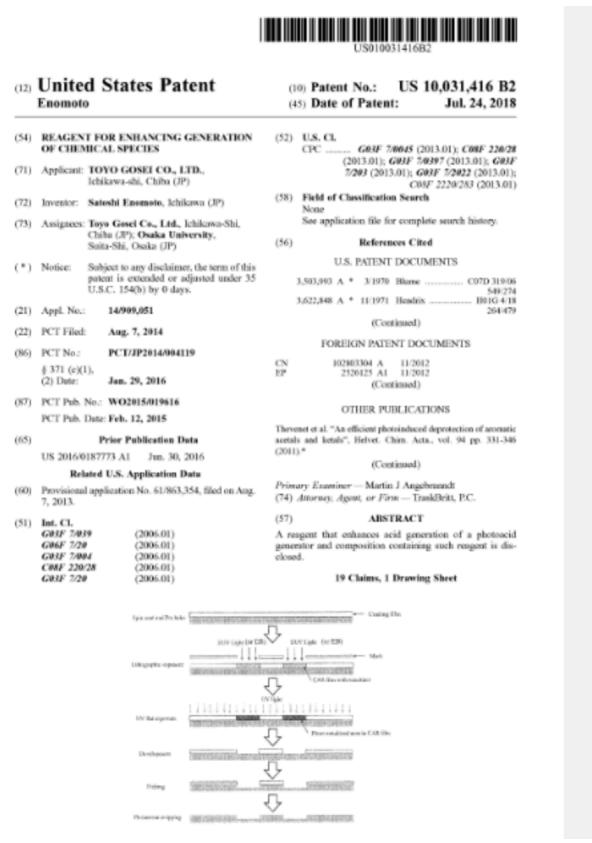
US Patents have complex title pages with a lot of meta-information about the patent. Fortunately, this is similar across most patents. Here is an example page.⁴⁴ Decisions are explained below.



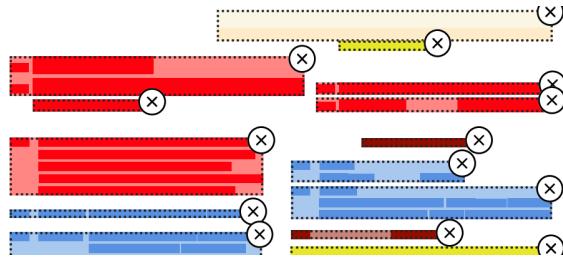
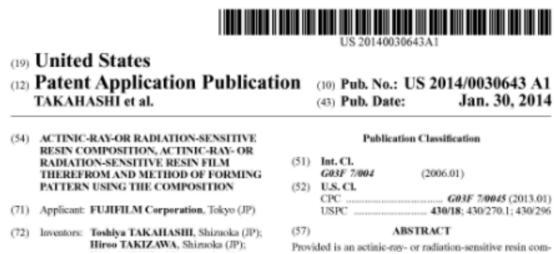
- Clearly, “United State Patents” gets label “Title”. However, it is the same for all US patents, and the main identifying information is the Patent No. and “NEGATIVE RESIST ...”. Thus we also make these paragraphs “Title”, and those in between with the same font size.
- There are several slightly untidy tables; we still made them “Table”.
- In the area starting “(71) Applicant ...” the multiline paragraphs are hanging, i.e., “List-item”. We made related 1-line paragraphs “List-items” too, as they are in the same list and look like they’d also become hanging if they were longer.
- However, those paragraphs where the main part is centered, like “Prior Publication Data”, instead of starting where the text of the list items starts, were choose “Section-header”.
- “12 Claims, No Drawings” does not really head anything, but it looks so much like a section-header on this page that we took the “label it as it looks” rule and made it “Section-header” too.

⁴⁴ Page 2cdd7147e85a2def39d461bf7181c4f6427b85d19f5b64bb7af80c59e5bcca5d

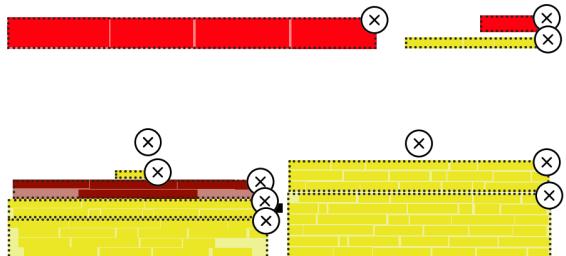
We show a second example so that one can see how the elements repeat.⁴⁵ Only here the meta-data were shorter, so that a real Picture fitted on the bottom of the page.



For a patent application (not yet a patent), the only difference is that “United States Patent Application” is a 2-line cluster:⁴⁶



An older patent from 1973 had a different title section.⁴⁷



⁴⁵ Page fc9fa610cef285c716ac466074a4016bad0d939f6513bd7b06d938b332477e6f

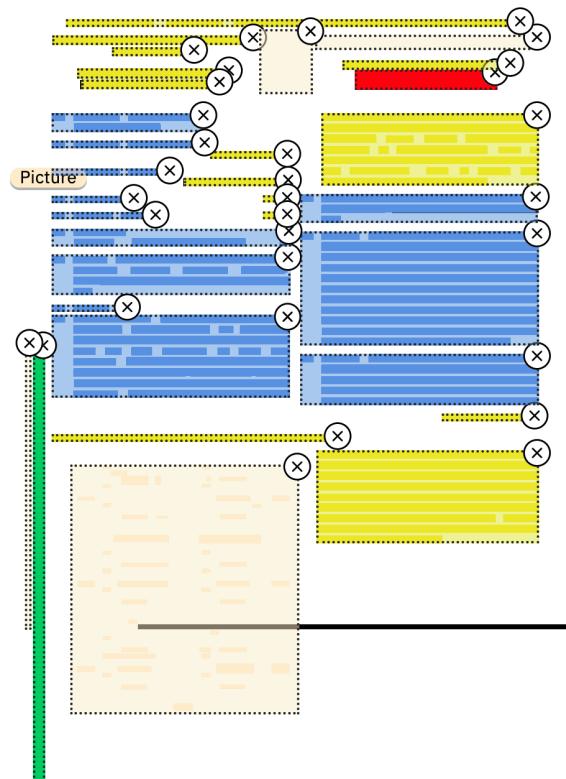
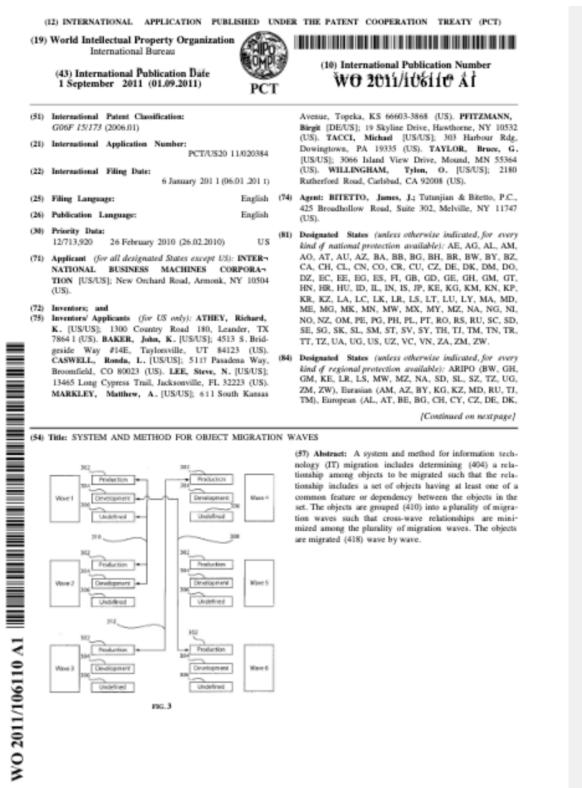
⁴⁶ Page 114768943e115fe05e8ff04b825b13adea5538af99f9e62b8efd6b8ca7211eb5

⁴⁷ Page 350329bd01f5f2ef9fcecc1d7a5eabd983f56940f7aa0867965e9000b6fb1c8ff

6.2 WIPO Patents

6.2.1 WIPO Patent Title Pages

For WIPO patents, similar considerations as for US patents apply, as shown in the next page:⁴⁸



- The patent number is significantly larger than any other text, so we made this the sole “Title”.
- The meta-data between the 2 horizontal lines seem to be hanging whenever they have more than 1 line, so we made them all “List-item”.
- However, where a part of such an item is geometrically too far from the start to be part of the same cluster, e.g., in a separate line and on the right, or the second column, this is “Text”.
- I would have preferred not to make “PCT” part of the world-picture, but there was no text box. We might watch once we have more WO-patents whether that is always the case.

7 Scientific Articles

Scientific articles do not have one single format, but there are many conventions that make them easier to annotate than some other document categories. You will already be aware of

⁴⁸ Page 684a4458ab643f303f95edc9f1e8b3257205c8f7dce31cdf4a4fd609d74fbef

most of them from the common part of these guidelines. Only a few special cases follow here.

7.1 Section-header vs. List-item

In the following example, “Section-header” is used.⁴⁹ It may look superficially like a list, but as there is always a line break after the first line (like after “Brightness Temperature”), such a line is a paragraph by itself and not hanging. Then its bold font makes it a section header.

- **Brightness Temperature**

The core brightness temperature (T_B) limit of all initial TANAMI sources was calculated. The high end of the distribution of calculated brightness temperatures is dominated by quasars and the low end by BL Lacertae objects and galaxies. Of the 43 sources in the sample, 13 have a maximum T_B below the equipartition value of 10^{11} K [17], 29 below the inverse Compton limit of 10^{12} K [10], putting about a third of the values above this limit. Doppler boosting is the most likely reason behind these high values though a variety of exotic mechanisms are also possible. There is no significant difference in the brightness temperature distribution of LBAS and non-LBAS sources. Many of the highest brightness temperature sources are not detected by LAT yet which is counterintuitive since they are expected to have higher Doppler factors.

- **Luminosities**

The core and the total luminosity was calculated for all 38 initial TANAMI sources that had published redshifts, assuming isotropic emission. There is no significant difference in the distribution of luminosities of LBAS and non-LBAS sources. On the other hand, there is a clear relationship between luminosity and optical type with quasars dominating the high luminosity end of the distribution, galaxies dominating the low luminosity end while the BL Lacertae objects fall in between,

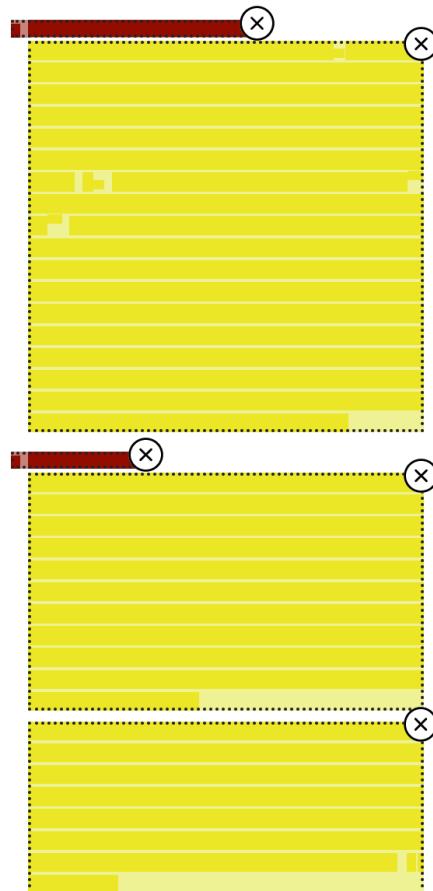
None of the five most distant and most luminous sources have been detected by Fermi in its first 3 months of operation. Intriguingly, none of the nine most luminous jets (difference of total and core luminosities) are detected. If this persists it would suggest a completely unexpected anti-correlation between jet luminosity and γ -brightness.

The TANAMI high resolution map of the sky. Such multiwavelength studies will revolutionize our understanding of the era of *Fermi*.

Data from the analysis of several individual studies of the growth of the galaxy, providing insight into the radiation from the component spectrum to greatly expand our help answer.

VI.

We are grateful to Julie McEnery, who contributed to the TANAMI program and the *Fermi*/LAT AG. This research has been funded by the German Federal Ministry of Education and Research (BMBF), Deutsches Zentrum für Luft- und Raumfahrt (DLR), number 50OR08, United States National Science Foundation, and the European Space Agency. Fermi has made use of the Galactic Database and Simulation Laboratory under contract with the Space Administration. (operated at CERN). Fermi has made use of the



7.2 Formula

7.2.1 Number on Separate Line

If a formula number got squeezed into an extra line, it still belongs to the formula:

⁴⁹ Page b7508cfbaa6ac3e79de7bd94c057d0171aac7eeb62a49354926e853de4b92261

can entropy productions. The approximation applied to Eq. A5 yields

$$\rho_{ij} = g_{ij} \exp \left(\frac{1}{2} (-2 + \mu_i + \nu_j + \mu_j + \nu_i + \Delta S_{ij} - \Delta S_{ij}^0) \right) \quad (\text{A8})$$

Using the result of Eq. A7 and the definition $\eta \equiv \sum_k \eta_k$ we obtain

7.2.2 Several Numbers in Multi-Line Formula

In the following example, logically the second line still belongs to the same formula, because it is preceded with the “ \approx ” sign; however, there are two formula numbers “(7)” and “(8)”. For follow-on processing, it would be good to treat this as one formula.

fluctuations using Widom's particle insertion

$$\beta \Delta \mu_v = -\ln P_v(0) \quad (7)$$

$$\approx \frac{\rho_B^2 v^2}{2 \langle (\delta N)^2 \rangle_v} + \frac{1}{2} \ln (2\pi \langle (\delta N)^2 \rangle_v), \quad (8)$$

where the second line is obtained using the Gaussian ap-

Formula

7.3 List-item

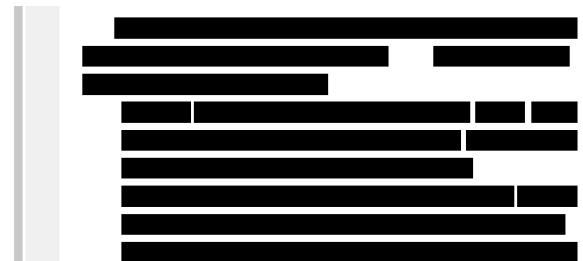
Scientific articles only contain few lists in the main text, and those are usually easy to treat with our general list rule.

7.3.1 Lists with Missed Hanging Shape: Error

The step that produces the black boxes from the original PDF is also an AI system. It sometimes misses list identifiers, as in the following example. Thus the black boxes don't look hanging, while the real text is. This should be reported as an error, and not submitted. (Because the AI system will see it as on the left, but you cannot draw the correct List-item rectangles.)

Like in a physical marketplace, the main purpose of an electronic marketplace is to bring potential *sellers* and *buyers* together:

- Sellers *offer* their goods and buyers *order* these goods; together this is a two-party *negotiation*, sometimes ending with an *agreement*.
- Both seller and buyer might need certain *certificates*. For instance, a buyer might only want to buy from sellers that are accredited with a well-



7.3.2 Citations: Treat with the Normal List Rules

Almost all scientific articles have references (citations) at the end (or before an appendix). We also treat those with the general list rules. A few examples follow, first three with the hanging shape:

ults
fol-
AL
is publicly available).

References

- [1] C. M. Bishop. *Pattern recognition and machine learning*, page 229. Springer-Verlag New York, 2006. [6](#)
- [2] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012. [9](#)
- [3] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *NIPS*, pages 2852–2860, 2012. [1, 2, 4, 7](#)
- [4] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. *arXiv preprint arXiv:1412.1283*, 2014. [9](#)
- [5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, et al. Long-term recurrent image models for street scene flow prediction. In *CVPR*, 2015. [10](#)

16

For more information

To learn more about this IBM Institute for Business Value study, please contact us at ibv@us.ibm.com. Follow @IBMVIEB on Twitter and for a full catalog of our research or to subscribe to our monthly newsletter, visit ibm.com/ibv.

Access IBM Institute for Business Value executive reports on your mobile device by downloading the free “IBM IBV” app for your phone or tablet from our app store.

The right partner for a changing world

At IBM, we collaborate with our clients, bringing together business insight, advanced research and technology to give them a distinct advantage in today's rapidly changing environment.

IBM Institute for Business Value

The IBM Institute for Business Value, part of IBM Global Business Services, develops fact-based strategic insights for senior business executives around critical public and private sector issues.

Notes and sources

- 1 Schumpeter, Joseph. *Capitalism, Socialism and Democracy*. Routledge, London & New York, 1994. 2003 edition, pp. 83-84.
- 2 For ease of reading, we have referred to CEOs who head enterprises that are torchbearers as Torchbearer CEOs, and to CEOs who head enterprises that are Market Followers as Market Follower CEOs.
- 3 Davison, Steven, Martin Horner and Anthony Marshall. “The new age of ecosystems: Redefining partnering in an ecosystem environment.” IBM Institute for Business Value, July 2014. www.ibm.com/business/value/ecosystempartnering
- 4 Moss, Stephen. “End of the car age: how cities are outgrowing the automobile.” *The Guardian*, April 28, 2015. www.theguardian.com/cities/2015/apr/28/end-of-the-car-age-how-cities-outgrow-the-automobile/
- 5 Wei Han Wong. “Financial giants take on fintech players.” *The Straits Times*, October 10, 2015. www.straitstimes.com/business/banking/financial-giants-take-on-fintech-players/; “Poland leading the way in consumer banking Fintech innovation.” *Daily Fintech*, June 2, 2015. <http://bankinnovation.net/2015/06/poland-leading-the-way-in-consumer-banking-fintech-innovation/>

their help and comments. We also thank Alexis Conneau, Duyu Tang and Zichao Zhang for providing us with information about their methods.

References

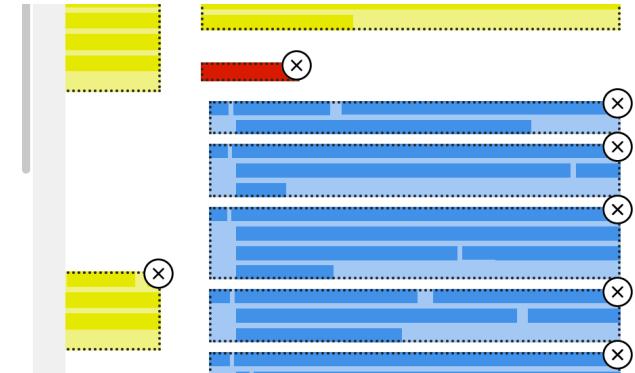
- [Agarwal et al.2014] Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. 2014. A reliable effective terascale linear learning system. *JMLR*.
- [Collobert and Weston2008] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multi-task learning. In *ICML*.
- [Conneau et al.2016] Alexis Conneau, Holger Schwenk, Loïc Barrau, and Yann Lecun. 2016. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*.

Correspondingly, if the individual references are not hanging, they are labeled “Text”:

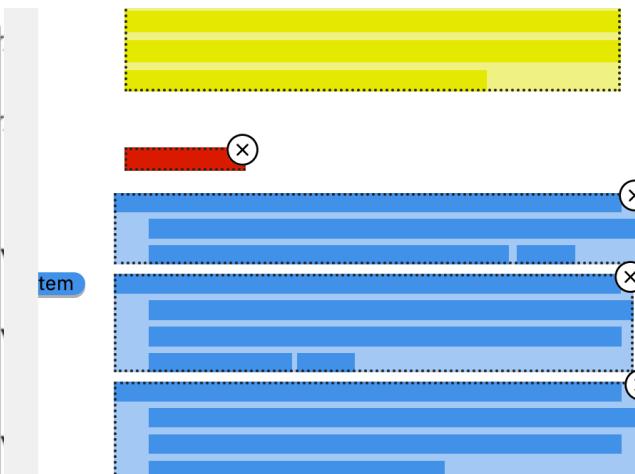
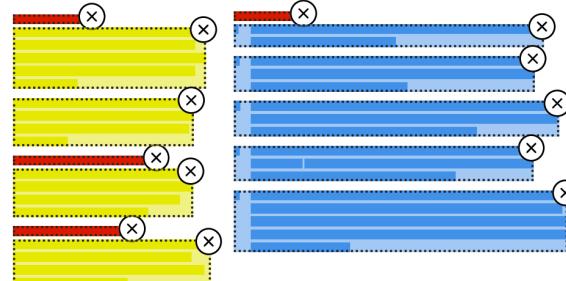
the Moore Sloan Foundation, NYU Center for Data Science, and NIH grant R01 EY027964 to DGP.

References

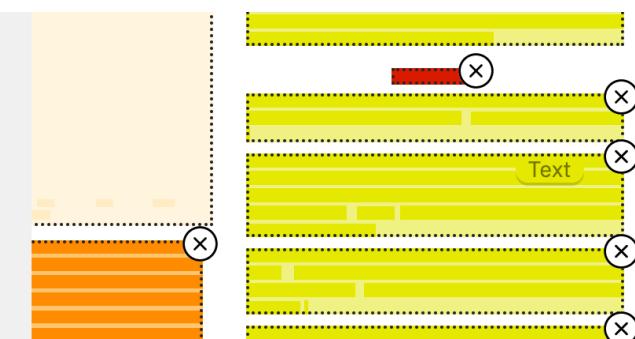
- Allard, R.; Faubert, J.; and Pelli, D. G. 2015. Editorial: Using Noise to Characterize Vision. *Frontiers in Psychology*.
- Asvestopoulou, T.; Manousaki, V.; Psistakis, A.; Smyrnakis, I.; Andreidakis, V.; Aslanides, I. M.; and Papadopoulou, M. 2019. DysLexML: Screening Tool for Dyslexia Using Machine Learning. *CoRR* abs/1903.06274. URL <http://arxiv.org/abs/1903.06274>.
- Basten, U.; Biele, G.; Heekeren, H. R.; and Fiebach, C. J. 2010. How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*.
- Cho, H.; Rybski, E. P.; Bar-Hillel, A.; and Zhang, W. 2012.



(x)

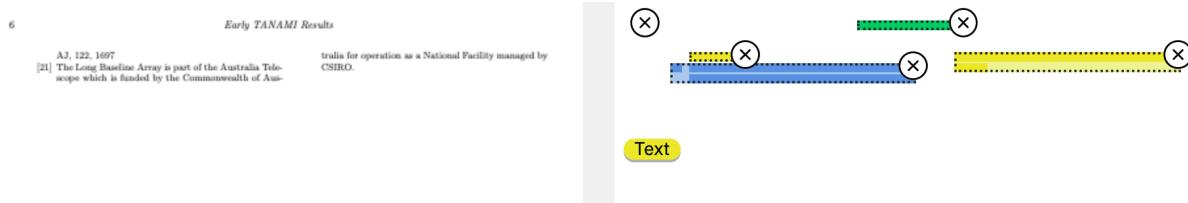


tem



7.3.3 Continuation of List-item at Top of Page: Text

Sometimes a page starts with the end of a paragraph that was likely a list item, like the small line “AJ ...” in the following example:⁵⁰



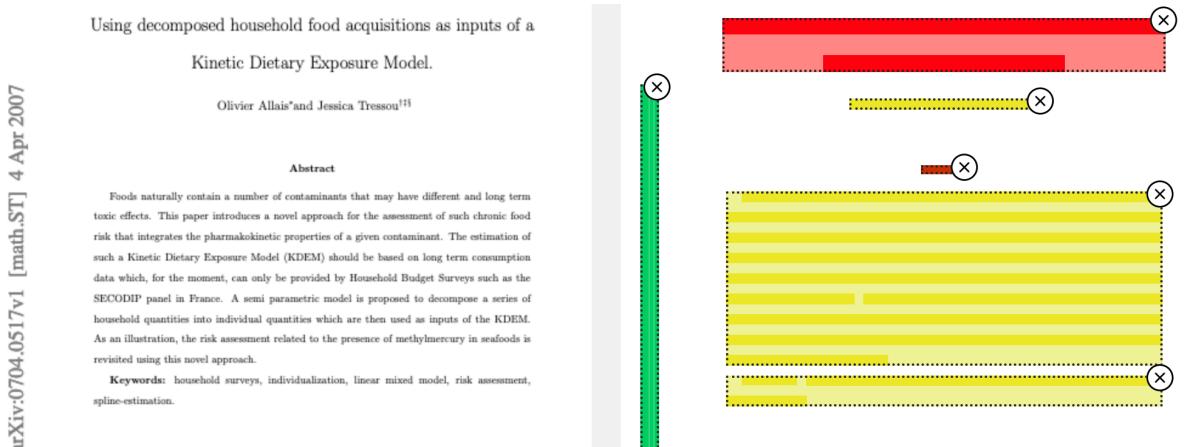
Although one can guess here that it is the end of another list item, it could also be the end of an inner paragraph in a list item. By “what it looks like”, it is therefore Text, and so is the (partial) paragraph in the right column.

7.4 Scientific Title Pages

In scientific articles, the title is typically at the top of the first page (except potentially for a small Page-header) and large, and followed by authors, as in the following two examples. The authors do not belong to the title.



Our next example has a large distance between the two Title lines.⁵¹ Nevertheless, they are a single Title cluster. The affiliations (addresses) of the authors are normal Footnotes, and we have an example of a sideways Page-header.



⁵⁰ INRA-CORELA, Laboratoire de recherche sur la consommation, Ivry sur Seine, France.
⁵¹ INRA-MétoRisk, Méthodologies d'analyse des risques alimentaires, Paris, France and Hong Kong University of Science and Technology, Department of Science and Sytem Management, Hong Kong.

⁵¹The second author research is in part supported by Hong Kong Research Grant #601906.

⁵¹Corresponding author: Dr Jessica Tressou; Mail address: Hong Kong University of Science and Technology, ISMT, Clear Water Bay, Kowloon, Hong Kong; Email: tressou@ust.hk.

⁵⁰ Page a7bdbfee9caf9f6439a572334b217a5e88633ab36128ebb7aabb917363225ac1

⁵¹ Page f09f2d1e43da93c41d87cb02e543946ef8d9eeeb83614b2a648ff5b81e039779

In the following page, we had to decide how many clusters to make among the authors and their affiliations (addresses). If authors are on one line, and affiliations on other lines, we separate those. Furthermore, as the 4 affiliations are written like footnotes (starting with little numbers), we made one cluster for each address.⁵²

arXiv:1001.0009v1 [q-bio.BM] 30 Dec 2009

Jamming proteins with slipknots and their free energy landscape

Joanna L. Sulkowska¹, Piotr Sulkowski^{2,3,4} and José N. Onuchic¹¹ Center for Theoretical Biological Physics,

University of California, San Diego,

9500 La Jolla Blvd, La Jolla, CA 92093

² Physikalisches Institut und Bethe Center for Theoretical Physics,

Universität Bonn, Nussallee 12, 53115 Bonn, Germany

³ California Institute of Technology, Pasadena, CA 91109,⁴ Institute for Nuclear Studies,

Hoza 69, 00-681 Warsaw, Poland

Theoretical studies of stretching proteins with slipknots reveal a surprising growth of their unfolding times when the stretching force crosses an intermediate threshold. This behavior arises as a consequence of the existence of alternative unfolding routes that are dominant at different force ranges. Responsible for longer unfolding times at higher forces is the existence of an intermediate, metastable configuration where the slipknot is jammed. Simulations are performed with a coarsened model with further quantification using a refined description of the geometry of the slipknot. The simulation data is used to determine the free energy landscape (FEL) of the protein, which supports recent analytical predictions.

PACS numbers: 87.15.ap, 87.14.E-, 87.15.La, 82.37.Gk, 87.10.+e

The large increase in determining new protein structures has led to the discovery of several proteins with complicated topology. This new fact has raised the question if their energy landscape and the folding mechanism is similar to typical proteins. One class of such proteins includes knotted proteins, which comprise about 1% of all structures deposited in the PDB database [1]. A related class of proteins contains more subtle topometric configurations called slipknots [3, 4]. Recent theoretical studies using structure-based models (where native contacts are dominant) suggest that slipknot-like conformations act like intermediates during the folding of knotted proteins [5]. This entire new mechanism is consistent with energy landscape theory (FEL) and the funnel concept [7, 8]. It was shown that the slipknot formation occurs at the transition between two conformations along the folding coordinate. Considering recent results on folding kinetics, additional information about the landscape was obtained by mechanical manipulation of the knotted protein with atomic force microscopy [9] both experimentally in [10, 11] and theoretically in [12, 13, 14]. For example, [12] it has been shown that unfolding proceeds via a series of jumps between various metastable conformations, a mechanism opposite to the smooth unfolding in knotted homopolymers.

Motivated by these early results, we now propose a unified framework for the mechanical unfolding of proteins with slipknots. In this Letter this question is addressed by explaining the role of topological barriers along their mechanical unfolding pathways. Supported by our previous results that knotted proteins can still have a minimally frustrated funnel-like energy landscape, structure-based theoretical coarse-grained models are used [15] to analyze the behavior of a slipknot protein under stretching. Studies are performed for the α/β class protein thymine

dine kinase (PDB code: 1e2i [17]).

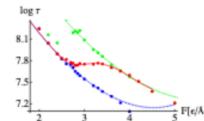
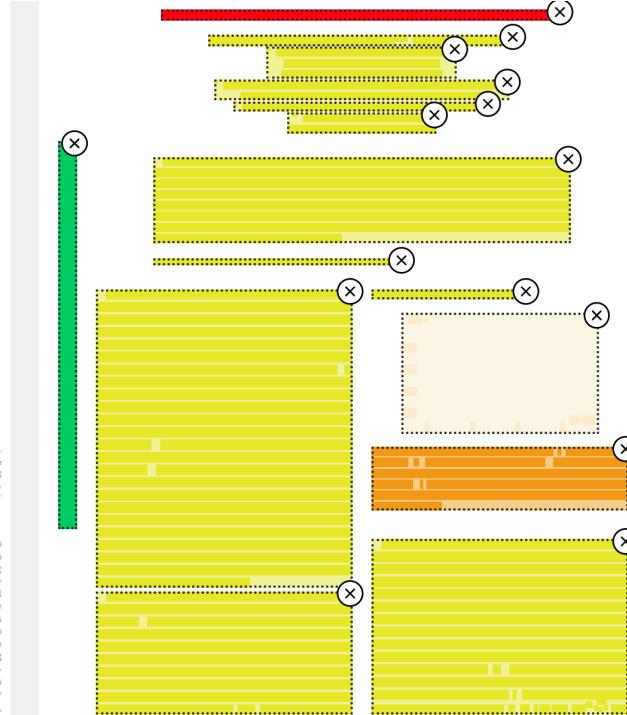


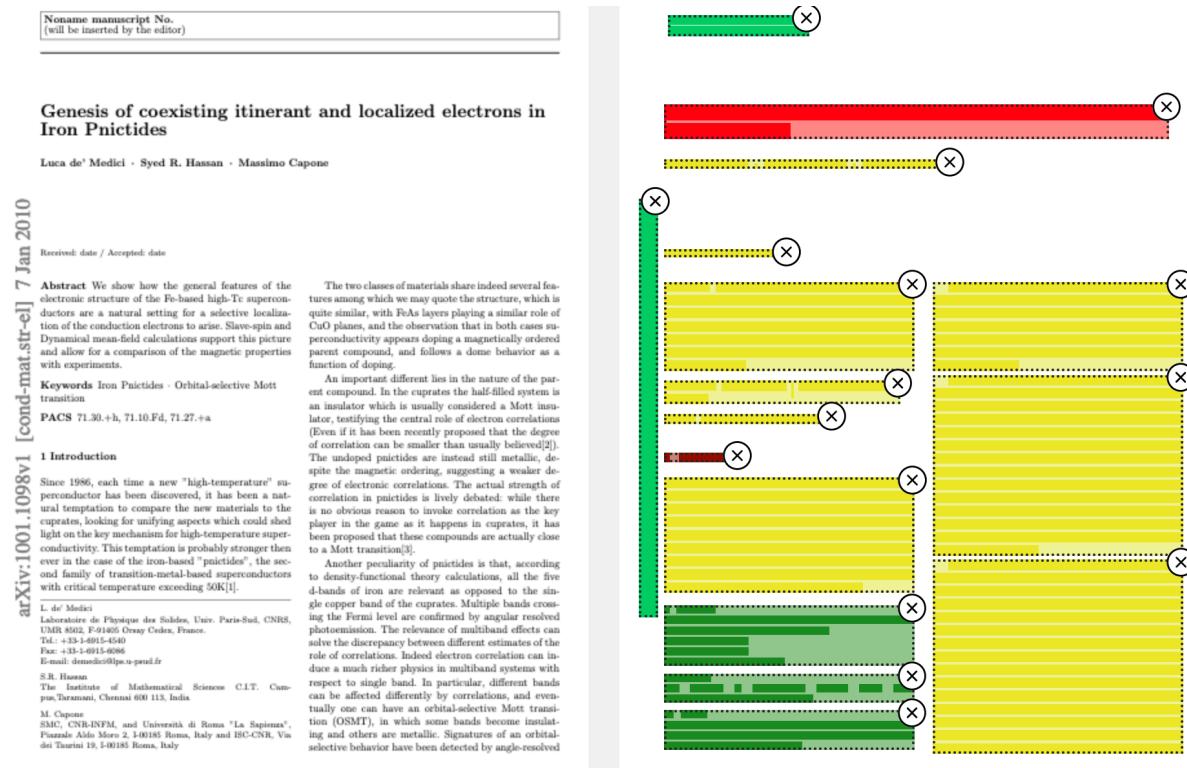
FIG. 1: Dependence of the unfolding times τ on the stretching force F for 1e2i (solid line, in red). In this Letter we describe this mechanism as a superposition of two unfolding pathways: I for small forces (dashed lower line, in blue), II for intermediate and large forces (dashed-dotted upper line, green).



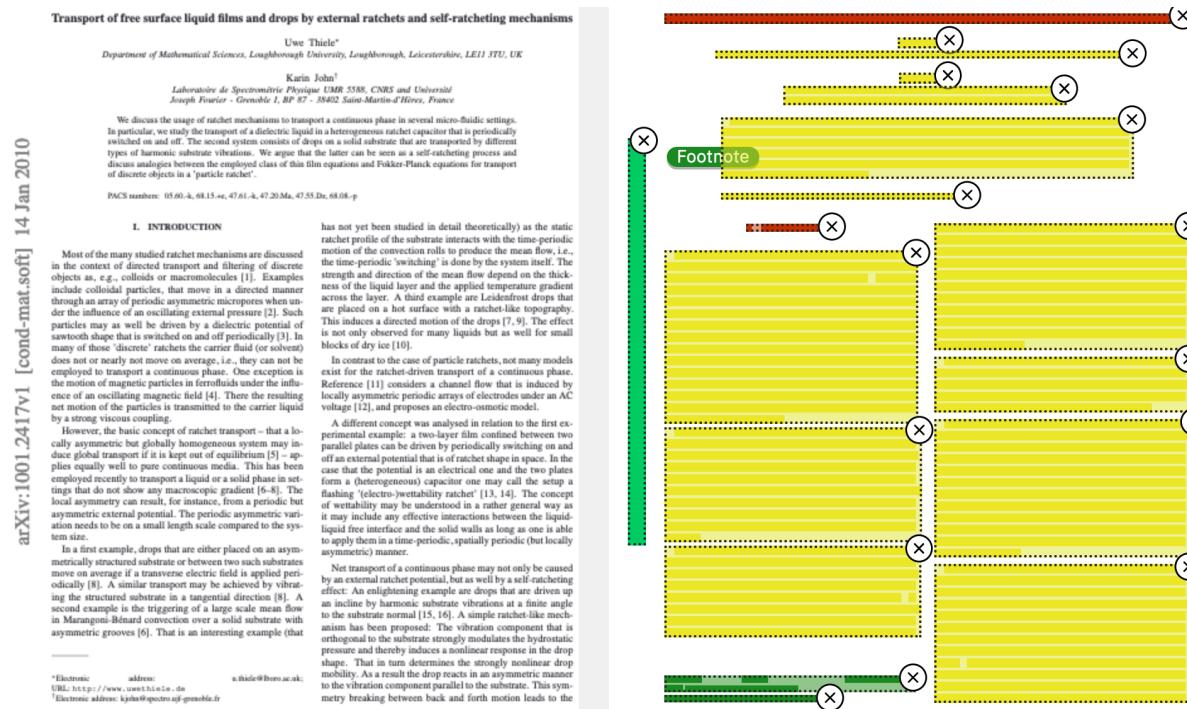
Most of our analysis is based on stretching simulations under constant force [16]. The crucial signature for this process is the overall unfolding time from the beginning of the stretching until the protein fully unfolds. Normally one expects that the transition between the folded state and the unfolded basin to be limited by overcoming the free energy barrier, which gets effectively reduced upon an application of a stretching force. The rate by which this barrier is reduced depends on the distance between the unfolded basin and the top of the barrier measured along the stretching coordinate x . This idea was first developed in the phenomenological model of Bell [18], which states that the unfolding time τ decreases exponentially with applied stretching force F as $\tau(F) = \tau_0 e^{-\frac{F}{k_B T}}$. A

⁵² Page 6609d3406660851312dd62c22f14abdb5ef38b37c1c8b2acfa5247eca0f36daf

In the third example, the addresses are in a footnote position and small font, even though they don't have the typical little footnote start symbols.⁵³



In another example, the authors are separate with their own addresses, plus footnotes about them.⁵⁴



⁵³ Page 782a1ae514a3e2b15668008fd8aa49c8018e2275068e38075d1532e2731bd2f3a

⁵⁴ Page 09e412156cf7772081d0031050906b1d35ce690c823364d5e5c7dfde5bc0d0a3

Now just the top part of a title page where abstract and keywords have Section-headers:

Concurrent consideration of cortical and cancellous bone within continuum bone remodelling

Ina Schmidt^a, Areti Papastavrou^a, and Paul Steinmann^b

^aFaculty of Mechanical Engineering, Nuremberg Tech, Kellerplatz 12, 90489 Nuremberg, Germany; ^bChair of Applied Mechanics, University of Erlangen-Nuremberg, Egerlandstraße 5, 91058 Erlangen, Germany

ABSTRACT

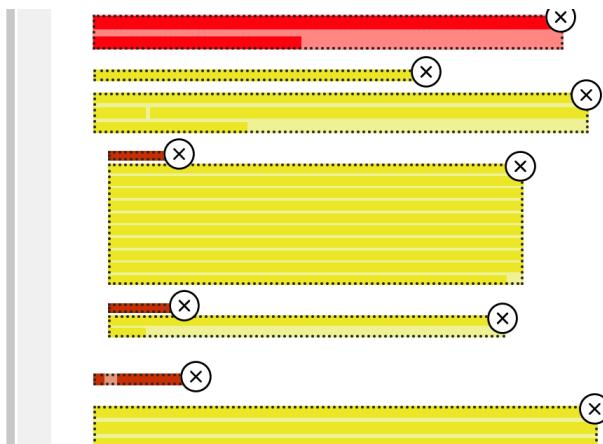
Continuum bone remodelling is an important tool for predicting the effects of mechanical stimuli on bone density evolution. While the modelling of only cancellous bone is considered in many studies based on continuum bone remodelling, this work presents an approach of modelling also cortical bone and the interaction of both bone types. The distinction between bone types is made by introducing an initial volume fraction. A simple point-wise example is used to study the behaviour of novel model options, as well as a more complex finite element, where the interaction of both bone types is demonstrated using initial density distributions. The results of the proposed model options indicate that the consideration of cortical bone remarkably changes the density evolution of cancellous bone, and should therefore not be neglected.

KEYWORDS

bone remodelling; cortical bone; cancellous bone; density change; finite element method

1. Introduction

Tubular bone is characterised by cortical bone surrounding the bone marrow in the shaft and the cancellous bone at the proximal and distal ends. Due to its special microstructure composed of beams and plates (trabeculae), cancellous bone is also



Some title pages have a date after the addresses. That should be split from the addresses too, in particular if it is in a different font, like here:⁵⁵

The Electronic Specific Heat of $\text{Ba}_{1-x}\text{K}_x\text{Fe}_2\text{As}_2$ from 2K to 380K.

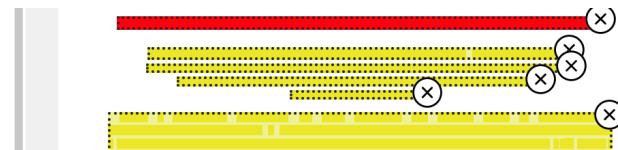
J.G. Storey¹, J.W. Loram¹, J.R. Cooper¹, Z. Bukowski² and J. Karpinski²

¹Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, U.K. and

²Laboratory for Solid State Physics, ETH Zurich, Zurich, Switzerland

(Dated: November 9, 2018)

Using a differential technique, we have measured the specific heats of polycrystalline $\text{Ba}_{1-x}\text{K}_x\text{Fe}_2\text{As}_2$ samples with $x = 0, 0.1$ and 0.3 , between 2K and 380K and in magnetic fields 0 to 13 Tesla. From this data we have determined the electronic specific heat coefficient γ ($= C_e/T$)



⁵⁵ Page fa4424c991ae8acf1e9710bea108ea33e22d45c394b4e1db5af8d884af24f860

Finally, a page where authors are written so far apart within a line that it's worth giving each author a separate cluster, plus a cluster for the address.⁵⁶

Attention Is All You Need

| | | | |
|----------------------------|-----------------------|-------------------------|------------------|
| Ashish Vaswani* | Noam Shazeer* | Niki Parmar* | Jakob Uszkoreit* |
| Google Brain | Google Brain | Google Research | Google Research |
| avaswani@google.com | noam@google.com | nikip@google.com | usz@google.com |
| <hr/> | | | |
| Llion Jones* | Aidan N. Gomez* † | Lukasz Kaiser* | |
| Google Research | University of Toronto | Google Brain | |
| llion@google.com | aidan@cs.toronto.edu | lukaszkaiser@google.com | |
| <hr/> | | | |
| Ilia Polosukhin* ‡ | | | |
| illia.polosukhin@gmail.com | | | |

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the initial work on the Transformer. Ashish worked with Ilia and Lukasz on implementing the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†]Work performed while at Google Brain.
[‡]Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

The diagram illustrates the Transformer architecture. It shows an input sequence (red dotted line) being processed by an encoder (yellow dashed boxes) and a decoder (green dashed boxes). The encoder consists of multiple layers, each containing multi-head attention (represented by circles with 'X' marks) and feed-forward layers. The decoder also consists of multiple layers, with attention heads that attend to the encoder's hidden states. The final output (red dotted line) is produced by a linear layer.

8 Free-form Document Categories: Financial Reports, Manuals, and Government Tenders

8.1 Text

8.1.1 Slightly Emphasized Text after Section-header: Text

Some section headers, in particular in company reports, are followed by an introductory statement in color or in bold, like “Merchant Energy ...” here:⁵⁷

⁵⁶ From another collection.

⁵⁷ Page e8ed6c775a7996a78f1cf70956706a4fd72bb94c30df50ed894e573994918e9b

65

Delivering gas and electricity to wholesale and retail markets

Merchant Energy manages the risks of procuring and delivering gas and electricity for AGL's wholesale and retail portfolios. It also manages AGL's compliance with mandatory renewable energy targets.

Merchant Energy's business groups are:

- > Energy Portfolio Management – which manages procurement of AGL's wholesale electricity and gas requirements, including management of the commercial aspects of operating AGL's own electricity generation assets and management of AGL's green product obligations.

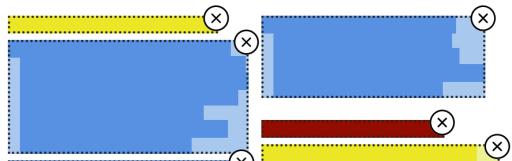
> Energy Services – which provides customers with expert advice on a range of energy related matters, including energy efficiency, and project development and management.

Energy Portfolio Management

Energy Portfolio Management manages

Yesterday tomorrow

From the old days of producing town gas, managed the gas supply, delivering supplies to local businesses.



This is nevertheless only “Text”, because it is clearly not a part of the Section-header, and it is also not heading a subsection.

In the following example, by its meaning the text “to the Fifth Edition” belongs to “Preface”, but it is so much smaller that by “what it looks like” it does not belong to the Section-header. Then it is just Text, because it does not introduce a subsection.⁵⁸

Preface

to the Fifth Edition



Do you have trouble with punctuation? Are you always using commas instead of full stops? Is your spelling weak?



8.1.2 Text Flowing around a Picture

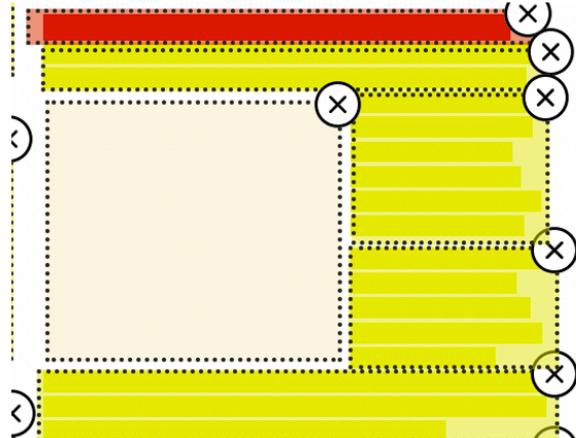
When text flows around a picture, you are constrained to build non-overlapping clusters. Break the logical paragraphs into the most reasonable rectangles as clusters (i.e., keeping lines intact). Here is an example with two paragraphs, broken into 4 clusters:

Leading the Way Back in Puerto Rico and USVI
Hurricanes severely damaged the power grids in Puerto Rico and the U.S. Virgin Islands in 2017. When Governor Cuomo led the first power restoration mission to Puerto Rico, NYPA President and CEO Gil C. Quinones was with him. NYPA has been there ever since.



In 2019, NYPA continued work with the Puerto Rican government and the island's government-owned utility—the

Puerto Rico Electric Power Authority—to rebuild the grid, provide technical assistance, consult on how to reform management and operations, and improve emergency preparedness..



⁵⁸ Page 8c5c21bb9acec37f2f540d3fc19e55e7cd88f9034ed863369e52b2b2c9e2a206

Exception: If there is only a small graphical element, not an actual figure, then instead make it part of the text, like below:

Reprints and permissions information is available at www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.



In the following example, the round picture would produce a Picture cluster that overlaps the text, so we had to omit it.⁵⁹ Then also what looked like a Caption for that picture (short text outside the normal text flow, different style) can't be a Caption, because there is no picture any more. Actually, it might be better not to submit such a page at all.



INVESTING IN TECHNOLOGY

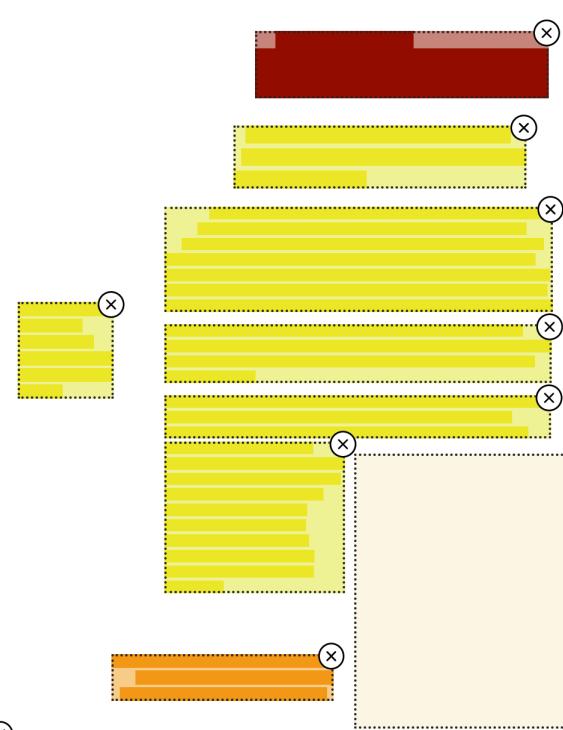
New systems and equipment help us operate more effectively and meet new market challenges.

In 2014, we continued to upgrade our technologies and production resources. We added four plants to our new Enterprise Resource Planning system, and will complete our Company-wide implementation in 2015. This system enables us to centralize many functions and respond more quickly to customer requests. It also provides insights to help us make more informed decisions on everything from resource allocation to product development. And it helps us share best practices among plants more easily.

In 2014, we also invested in a new Customer Relationship Management system, which will provide a new level of visibility into customer information. This will make account planning and forecasting easier, and help us serve customers better.

In addition, we invested in new production resources for our fastest growing markets. For example, our \$7 million investment in new molded fiber equipment in Texas will help us meet the rising demand for eco-friendly packaging solutions. We also expanded clean room capacity and added new high-speed die cutting and thermoforming equipment for our medical and biotech customers. These initiatives are all aimed at accelerating growth, enhancing efficiency, and improving our long-term profitability.

We added four plants to our Enterprise Resource Planning system, and will complete the implementation in 2015.



8.1.3 Text in Boxes

Usually, if there is a box around text (only one, not a table with many boxes), it remains just Text. This is usually a way to highlight certain text.⁶⁰

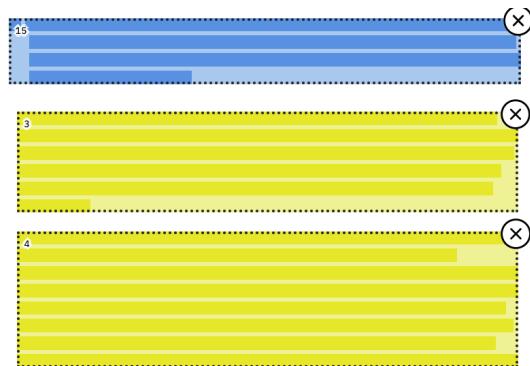
⁵⁹ Page ebf8858be842f76c6b83b62dd035204113c0246db889fde90098c8ccb40e5382

⁶⁰ Page ad5317e53fa047c316fb5848e58db01edc0d806a463d0592866f73d60b2d5ad8

6. Complete the section "Tips for avoiding installation problems" under the section entitled "Performing pre-installation tasks for the Records Enabler servers" in *IBM DB2 Content Manager Records Enabler V8.3: Installing and Configuring*, GC18-7570.

Note: CMRE asks for the Content Manager administrator user ID, such as ICMADMIN and its password, when you add a new Library Server connection to the Records Enabler System. CMRE uses the ICMADMIN ID to create the new CMRE connection ID (CMREID) in Content Manager. CMRE does not save the ICMADMIN ID and it does not use the ICMADMIN ID again after installation.

The CMRE creates the CMREID user ID in Content Manager. CMREID is a Content Manager administrator. CMRE saves this user ID in its host configuration and uses it to configure the Content Manager host. For example, CMRE uses the CMREID ID to create the necessary item types, privilege sets, and ACLs during the installation process. The CMREID is also stored in the WebSphere Data Source. This user ID is used to add and remove triggers to the records-enabled item types. The CMREID should not be imported into Records Manager. It is meant to be used under the covers by the CMRE code.



8.1.4 Text Boxes as Pictures: Don't Submit

We have an example document (about improving in English) where text in boxes serves as examples; it does not belong to the main text flow, e.g., on this page.⁶¹

112 / PART TWO: ENGLISH IN ACTION

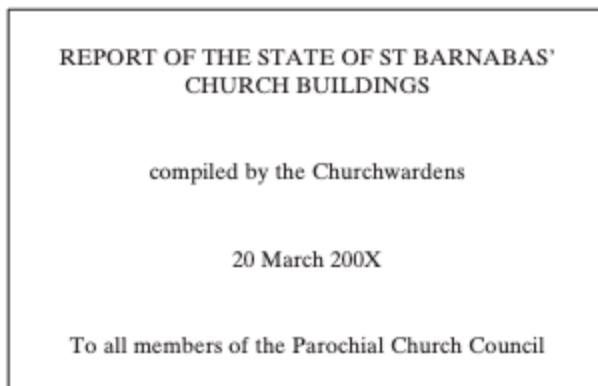


Fig. 2. Title page of report.

Providing the contents table

A contents table follows the title page. This may not be necessary if the report is short. However, if it is a long report, it is useful to list the paragraph headings and the pages on which they appear.

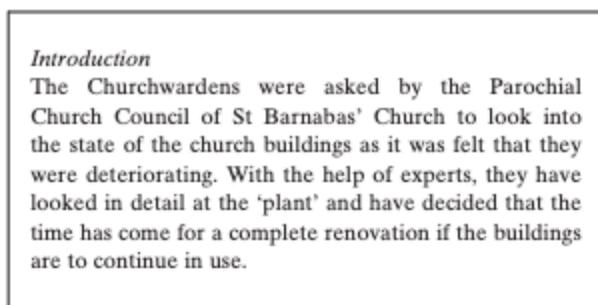


Fig. 3. Introduction to report.

⁶¹ Page 925b8c73c012936f497d6c7b35a6a8db55164c693324c34480144dc1a903a8b8

From “what it looks like” the part in the boxes is “Text”, and underneath there a Captions. However, we also have a rule that a Caption should belong to 1 Picture or Table. As we can’t fulfil all our rules, it is best to report a problem with “I am unsure how to label this page”.

The following example shows bullet items, with intermediate headings and lists, but it has a Figure-caption. Also, no black cells were recognized, so you cannot annotate the text. Such a case should be reported as an error, and not be submitted.

FIGURE 1

Executive Snapshot: Public Cloud Adoption in Global Banking

IDC's 2019 Industry Cloudshift Survey results for the global banking industry show continued strong demand for public cloud services. In fact, both public and private cloud deployments continue to attract attention and budget over noncloud workload models. IDC believes this will continue to grow as institutions continue to modernize and take advantage of the opportunity to create hybrid environments of location-agnostic architecture.

Key Takeaways

- Two-thirds of banks surveyed have deployed more than one or two applications on public cloud.
- Software as a service (SaaS) leads the way in deployments, but intent to move to platform as a service (PaaS) and infrastructure as a service (IaaS) shows an appetite to leverage other public cloud models.
- Technology spending on noncloud deployments has dropped by 70% since 2018 while spending on public cloud services rose 133% over the same period.
- Worldwide bank spending on public cloud services will be \$21.5 billion in 2019 and is expected to grow over 22% annually.

Recommended Actions

- Bank spending on public cloud services shows no signs of tapering off and is in fact growing at a 22% annual pace into 2027. Banks that have not yet done so must begin leveraging public cloud to avoid lagging behind their competitors in speed, efficiency, and agility.
- The increased use of public cloud for SaaS deployments isn't surprising, but the intention to increase the use of PaaS and IaaS signals an increased trust of public cloud services for more critical workloads. Institutions should begin to evaluate whether this strategy makes sense for their own workloads.
- For some period of time, and as investments in noncloud deployments shrink, banks will be challenged to maintain and govern a hybrid cloud/noncloud environment. This should act as motivation to re-justify legacy systems and modernize or outsource them as befits the business.

Source: IDC, 2019



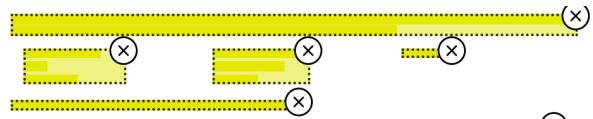
8.1.5 Not-hanging Multi-column Lists: Text

Most Redbooks contain a list of trademarks or other products mentioned, formatted like this:

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo) IBM®
IBM®
IBM Cloud® IBM Elastic Storage® Redbooks®
IBM Spectrum®
POWER®

The following terms are trademarks of other companies:



While it looks a bit like a table, we select “Text” because this is just a list. Which trademarks end up in the same row is meaningless, so there is not really a row-and-column structure.

8.1.6 Code: Text, except if very Tabular

Some computer-related manuals, e.g., in the collection “Redbooks”, contain a lot of code. Usually, this is annotated as “Text”, see Section 3.1.9.

Only when computer output is in table format with proper column headers and rows of related information, we use Table, e.g., here:⁶²

⁶² Page f4da878e1ce12eb36ffc2f396be12db3416ba34d4083f65960addb4dcfdc88f5

Display, then an average of the single raw values

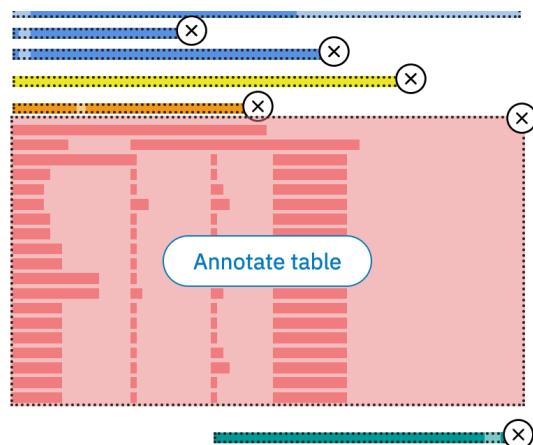
- IOPS: Total IOPS of all nodes
- CPU percentage: Average CPU percentage of all nodes

Example A-5 shows the resulting output of the `lssystemstats` command.

Example A-5 The `lssystemstats` command output

```
IBM_2145:ITSO-SV1:superuser>lssystemstats
stat_name      stat_current stat_peak stat_peak_time
compression_cpu_pc 0          0          181014221539
cpu_pc         2          2          181014221539
fc_nb          0          14         181014221504
fc_io          566        690        181014221504
sas_nb          0          0          181014221539
sas_io          0          0          181014221539
iscsi_nb        0          0          181014221539
iscsi_l0        0          0          181014221539
write_cache_pc 0          0          181014221539
total_cache_pc 21         21         181014221539
vdisk_nb        0          0          181014221539
vdisk_l0        0          0          181014221539
vdisk_ms        0          0          181014221539
mdisk_mb        0          13         181014221504
mdisk_l0        5          100        181014221504
mdisk_ms        0          2          181014221504
drive_nb        0          0          181014221539
```

Appendix A. Performance data and statistics gathering 773



This is really a table, with column headers “stat_name” etc., and then 3 numbers belonging to each of those stat_names. We don’t worry that the font is a typical computer font and the caption says “Example” instead of “Table”.

8.1.7 Diagonal Text over a Page

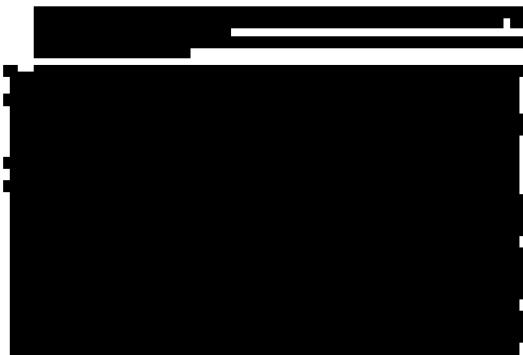
In particular in the Bavarian tender collection, there are documents with diagonal text over them, saying things like “Draft” or “Cannot be submitted electronically”. Initially such a page looks like this:⁶³

- In der alle Mitglieder aufgeführt sind und der für die Durchführung des Vertrags bevollmächtigte Vertreter bezeichnet ist.
- dass der bevollmächtigte Vertreter die Mitglieder gegenüber dem Auftraggeber rechtsverbindlich verfügt und dass alle Mitglieder als Gesamtschulden haften.
Auf Verlangen muss der Bevollmächtigte von allen Mitgliedern unterzeichnet bzw. fortgeschritten oder qualifiziert signierte / mit Siegel versehene Erklärung abzugeben.

5.2 Sollten nicht öffentlich ausgeschrieben wird, werden Angebote von Bietergemeinschaften, die sich erst nach der Aufforderung zur Angebotsabgabe aus aufgeforderten Unternehmen gebildet haben, nicht zugelassen.

6 Nachunternehmen
Baublatt der Bieter Teile der Leistung von Nachunternehmen ausführen zu lassen, muss er in seinem Angebot Art und Umfang der durch Nachunternehmen auszuführenden Leistungen angeben und auf Verlangen die vorgesehenen Nachunternehmen benennen.

7 Eignung
7.1 Öffentliche Ausschreibung
Präqualifizierte Unternehmen führen den Nachweis der Eignung durch den Eintrag in die Liste des Vereins für die Präqualifikation von Bauunternehmen e.V. (Präqualifikationsverzeichnis) und ggf. ergänzt durch geforderte aufragspezifische Einzelnachweise. Bei Einsatz von Nachunternehmen ist auf gesondertes Verlangen nachzuweisen, dass diese präqualifiziert sind oder die Voraussetzung für die Präqualifikation erfüllen, ggf. ergänzt durch geforderte aufragspezifische Einzelnachweise.
Nicht präqualifizierte Unternehmen haben ein vorläufiges Nachweis der Eignung mit dem Angebot die ausgewählte „Eigenerkündigung zur Eignung“ vorzulegen, ggf. ergänzt durch geforderte aufragspezifische Einzelnachweise. Bei Einsatz von Nachunternehmen sind auf gesondertes Verlangen die Eigenerkündigungen auch für diese abzugeben ggf. ergänzt durch geforderte aufragspezifische Einzelnachweise. Bei Einsatz von Nachunternehmen präqualifiziert, reicht die Angabe der Nummer, unter der diese in der Liste des Vereins für die Präqualifikation von Bauunternehmen e.V. (Präqualifikationsverzeichnis) geführt werden ggf. ergänzt durch geforderte aufragspezifische Einzelnachweise.
Gelangt das Angebot in die engere Wahl, sind die Eigenerkündigungen (auch die der benannten Nachunternehmen) auf gesondertes Verlangen durch Vorlage der in der „Eigenerkündigung zur Eignung“ genannten Bezeichnungen zuständigen Stellen zu bestätigen. Bezeichnungen, die nicht in deutscher Sprache abgefasst sind, ist eine Übersetzung in die deutsche Sprache beizufügen



Currently this cannot be properly labeled. You have to report a problem as follows:

Other (please describe problem below)

Description

Diagonal text over page

8.1.8 Wrong Text that Can Still be Labeled

In general, if one sees wrong text boxes, it is good to report the error. Often, one can also not label the page, but sometimes one can:

⁶³ Page f481e35c94c07a2bb825079fcec2d61bbb59371af36dbfff1135207250293ec6

- If the error is in a Picture or Table, because there you can draw the cluster independently of the text boxes.
- If a middle part in a cluster is missing, so that you can still get the cluster shape you want. Here is an example:⁶⁴



You have to be rather careful in this example not to include the green and yellow star. The correct box looks like this:



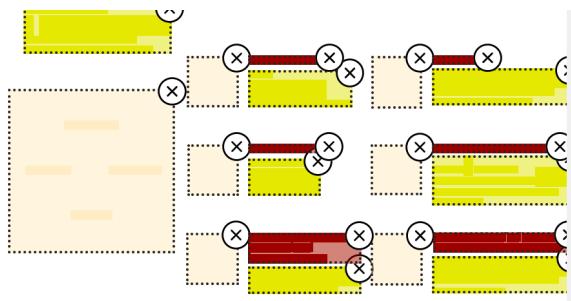
8.2 Picture

8.2.1 Multiple Small Pictures

On pages with many small pictures and text, one can wonder if the overall group is a picture or table, or each part is a separate cluster. Scientific articles usually have clear captions to resolve this, but other documents often don't.

We try to go by a similar rule: If there is text between the pictures, or each picture has its own accompanying text on the side, then we make several pictures. Whereas if they are close together without separate texts, we make them one picture.

First we show an example with separate Picture and Text clusters:



In the next example, we make only one picture, because there is no separate text, and the two photographs look as if they belong together.⁶⁵ One could imagine a joint caption "Wrong and correct riding clothes" for them.

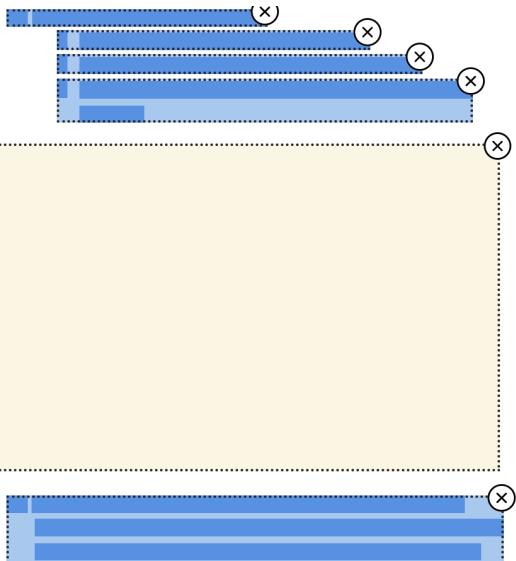
⁶⁴ Page 06c327a755df058276b089462aad1664e5626c98a819b06475d3c383899ac9fb

⁶⁵ Page 3b1baa5fe926d467a2de593125d2958eba4d805eb9974ad0b9a36b7323ae5b24

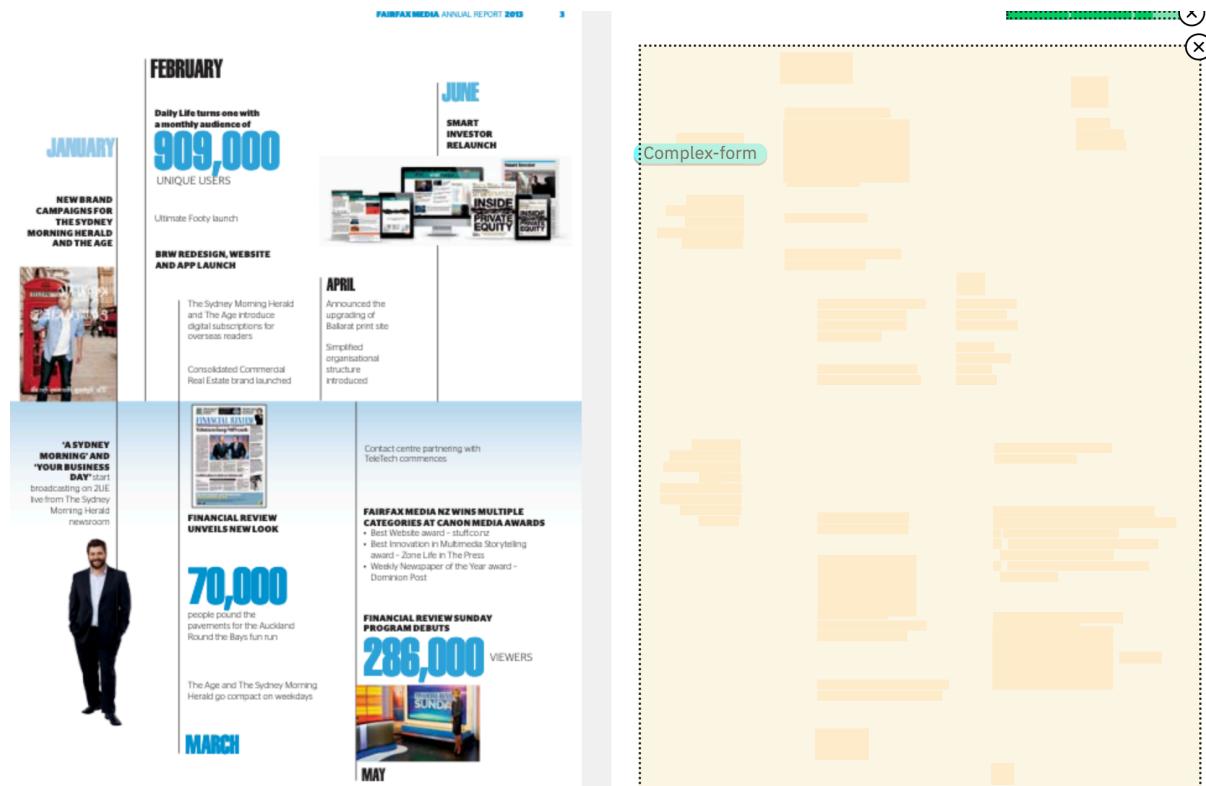
- 50. Other clothing.** You should wear
- boots or shoes with hard soles and heels
 - light-coloured or fluorescent clothing in daylight
 - reflective clothing if you have to ride at night or in poor visibility.



- 51. At night.** It is safer not to ride on the road at night or in poor visibility, but if you do, make sure you wear reflective clothing and your horse has reflective bands above the fetlock joints. A light



For more chaotic interleaving, one large picture is better, e.g., here (where there are even additional lines in the intermediate spaces):⁶⁶



8.2.2 Background Images behind Text

Sometimes you cannot create both a “Text” cluster and a “Picture” cluster without overlaps. In such a case:

- If the text is the key information, and there is just some highlighting, or a general picture in weak colors in the background, then use appropriate clusters like “Text” or “Section-header” and ignore the picture.

⁶⁶ Page 51337b1626f3c40af8ac85c6959f7f5aea7ba4d7bf1388a8b14ed2c6e8909f78

- Do not annotate *parts* of a background picture – if the entire picture overlaps with important text, then also don't annotate any picture parts outside the overlap.

Below is an example of such a page:



More examples are under “Title Pages”.

If you find text or other elements that you cannot annotate (e.g., no cell is present), please use the “Report error” feature in the lower right corner. If the problem is only minor, then submit the page anyway, while if it concerns the title itself or other large text, do not submit the page.

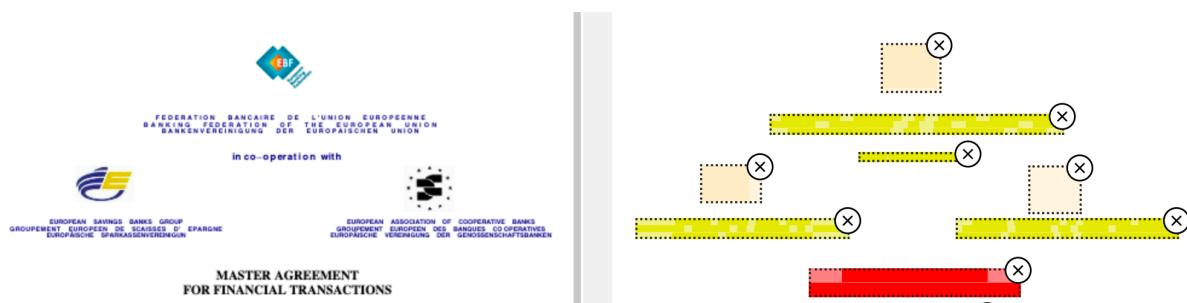
8.2.3 Logos

Logos are treated as what they primarily look like. This can be “Text” or “Picture”, or “Page-header” or “Page-footer” if they are in those positions.

The first example is best labeled “Text”, because there are mainly letters. (Or “Page-footer” if at the bottom of a normal page.)



In the next example, the logos are pictorial. This is the top of a title page, and the text and logos describe the three authoring organizations.



Many logos spread around on a page without order (e.g., “all our partners”) better become one big Picture:

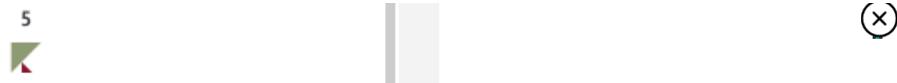


If there is no text cell for the logo, and this is only in a Page-header (in contrast to a title page where we really want to know the company name), then making it a Picture is OK, like here:



This is better than just omitting the logo, because there clearly is something that our image-based AI will see.

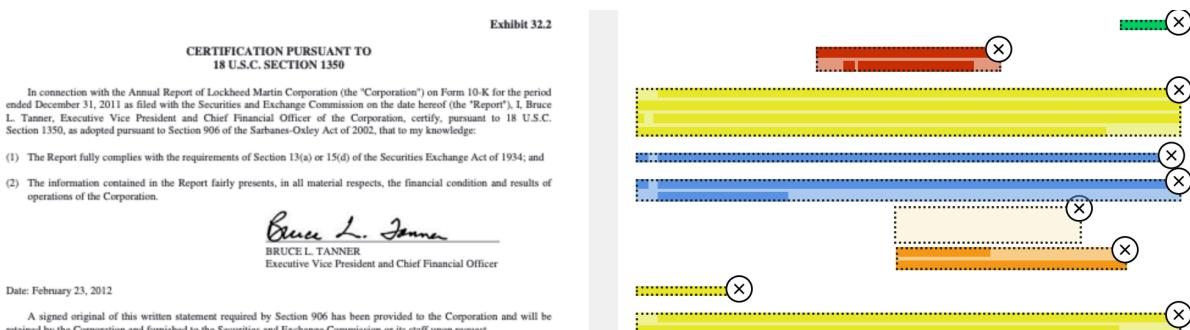
Tiny logos in a Page-header or Page-footer position that don't fit inside the text of that header are better ignored, like here under a Page-footer:⁶⁷



As a rule, one might say: Only make Picture clusters if the picture is taller than 2 rows of normal text (not inline).

8.2.4 Signatures

Company reports and contracts often contain signatures. We treat these as Pictures, and the printed name under them as Caption for them, wherever possible. Here an example:



In the second example, the signature reaches into the text, but we still decided that the separation was possible:

⁶⁷ Page 5c82776fc48f15a55f76534a899b3cd64d5c320d3ad2f8604162306b592f267c

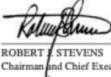
Exhibit 32.1

**CERTIFICATION PURSUANT TO
18 U.S.C. SECTION 1350**

In connection with the Annual Report of Lockheed Martin Corporation (the "Corporation") on Form 10-K for the period ended December 31, 2011 as filed with the Securities and Exchange Commission on the date hereof (the "Report"), I, Robert J. Stevens, Chairman and Chief Executive Officer of the Corporation, certify, pursuant to 18 U.S.C. Section 1350, as adopted pursuant to Section 906 of the Sarbanes-Oxley Act of 2002, that to my knowledge:

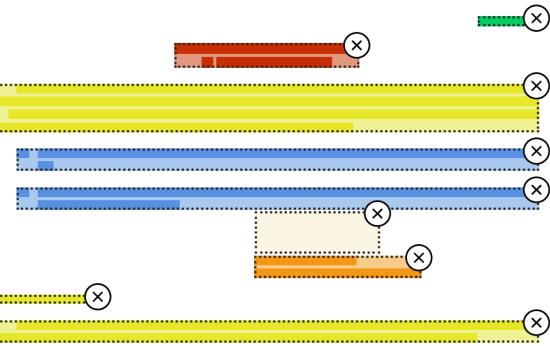
(1) The Report fully complies with the requirements of Section 13(a) or 15(d) of the Securities Exchange Act of 1934; and

(2) The information contained in the Report fairly presents, in all material respects, the financial condition and results of operations of the Corporation.


ROBERT J. STEVENS
Chairman and Chief Executive Officer

Date: February 23, 2012

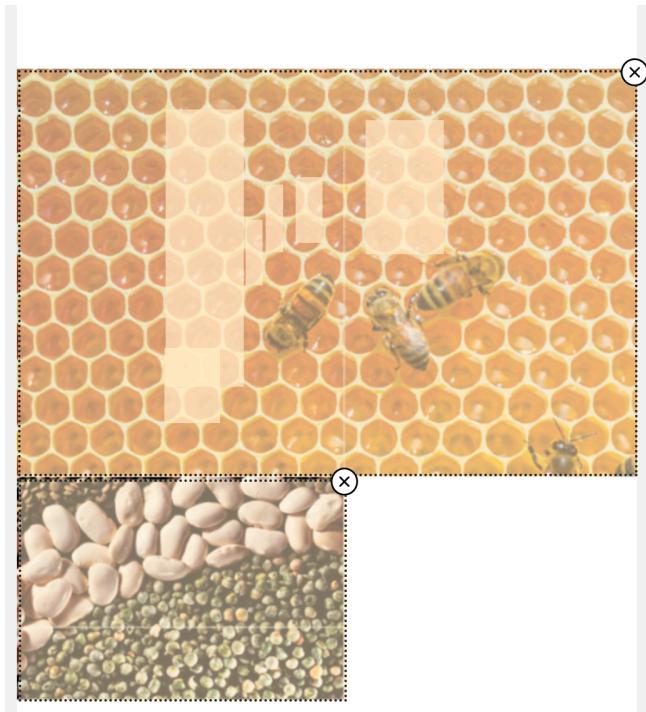
A signed original of this written statement required by Section 906 has been provided to the Corporation and will be retained by the Corporation and furnished to the Securities and Exchange Commission or its staff upon request.



8.2.5 “Irrelevant Pictures”: Still Pictures

Something that looks like a picture should be annotated as a Picture, unless it conflicts with foreground text. In particular this may apply for pictures on title pages of enterprise documents.

The following example is the second page of a document and without text, and might not be very important, but as you cannot know that, you should annotate it as two Picture clusters:



8.2.6 Table in Picture

If there is a table in a picture, one has to decide if the picture aspect or the table aspect is more important. In the following example, we decide for Picture for these reasons:⁶⁸

- There is interesting text in the picture beyond that of the table, and it may be needed to understand the table.
- That text would be hard to annotate piece-by-piece.
- There is a single caption for the entire picture.

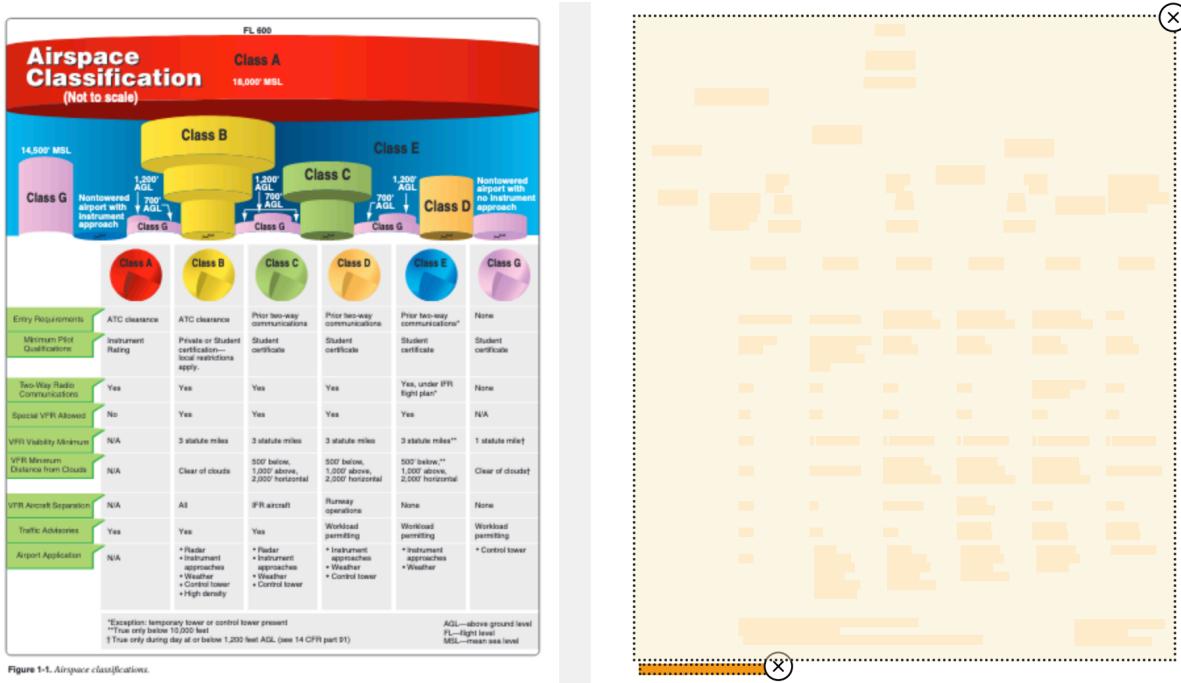


Figure 1-1. Airspace classifications.

8.3 Caption (or Not?)

In scientific articles, captions are usually easy to recognize by their label “Figure 3: ...” or “Table 5: ...”, but in the document categories we are considering here it is more difficult.

8.3.1 Indicators of Captions

Key indicators for captions without label are:

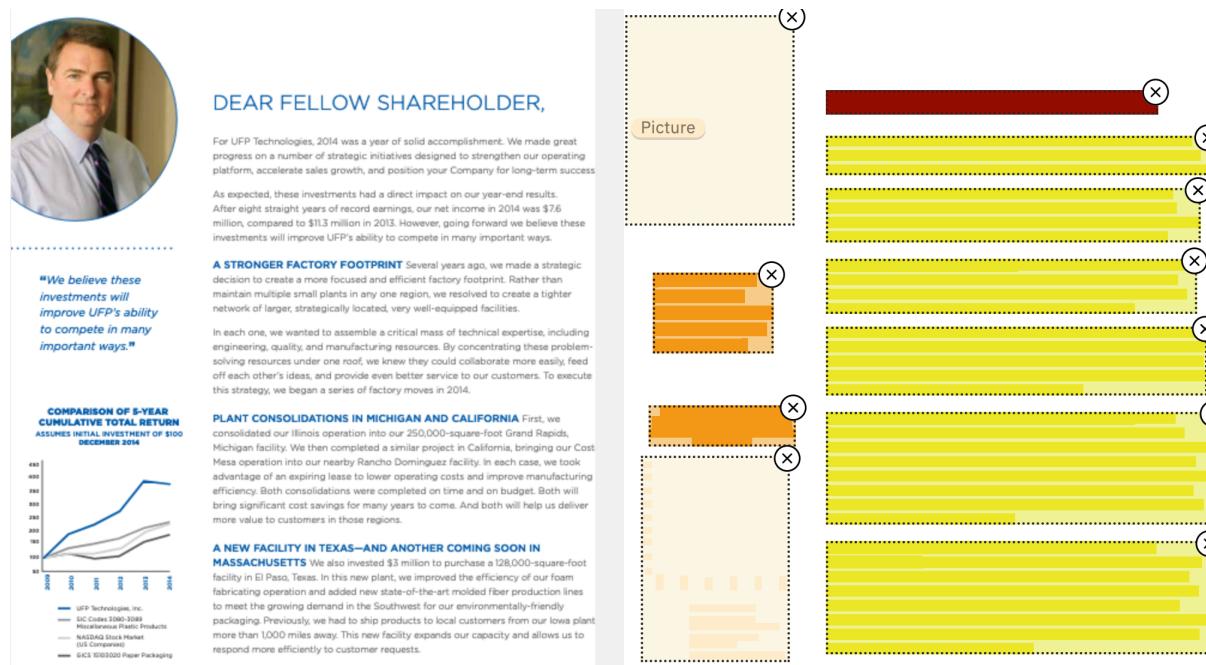
- A kind of name for a Picture or Table (as you might know it from newspapers too).
- Not in the normal text flow.
- Different style than normal text.
- Only 1 per Picture or Table.

⁶⁸ Page 109d5f22725c5776e14226eded14f3dc9b98a31064cf0618b37db16aa681c026

8.3.2 Special-looking Paragraph near Picture or Table: Caption

The following example contains two Pictures.⁶⁹

- The legend of the graphic belongs inside the “Picture” cluster, as it explains the differently colored lines.
- The text above the graphic is quite clearly a kind of header for that picture, thus “Caption”.
- The text under the upper picture is a bit more doubtful. It would clearly be a Caption if it said “Our CEO xyz”. But while it has contents that could also be a sidebar (and thus “Text”), here it seems to belong to the picture and has similar coloring as the clear caption below. Thus we decided on “Caption”.



8.3.3 Normal-looking Paragraph near Picture or Table: Text

If the text looks like any other text, then even if it clearly introduces a picture (e.g., with a colon), we keep it as “Text”, e.g., here:⁷⁰

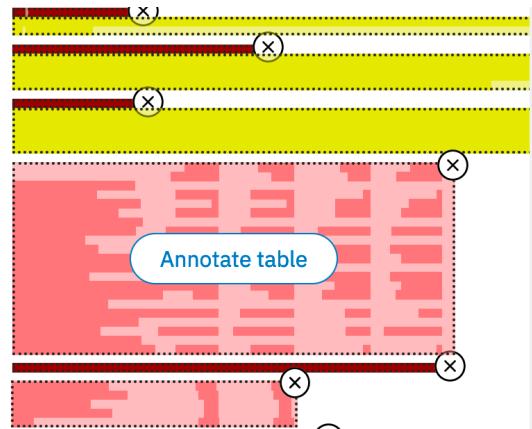


⁶⁹ Page e3f7d52143efc9bc60eef8080235457fc7f95abc06fa1b13915468ef262785c1

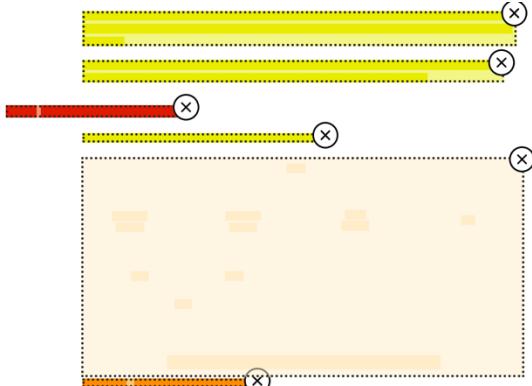
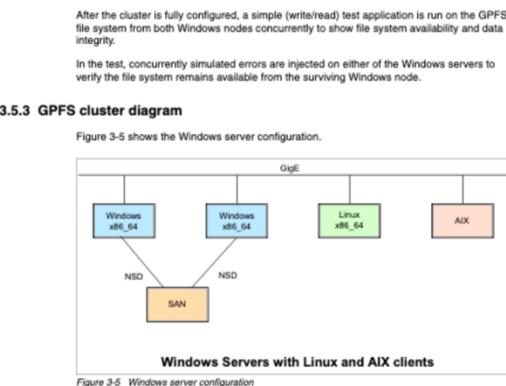
⁷⁰ Page 0a99340a323a266edc4baf8e28d195cfdb4903b0d0a855b924d52f80ccb5358f

In the next example, “Text” is clear for the paragraph above the table, because this paragraph looks like the 2 Text paragraphs before, is rather long, does not start “Table x” and doesn’t even mention the table a lot.

| 1.1) Shareholding Board members | | | |
|---|---------------|---------------|-------------|
| During 2008, Board member Jo Lunder held the position of President in Ferd Industrial Holding, which had a holding of 2,900,000 SEK 31.12.2008. | | | |
| Extract from principles for remuneration of Group Management | | | |
| Salary should include both a fixed and a variable part. The variable salary may amount to a maximum of 50% of the fixed salary. Bonus should be moderate and only account for a limited part of the remuneration package. There should be no special pension plans for CEO members of the Board. In 2008, TOMRA was involved by a long term incentive plan for Group Management members (see the principles for remuneration of Group Management are found under the Corporate Governance section of the annual report). | | | |
| Long-Term Incentive Plans (LTIP) | | | |
| At the end of 2008, TOMRA established a long-term, cash-based incentive plan, where managers receive bonuses based upon annual growth and local unit's profit and performance. The bonus for each year is placed in an interest-bearing account in a virtual bonus bank, from which any bonus can be withdrawn at any time. The bonus bank is invested in a diversified portfolio of equities and bonds. The bonus bank is invested in a group of comparable companies. Vesting will only be achieved if TOMRA beats the index, and has a positive share price over vesting during any given year for each of the participants. Vesting will be their annual salary, and 50% of the earnings after taxes must be invested in TOMRA. | | | |
| Balance | Paid out | Earned | Balance |
| 31.12.2008 | 2009 | 2009 | 31.12.2009 |
| Arneid Skarholt (President/CEO) until 10 August 2009) | 3,822,923 | 3,822,923 | 0 |
| Steffen Gundersen (President/CFO) | 0 | 0 | 144,000 |
| Leifur Gundersen (CFO) | 3,822,923 | 2,018,742 | 433,699 |
| Harald Henningsen (SVP Technology) | 3,822,923 | 2,018,742 | 433,699 |
| Heidi Kjeldsen (VP, Tomra Nordic) | SEK 933,685 | SEK 933,685 | SEK 324,000 |
| Hans Bevers (R&D), Tomra Systems GmbH* | EUR 470,529 | EUR 248,392 | EUR 52,271 |
| Bjørn H. Hagen (R&D), Tillech) | 3,474,643 | 1,670,463 | 433,699 |
| Ton Klumper (VP, Tomra Western and Eastern Europe) | EUR 250,214 | EUR 92,199 | EUR 50,840 |
| Markus Kjeldsen (Business Development) | 1,913,414 | 1,067,072 | USD 15,556 |
| Michael Læsø (President, Tomra US East) | USD 0 | USD 0 | USD 16,965 |
| Gregory Knoll (President, TOMRA Solutions) | USD 676,545 | USD 676,545 | USD 0 |
| Håkan Ergegen (VP, Tomra Nordic until 31 October 2009) | SEK 4,440,633 | SEK 2,321,078 | SEK 0 |
| Tomra Systems (GP Business Development until 1 November 2008) | 3,822,923 | 3,822,923 | 0 |
| The collective compensation for key management personnel is as follows (26 managers in 2009 and 21 in 2008): | | | |
| Amounts in NOK million | | | |
| Short-term employee benefits | 57.5 | 44.6 | |
| Severance payments | 9.7 | 0.3 | |
| Post-employment benefits | 3.9 | 3.3 | |
| Total | 70.7 | 47.4 | |



In the following example, one might be confused what is the correct caption, as both texts above and below the figure start “Figure 3-5”. However the text above has the same font as the longer paragraphs before and is a full sentence, while the text below the figure is in italic and no full sentence. Thus only the line below is the Caption.



Another example shows the same principle for a table.

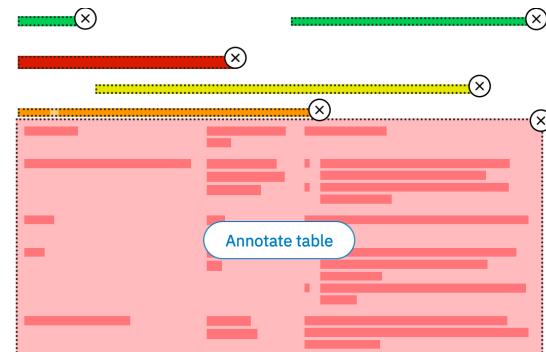
5608paper.fm Draft Document for Review December 17, 2020 10:54 am

Minimum software version levels

Table 1 shows the minimum software version levels for CDP Private Cloud Base.

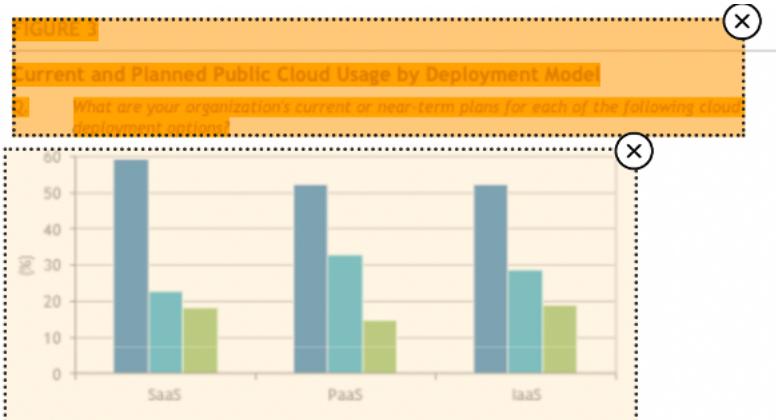
Table 1 Minimum software version levels for CDP Private Cloud Base

| Component | Minimum release level | More information |
|--------------------------------------|--|--|
| Cloudera supported operating systems | 64-bit Red Hat Enterprise Linux (RHEL) 7.7 | <ul style="list-style-type: none"> Supported OS version for both CDP Private Cloud Base and IBM Spectrum Scale. CDP Private Cloud Base currently does not support RHEL 8. |
| Python | 2.7 | CDP Private Cloud Base currently does not support Python 3. |
| Java | Java 8/OpenJDK 1.8 | <ul style="list-style-type: none"> Supported Java version for both CDP Private Cloud and IBM Spectrum Scale HDFS Transparency. HDFS Transparency currently does not support Java 11. |
| CDP Private Cloud Base | CM 7.2.3 CDH 7.1.4 | IBM Knowledge Center Big Data and Analytics support CDP Private Cloud Base Support Matrix for more information. |



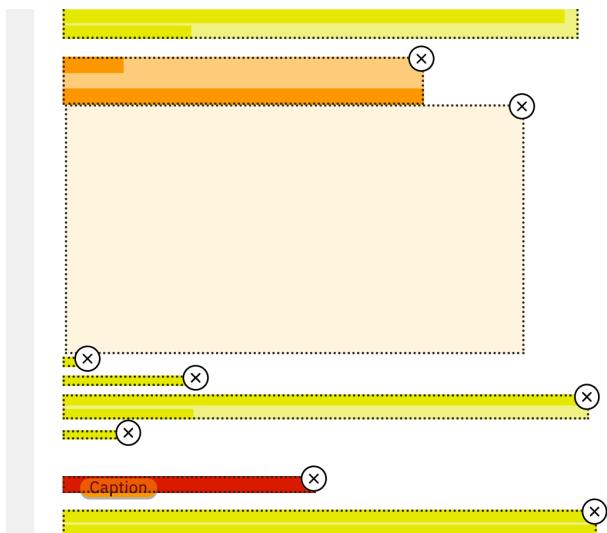
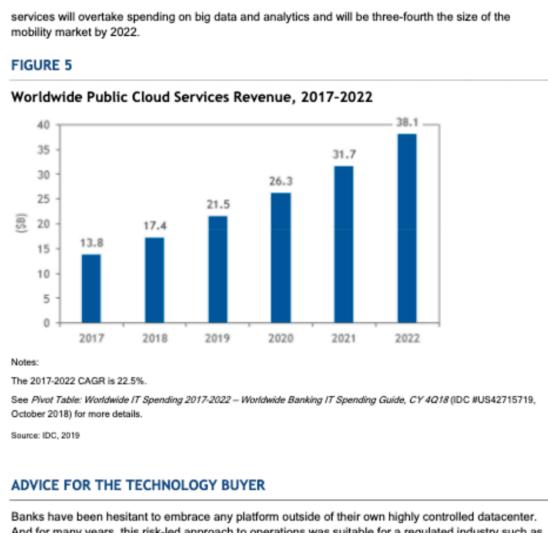
8.3.4 Multi-paragraph Caption

Some captions are longer and contain hard paragraph breaks or indents, which would indicate a new paragraph in normal text. However, please still make this only 1 cluster, because we prefer to have exactly 1 Caption for each picture or table. In the following example, the text “FIGURE 3” in the first line starts a Caption, and then the rest of the text up to the picture must also belong to this Caption.

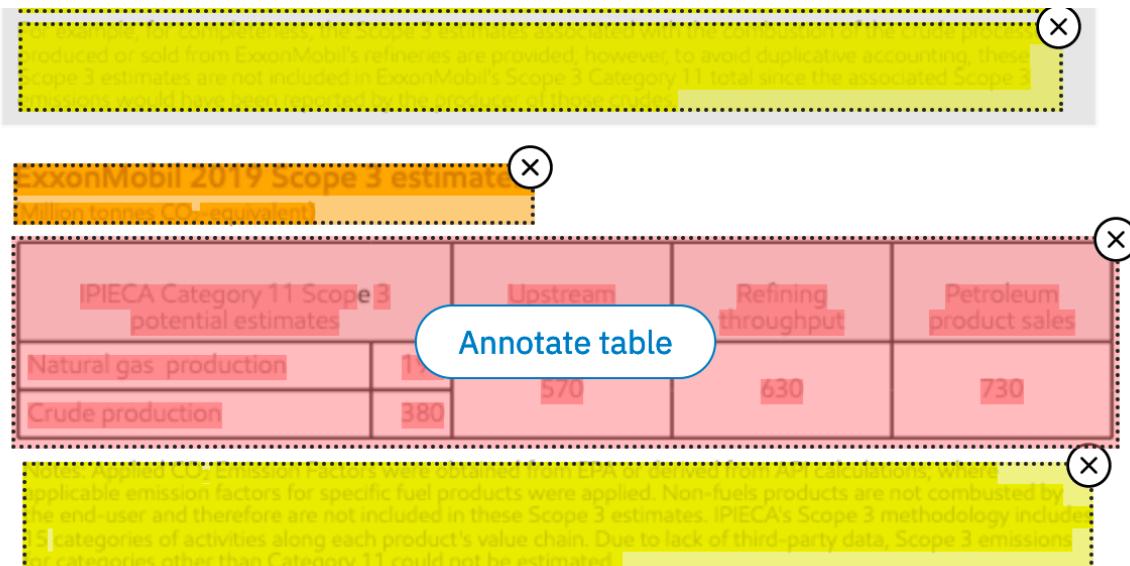


8.3.5 Additional Context for Pictures or Tables: Not a Second Caption

Some pictures have a caption above, and some notes, or source statement, or footnote-like text below. These cannot be a second caption. One has to decide whether to consider it part of the Picture or normal Text. In the following example, the notes below the figure look like Text, and are not key to understanding the bars in the Picture, so they become Text clusters.

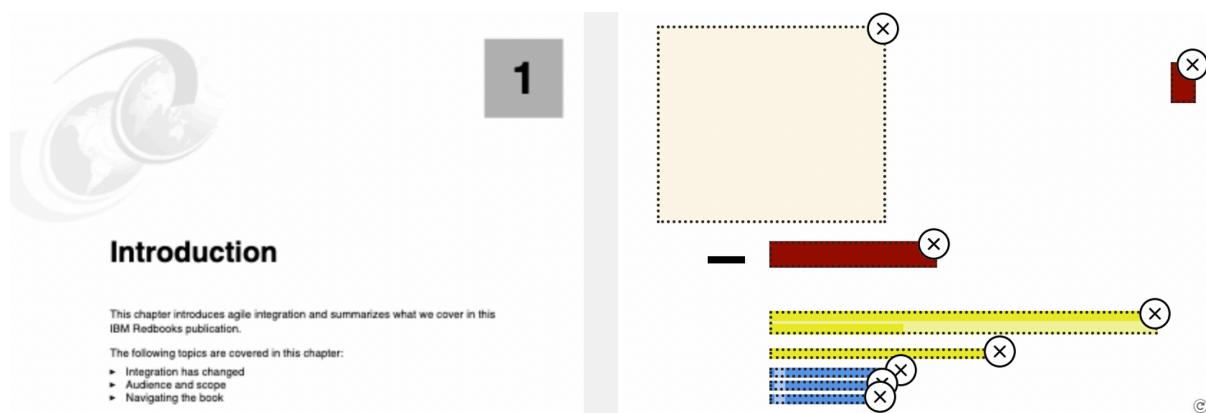


The same rules apply for tables. In the following example, the Caption has no start word like “Table 6”, but clearly is the caption of this table, i.e., a separate line without a full sentence, giving a name to this table. The next line fits well into a multi-line table caption, while the text below the table has to become “Text”, as it is not inside the table grid.



8.4 Section-header

Some long documents have major chapters with a separate introductory page. In particular this is true in the “Redbooks” collection within the document category “Manuals”. Their chapter introduction pages should be labeled as in the following example:⁷¹



Here “1” is the chapter number and “Introduction” the corresponding text. Thus we label them both as “Section-header”, but as they are in different styles and far apart, we make two separate clusters.

Similar principles can be applied to other individual manuals or financial reports.

8.5 Formulas

Some manuals are technical, but not deeply mathematical. You will not find many formulas, but there can be some. For instance, an equality sign may be used to introduce a term or an equation, e.g., like this:

$$\text{Availability} = \text{Uptime} / (\text{Uptime} + \text{Downtime})$$

⁷¹ Page 385982667ac8903f6f0c3b81299a83cf806315570c0cabd615d9433a5e6aeb3

This should be a Formula cluster, as it is on its own line and contains several mathematical symbols.

In contrast, if it would appear only as:

Av = Availability,

you should mark it as “Text”, since it doesn’t look sufficiently formula-like. The exact borderline is up to your judgement.

8.6 Table

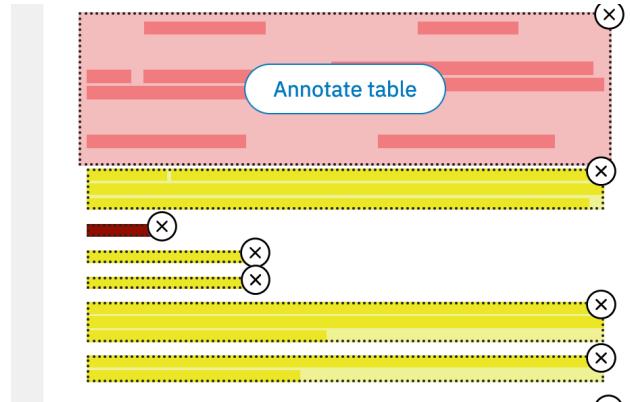
8.6.1 One-column Paragraphs in Table Grid

In some complex tables, not all rows have the same number of columns. For instance, multi-column headers are quite common in the first row (before a row with more detailed headers per column), and belong to the table.

In the following two examples, there is text without column structure at the end, but still inside the gridlines. So one needs to decide whether one believes the gridlines or the text format on where the table ends.

In the first example, the grid comprises not only the real table with 2 columns, but several paragraphs of normal text.⁷² Here we stop the table after the 2-column part.

| Subject of the tender | Maximum budget |
|--|---|
| LOT 2 - Provision of infrastructure for Operational Cooperation and exercises | A maximum budget of €600.000,00 (six hundred thousand euro) over the maximum possible period of 3 years |
| Last date for dispatch of offers | 13 th March 2020 until 18:00 CET |
| PLEASE NOTE: This tender procedure is limited to tenderers which are legally incorporated in a member state of the European Union/EEA, or which have an incorporated subsidiary in one of the EU/EEA member states. (The Agreement on Government Procurement (GPA) does not apply to EU Regulatory Agencies.) | |
| IMPORTANT! | |
| Provisions relating to BREXIT | |
| For British candidates or tenderers: | |
| Please be aware that after the UK's withdrawal from the EU, the rules of access to EU procurement procedures of economic operators established in third countries will apply to candidates or tenderers from the UK depending on the outcome of the negotiations. | |
| In case such access is not provided by legal provisions in force candidates or tenderers from the UK could be rejected from the procurement procedure. | |



⁷² Page 510f18f8318147c28922afcf11e47d8b01ffccb89e3097c1e8d25a67c7d80c02

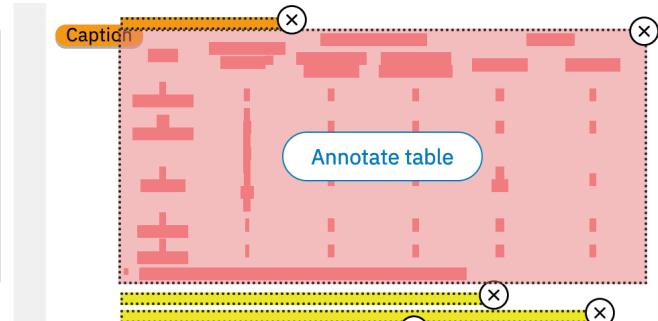
In the next example, a single last line in the grid is across all columns, and thus also somewhat like Text. However, it is kind of a footnote to the table (linked by the “a”), and the add-on to the table is much shorter than in the previous example. So here we still made it part of the table, similar as multi-column headers exist.⁷³

| Child seats with the ISOFIX system | | | | | | |
|------------------------------------|--|-----------------------------|-------------------------------|-------------|-------------|--|
| Group | Size category of child seat ^a | Front passenger's seat | | Rear seats | | |
| | | With activated front airbag | With deactivated front airbag | Outer seats | Centre seat | |
| 0 (up to 10 kg) | E | X | X | IL | X | |
| 0+ (up to 13 kg) | E D C | X | X | IL | X | |
| 1 (9-18 kg) | D C B B1 A | X | X | IL IUF | X | |
| 2 (15-25 kg) | - | X | X | IL | X | |
| 3 (22-36 kg) | - | X | X | IL | X | |

^a The size category of the child seat is indicated on the child seat plate.

IL The seat is suitable for the ISOFIX child seats with "Semi-Universal" approval.

IUF The seat is suitable for forward-facing child seats and is permitted for use in this weight category.



If there is more such text than in the second example but less than in the first example, you have to use your judgement.

8.6.2 Glossaries: Usually Table

A glossary is an explanation of terminology.

- If it comes within gridlines, it is clearly a table.
- Below is an example glossary without gridlines. One might consider it a list (hanging paragraphs). But typically there are some terms on the left, like “Binary classification” here, which wrap around in their invisible table cell. Thus the reading order implies that this should be labelled as “Table”, too, here with 2 columns and 4 rows.

Bag of words Text representation as an unordered set of its words, but with frequencies.

Balanced In classification, the training set is called balanced if there are approximately the same number of examples for each class.

A very unbalanced training set leads the classifier to prefer the majority class too much. E.g., if there are 900 examples of class A, and 100 of class B, the simple algorithm that always outputs “A” already has 90% accuracy.

A countermeasure is subsampling.

BERT The method is usually called by this abbreviation.

Binary classification Classification into only 2 classes. In CDRA, primarily the classification “risk or not”.

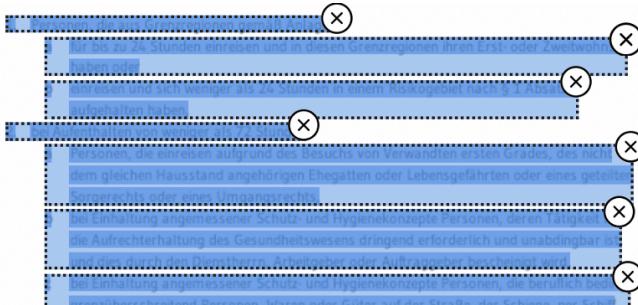
8.7 List-item

Actually, most of the following cases are *not* difficult decisions, but follow from the key rule in Section 3.9. Anyway, we provide examples of typical lists and non-lists.

⁷³ Page 700a63afc816281d29a5474119076d589153a026a2d705b893ebd7c8f6e4e4c0

8.7.1 Multi-level Lists

List items of different depth are not distinguished:



8.7.2 Each “Hanging” Line Starts a List-item

There is not always vertical space between list items, but always every new hanging line (here because of a new bullet) starts a new List-item, as in the following example:

1.5 Resulting Goals of *SEMPER*

1.5.1 Security Requirements

Generalizing from concrete scenarios and concrete threats, and from user interviews in *SEMPER* (see Chapters 7 and 8) and outside surveys, one can see the following main security requirements for electronic commerce:

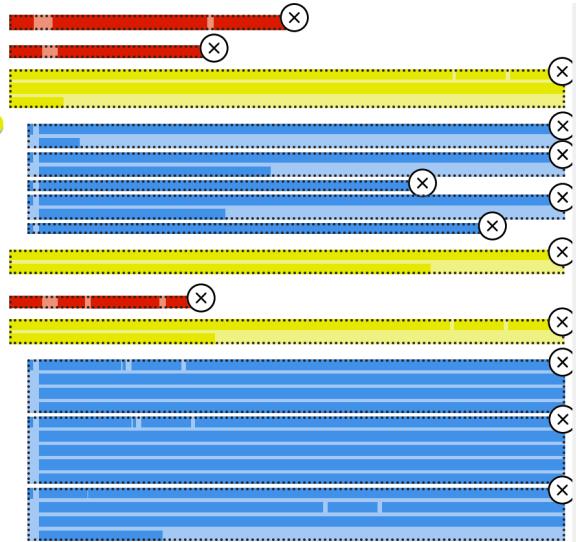
- Fairness: Ensure that situations like payment without delivery, or delivery without payment, do not happen.
- Authenticity: Ensure that partners cannot impersonate others or, where anonymous, ensure that they cannot act without having the proper rights.
- Availability of service: Ensure that all contracts and promises are fulfilled.
- Privacy: Ensure that partners do not collect or use data for unintended purposes. Ensure that outsiders do not get unnecessary information.
- Prevent misuse of goods, like infringement of copyright and illegal resale of information.

All these requirements are not only made on the internal technical system, but relative to how the real users interact with the system, from the user interface design to the legal environment.

1.5.2 The *SEMPER* Focus

As a result of the view on the current situation and future needs as described so far, *SEMPER* has set the following goals for its specific approach:

- *Entire processes*: *SEMPER* should support all the standard steps, linked into complete business processes as described in Section 1.4, such that the security requirements are fulfilled for the entire processes. For extreme cases, disputes must be supported, i.e., the tool must help to find the necessary evidence and must also help arbitrators to evaluate it. Privacy must also be supported for entire processes.
- *Multiple scenarios*: *SEMPER* should be usable in a variety of person-to-business, business-to-business, and person-to-person scenarios. In particular for large businesses, it should be suitable for integration into back-end systems, i.e., for use without human intervention. In other cases, only export and import of data with other programs is necessary. The benefits of using one framework for all these scenarios is decreased development cost and increased confidence in its security due to more intensive scrutiny.
- *Openness*: For many of the services (e.g., payments, signatures, encryption) several protocols and products already coexist and will keep coexisting. The *SEMPER* architecture should be able to integrate them easily. It must also support business partners in selecting which protocol and product to use in a given situation.



8.7.3 Bullets or Numbers Not Needed

If paragraphs have the hanging shape, then even if they have no list-identifiers (bullets, numbers etc.), they are List-items.

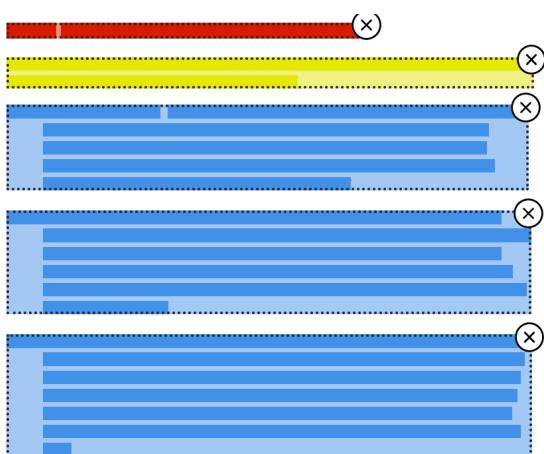
4.1.1.1 Commerce transaction service rationale

The rationale for the commerce transaction service in the *SEMPER* architecture is based on the following high level requirements:

Business commonalities Although the business applications will differ in layout and functionality, there will be common features. The commerce layer should provide these common business features. These common features will be offered as a set of abstract business services built on top of the lower layer services such as connection and transport services.

Business context One of the key common features of business applications is the requirement for maintaining a business context. The commerce layer transaction service should provide means for maintaining such a business context. This context generally comprises of an association between the parties involved in the business, the transactions exchanged between these parties, and private data stored by each party.

Secure service access point Both trusted and untrusted business applications will be built on top of the commerce transaction service. For untrusted applications the commerce transaction service represents the perimeter between the trusted and untrusted functionality of *SEMPER*, and a secure service access point must be provided. It must be impossible for untrusted business applications to use the commerce transaction service to perform actions that are not sanctioned by the user.



8.7.4 Normal Paragraphs Inside a List Item: Text

Some documents contain lists with items that are long and have internal paragraphs. As those internal paragraphs are not hanging, they become “Text”, e.g., here the additional paragraphs in item 6

5. The GK is wrapped with the DK and stored in a structure that is referred as the Encrypted group key (EGK).
 6. The EGK is persistently stored in the key repository (KR) of the DS8000. Both the EEDK and the EGK are stored in multiple places in the DS8000 for reliability.
- This dual control (from the DS8000 and IBM Security Key Lifecycle Manager) improves security. The DS8000 does not maintain a persistent copy of the DK on disk in the clear, and cannot encrypt or decrypt data without access to IBM Security Key Lifecycle Manager.
- The DK is *erased* by the DS8000 at power off, such that each time it is powered on, the DS8000 must communicate with IBM Security Key Lifecycle Manager to obtain the DK again.



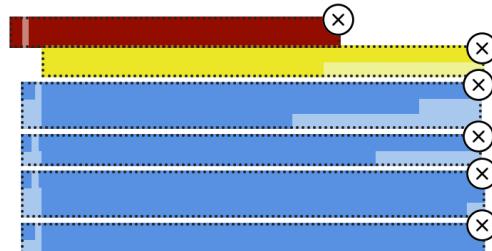
8.8 Page-header

8.8.1 Very Large Page-header can become Section-header

In the following example, hopefully only in this one document, the header of each section is repeated in the same large font on each column in that section.⁷⁴ You can only see from the strange start of the next paragraph (“to the release ...”) that this can’t really be a section start. By “what it looks like”, we make this a Section-header.

I. ATP CIRCUIT REGULATIONS

- to the release to ATP of their Anti-Doping results obtained by the ITF at ATP events, including missed tests and/or filing failures.
- 3) To comply with and be bound by all of the provisions of the 2021 Tennis Anti-Corruption Program Rules, the full text of which can be found at: <https://www.tennisintegrityunit.com/education>.
 - 4) To review and agree to the terms and conditions contained in the Notice of Privacy Practices set forth in “Exhibit P - Privacy Notice (“HIPPA”).
 - 5) Each calendar year all players shall, as a condition of entering or participating in any event organized or sanctioned by ATP, deliver to ATP a signed Consent and Agreement in the form set out in ‘Exhibit O - Consent and Agreement Form’.
 - 6) For entry into an ATP Tour or ATP Challenger Tour tournament, all players must be an ATP Player Member (“Member”) or an ATP Registered Player. Wild cards



8.9 Title Pages

Title pages in company reports, manuals etc. come in great variety, so we can only provide general rules and a few examples.

Wherever possible, one or more “Title” clusters should be identified with the title of the document, and possibly other information that is key to identify the document, or is also very large. One has to balance “what it looks like” with the meaning of the text elements. Thus if a company name is much larger than the real title of a document, we make them both “Title”. With financial reports, the company is also really important, because otherwise the documents from many companies would have the same title “Annual Report” or “Annual Report 2019”.

Only use “Section-header” on a title page if real text starts on that page and is headed by this header. Otherwise decide between “Title” and “Text”. This is because a Section-header should typically be above something else, which it introduces.

⁷⁴ Page 4ed449ade591d365e986d46091ee707d2f5fa6b08374c5da8087da667fd3a148

8.9.1 Title Pages with Mostly Text

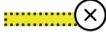
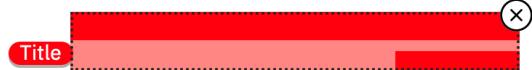
If there is mostly normal text, recognized as text boxes, similar rules hold as for scientific articles: The overall title is “Title”, while authors, addresses, dates etc. are “Text”. What might need decisions is whether a smaller subtitle still belongs to the title. If it is closely linked to the title, relatively large, and seems important as part of the title, then making it part of the title seems useful, like here:⁷⁵

The Python Language Reference
Release 3.9.5

Guido van Rossum
and the Python development team

June 01, 2021

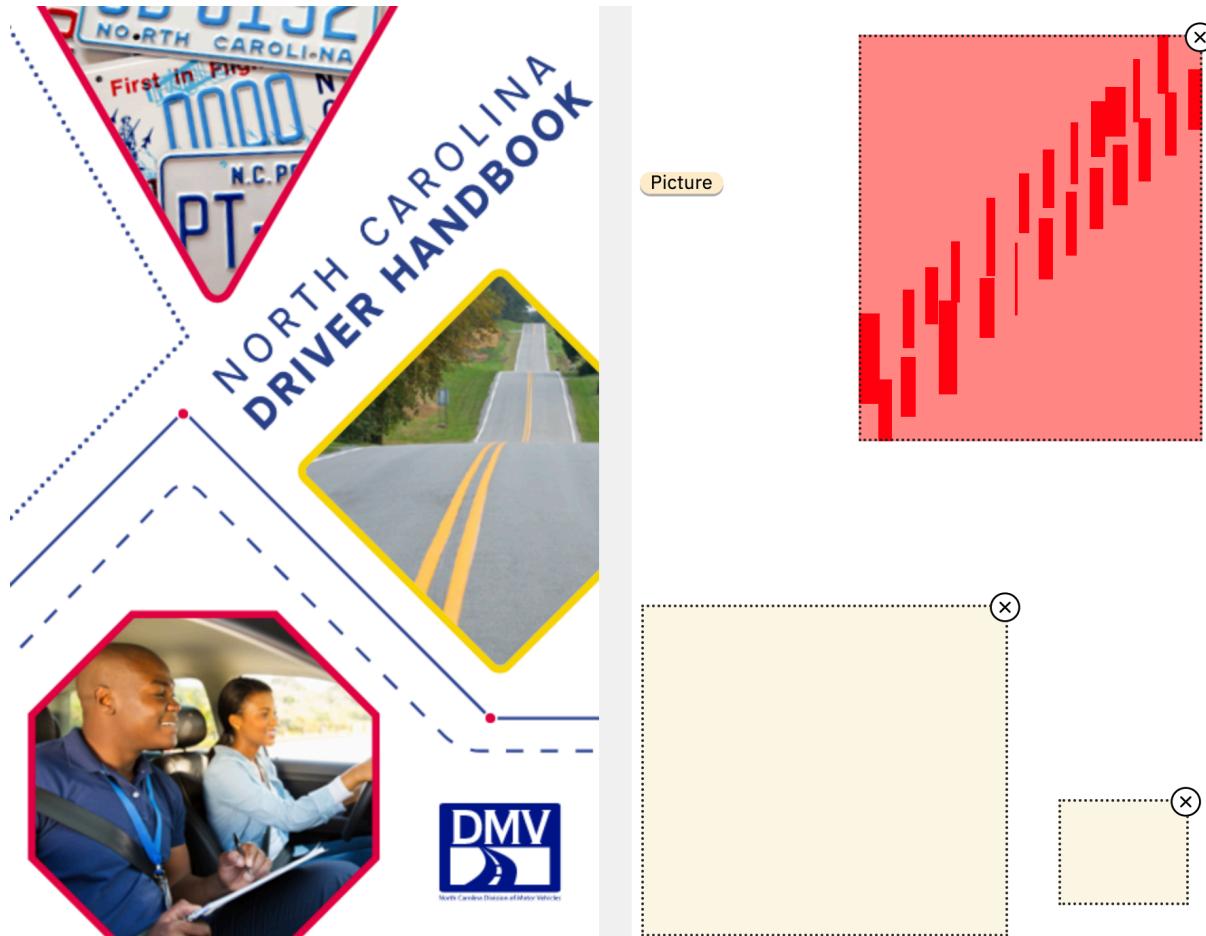
Python Software Foundation
Email: docs@python.org



⁷⁵ Page 4b1acce9c1b3114de9948591a490657f9d09becc5383d23cdfaf38333b263500

8.9.2 Text in Unusual Orientation

In the following page⁷⁶, one can select the text, and it seems to actually contain the letters. However, from a geometric point of view, a whole-page picture would seem more plausible.



⁷⁶ Page 1b93e3948e7f429a8a165d06016d72ac362400d1c2cd1aae94b47f80c5d65757

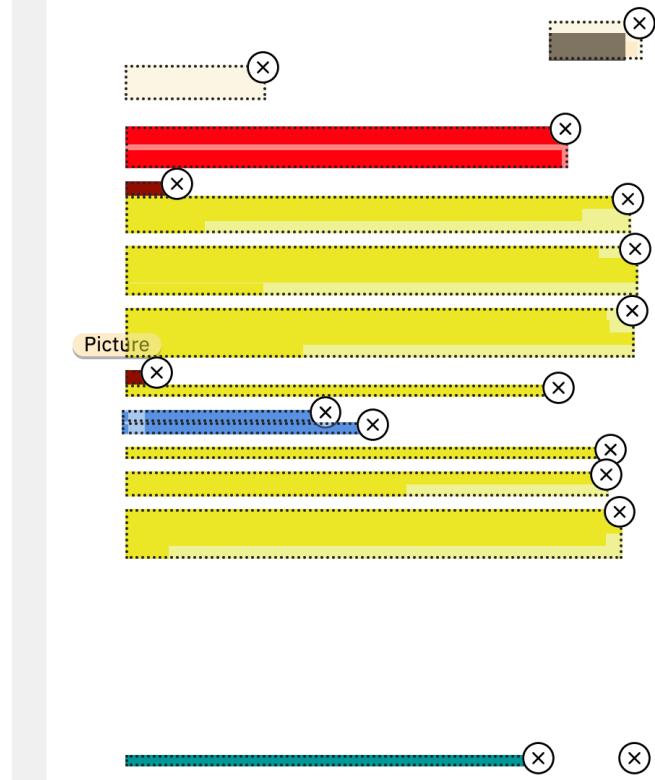
8.9.3 Manuals

8.9.3.1 Redbook Title Pages

Within the document category “Manuals”, we have a special collection “Redbooks”. Redbooks all have rather similar formats. Hence we can make precise rules how to label their title pages.

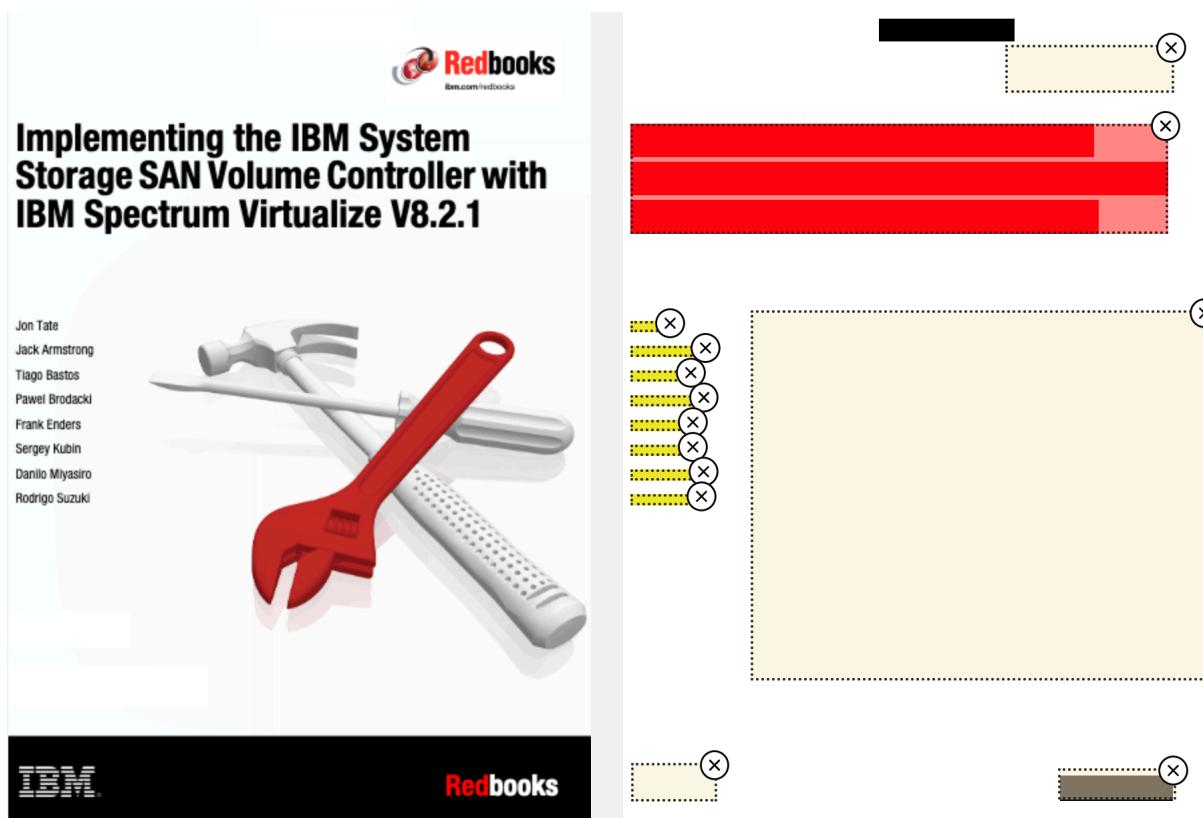
First we show a title page of a very short Redbook (sometimes also called Redpaper).⁷⁷

The image shows the title page of a Redbook. At the top is the IBM logo and the Redbooks logo. The title is "Assessing Hardware Failures on IBM PureData System for Analytics N2001, N2002, and N3001". Below the title is a section titled "Overview" with a detailed description of the document's purpose. There is a "Notes" section listing specific models that do not apply, and a note about architectural differences between N1001 and N2001/N3001 models. The page is dated "1" at the bottom center.

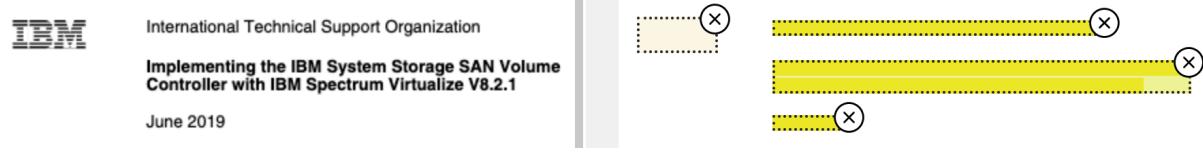


⁷⁷ Page bf2e78472a59d72a6dbd9177e2096b0e0698becf8d24a49d4049b8385cd20243

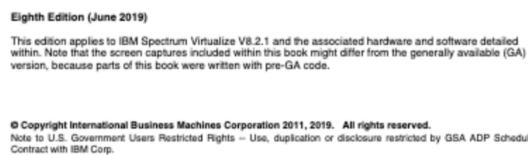
The next example belongs to a real book:⁷⁸



These large redbooks often have additional standardized pages at the front: An empty page 2, then a page with this top part:

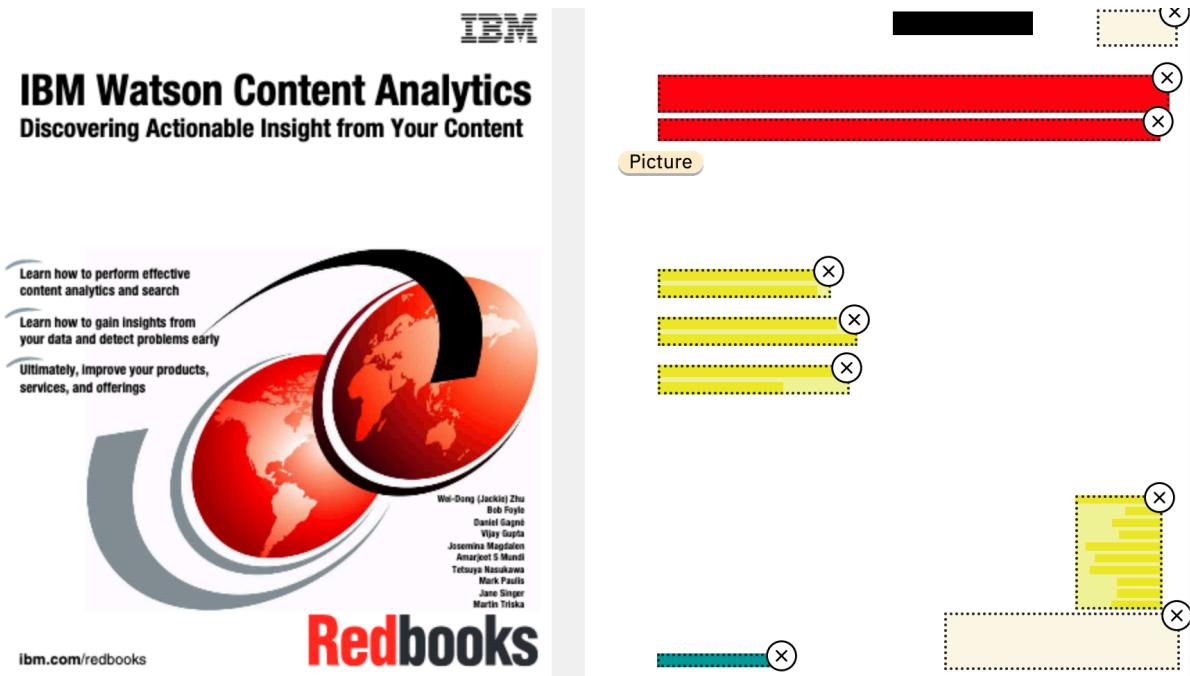


We use “Text” also for the bold middle part, because it is a repetition of the book title, not heading a certain section of the book. Similarly, we use Text at the bottom of page 4:



⁷⁸ Page 4b26bd9c094a0dd14f3a291588b2669557977eefdfcff4da23a2e93a528d1e2

Our third example has a different background picture, and this one cannot be annotated without overlapping the text, so we have to leave it out:⁷⁹



Another difference here is that the authors are so close together that we made them 1 Text cluster only.

8.9.4 Government Tenders

8.9.4.1 Title Pages of EU Tenders

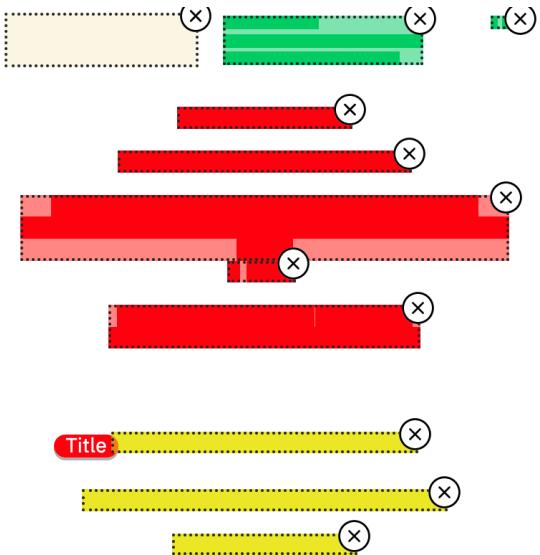
Tenders are requests for services, from small service like a cleaning service for a building to huge services like building an airport. It is also useful to know that one tender often contains many documents. The first documents may have large title pages, while appendices look more like chapters. However, as they are separate documents, we also give them a Title.

European tenders come in various formats, but there are similarities within each agency. We show examples from several agencies.

⁷⁹ Page 7987fabe63db6ed75290a16c4149bb5c9caa0bf4c2117ef7bcd2cc0e1ef71ac4

8.9.4.1.1 ECHA

In the first example, all the text on the title page is in the same format. The real title is only the 3 lines starting “Title:”. We decided to give everything in the same grey box the label “Title”.⁸⁰



8.9.4.1.2 EIB

A relatively clear page: Everything is large and seems key as the title.⁸¹

⁸⁰ Page 84e8928a907a1d831359f0dbc5a849828e66b81dc1283ddc18faf57fd1ac024d

⁸¹ Page 32153454a310f21211262bf0914ba9bd409cd2d82c412d088a5d76049600a93b



Terms of Reference

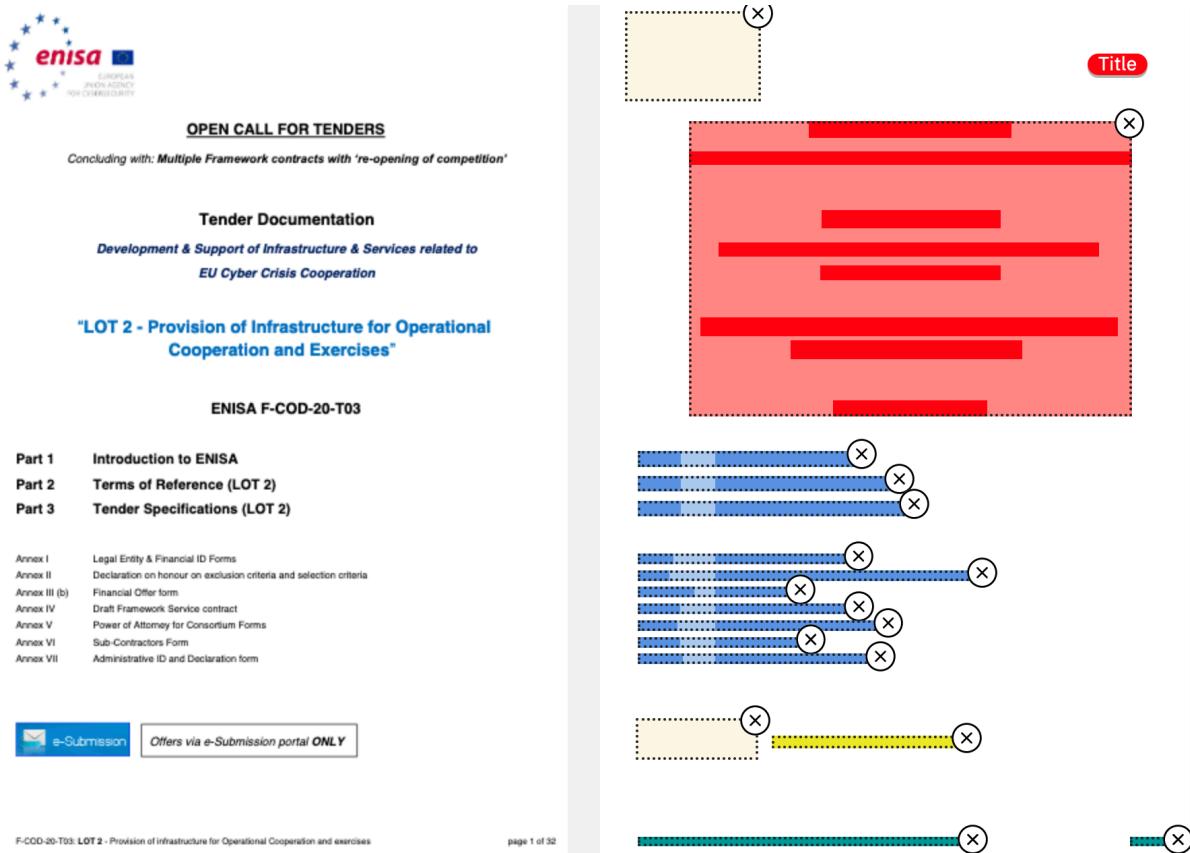
Ref: CFT-1626

**Banking and Financial Skills
Training Services**



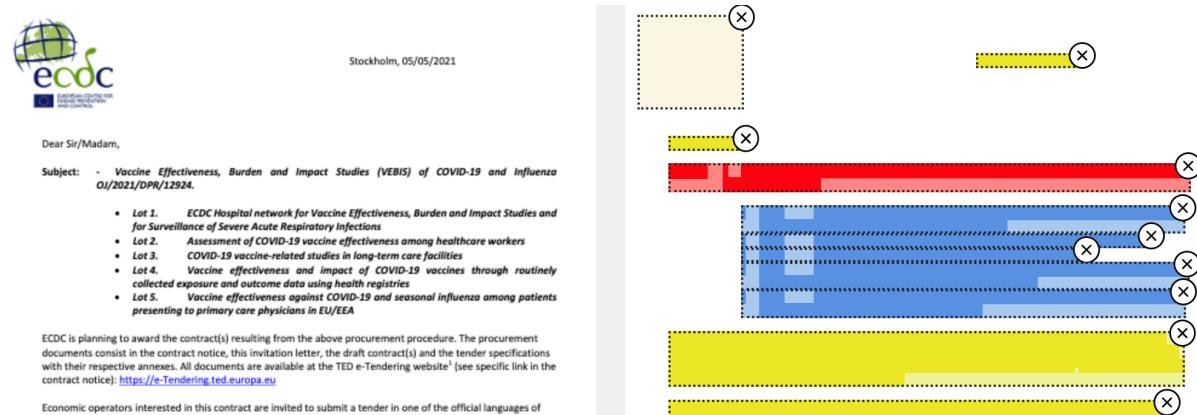
8.9.4.1.3 Enisa

Also the following title page contains a lot of large text. The real title of this document is the light-blue “LOT2- ...”, while the name of the overall tender it belongs to is the relatively small “Development and Support of ...”. The number “ENISA F-COD-...” is also characteristic. Hence we decided to make this entire area the title.⁸²



8.9.4.1.4 Title Page of Letter-style Documents

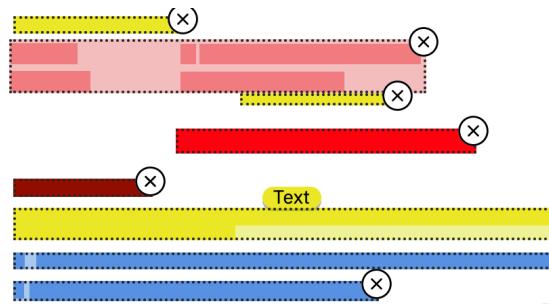
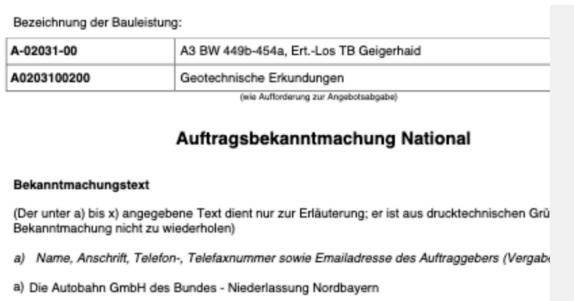
Some documents in tender packages are formulated as letters, as in the following example.⁸³ The subject line fits well as a Title; we did not count the long bold list afterwards as title. We considered the date “Text”, while on other pages something in that position might be a Page-header.



8.9.4.2 Title Pages of Bavaria Tenders

Like EU tenders, Bavaria tenders typically consist of multiple documents. One has to find reasonable solutions, including one or a few “Title” clusters if at all possible.

Example:⁸⁴



8.9.5 Financial Report Title Pages

In the category “Financial Reports”, we have one collection “sec_filings” with more or less fixed format, and two collections “ann_reports...” with free-form, more marketing-oriented reports.

8.9.5.1 SEC Filings

Title pages of SEC filings are quite full of small information pieces.⁸⁵ The top lines are always the same; the key information for this particular filing is the company name “ALICO, INC”. In this case, it is also largest, so it was easy to select it as part of the title. On some other title pages it is smaller, but should still be made “Title”.

⁸³ Page 3920ddef794557a9e0de59d50cd18359d0e5b4eecdadcb5abf5a1d642f59d434

⁸⁴ Page 8b2871cf1d5a73838fc219b4a66c509e9f43e201026e1cb56fe129a85b77c7d8

⁸⁵ Page 32d600b701fe8e5285fde1670e9313b14eefb287daa0c50236378b308d486450

**UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549**

FORM 10-K

Annual Report Pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934
For the Fiscal Year Ended September 30, 2019
or
 Transition Report Pursuant to Section 13 or 15(d) of the Securities Exchange Act of 1934
For the transition period
from _____ to _____
Commission File Number: 0-261

ALICO, INC.
(Exact name of registrant as specified in its charter)

| | |
|--|------------|
| Florida | 59-0906081 |
| (State or other jurisdiction of incorporation or organization) | |
| 10070 Daniels Interstate Court Suite 100 Fort Myers FL | 33913 |
| (Address of principal executive offices) | |
| (239) 226-2000 (Registrant's telephone number, including area code) | |

Securities registered pursuant to Section 12(b) of the Act:

| Title of each class | Trading Symbol(s) | Name of each exchange on which registered |
|---------------------|-------------------|---|
| Common Stock | ALCO | NASDAQ Global Select Market |

Indicate by check mark if the registrant is a well-known seasoned issuer, as defined in Rule 405 of the Securities Act.

Yes No

Indicate by check mark if the registrant is not required to file reports pursuant to Section 13 or Section 15(d) of the Act.

Yes No

Indicate by check mark whether the registrant has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days. Yes No

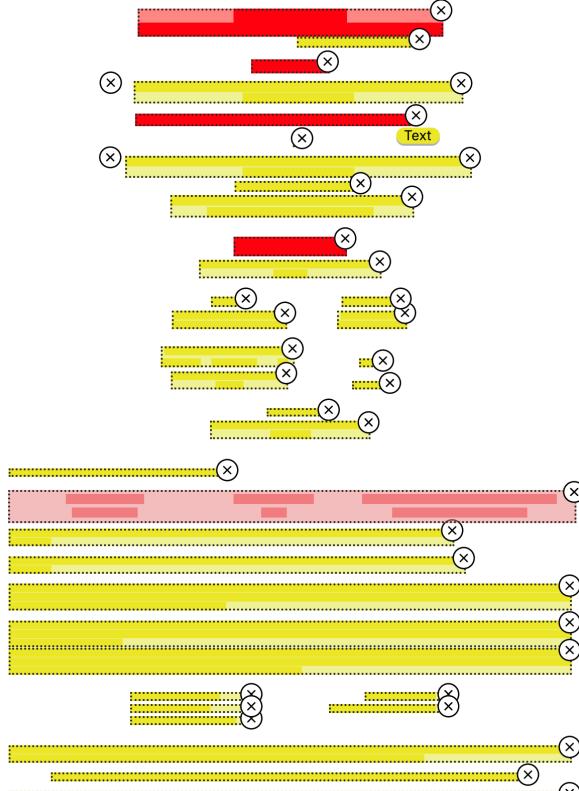
Indicate by check mark whether the registrant has submitted electronically every Interactive Data File required to be submitted pursuant to Rule 405 of Regulation S-T (Rule 405 of this chapter) during the preceding 12 months (or for such shorter period that the registrant was required to submit such files). Yes No

Indicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, a smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer", "accelerated filer", "smaller reporting company" and "emerging growth company" in Rule 12b-2 of the Exchange Act. (Check one):

Large Accelerated Filer
Non-accelerated Filer
Emerging Growth Company Accelerated Filer
Smaller Reporting Company

If an emerging growth company, indicate by check mark if the registrant has elected not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange Act. Yes No

The aggregate market value of the voting and nonvoting common equity held by non-affiliates based on the closing price, as quoted on the Nasdaq



8.9.5.2 Free-form Annual Reports

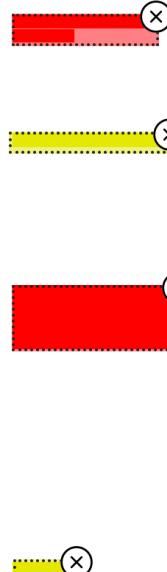
In our first example of free-form annual reports, "Suncor" is the company, and "Meadow Creek East Project" is the title. Both are important, and the company name is larger, so we made them both "Title" and the rest "Text".

MEADOW CREEK EAST
PROJECT

2019 Annual Project Report

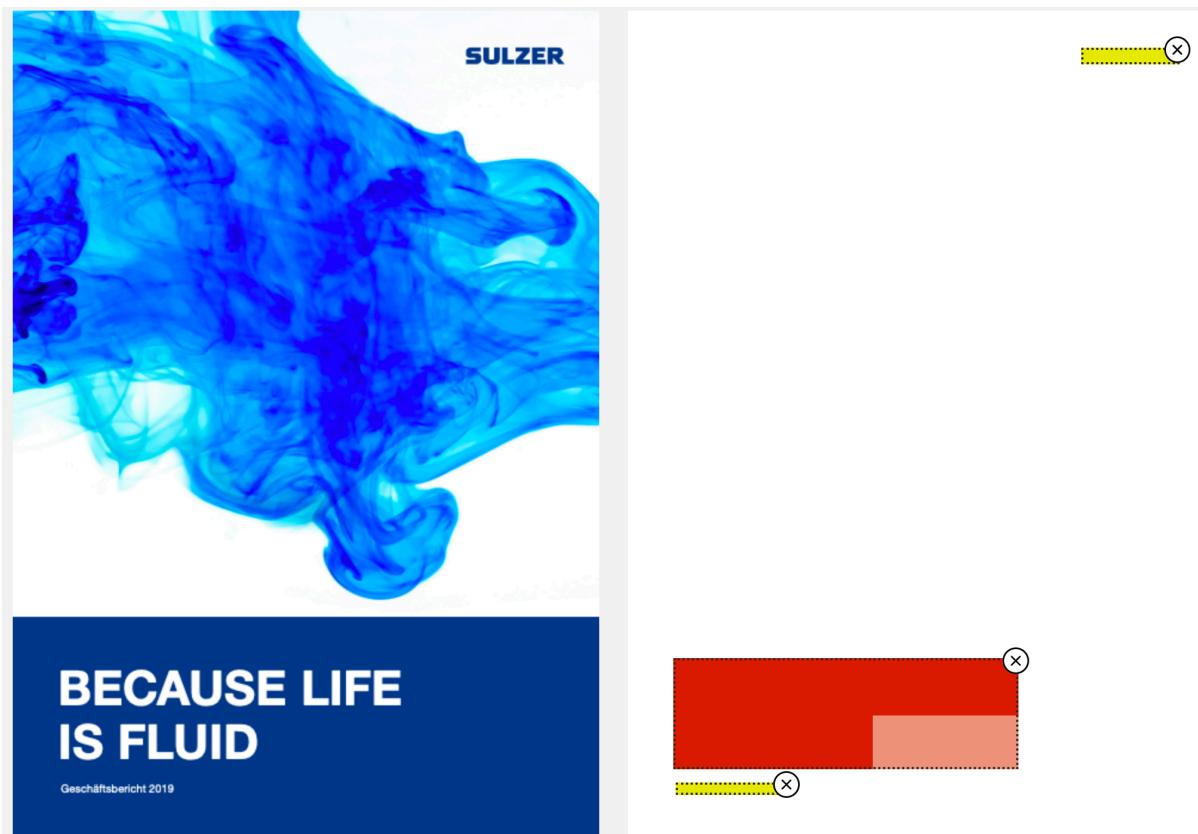
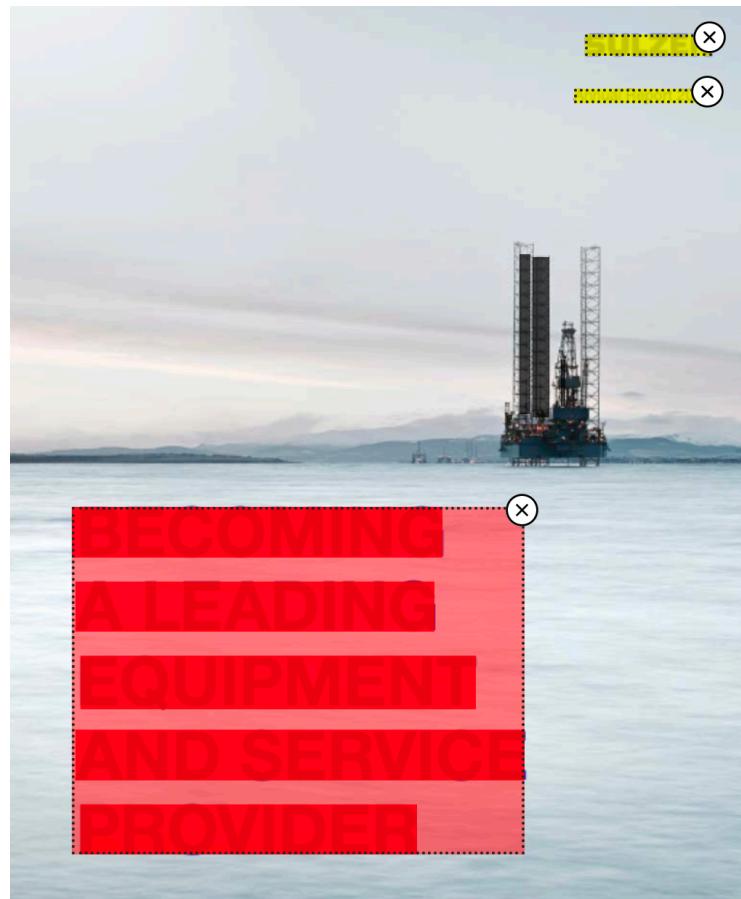


June 28, 2019

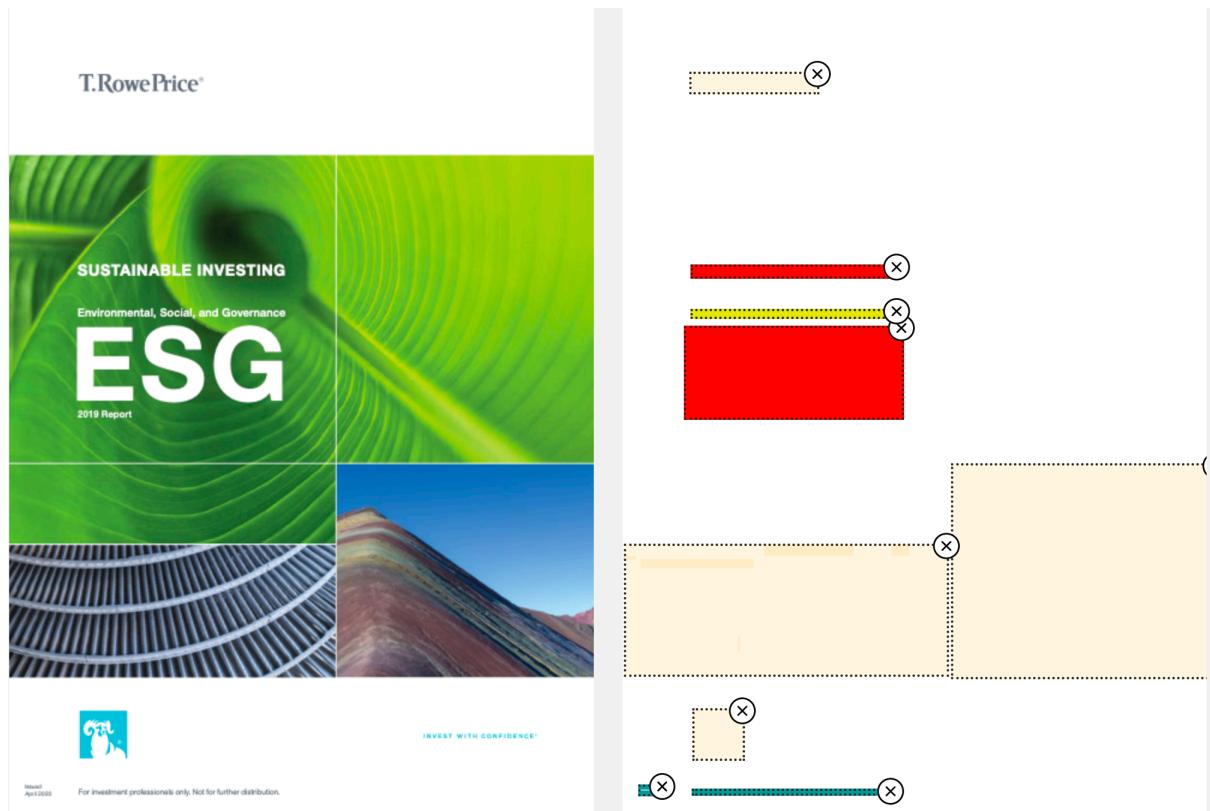


Many reports have pictures on the title page. If one can draw a picture rectangle around the entire picture without overlap with the important title and company name text, then that is

good. But do not annotate *parts* of a picture if another part is behind the text. Two examples of that follow:



Here another example:



"ESG" is by far largest, but only 3 letters. So we also made the next-largest, more title-like text "Sustainable Investing" a title cluster. It would be nicer if "T.RowePrice" were an annotatable text cluster (so one should input an error report too, also for the missing "Invest with confidence"), but we made it at least a picture, because it is the company name and should not get lost.

In the next example, we did not add "IDC Perspective" to the title (but we had a little background knowledge that this is the name of a series of such articles, not the title of the individual article).

IDC PERSPECTIVE
Adoption of Public Cloud in Global Banking: Findings from
IDC's Industry CloudPath Survey, 2019
Jerry Silva



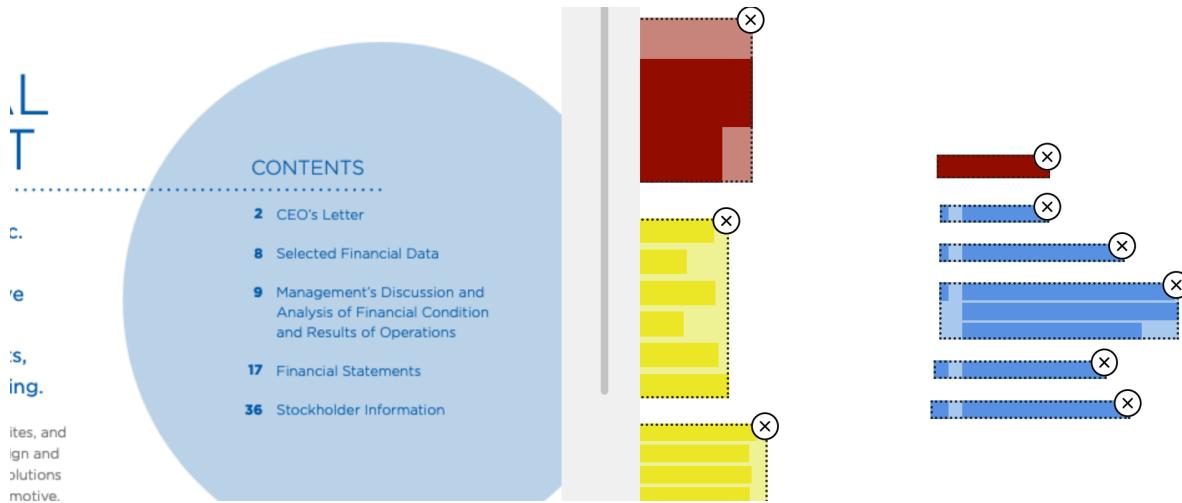
If you find text or other elements that you cannot annotate (e.g., no cell is present), please use the "Report error" feature in the lower right corner. If the problem is only minor, then submit the page anyway, while if it concerns the title itself or other large text, do not submit the page.

8.10 Tables of Content

8.10.1 Neither Page Numbers nor Lines: List-items

If tables of content do not have page numbers, then they usually look like lists, possibly with section headers in between. We saw several such examples among the laws above already.

The following example from an annual report is a clear list: There are hanging paragraphs pretty much as in any other list.⁸⁶



8.10.2 With Gridlines between Items: Table

For the following table of contents, "Table" seemed more suitable because of the complete lines between all rows. This seems to be a rare case though.

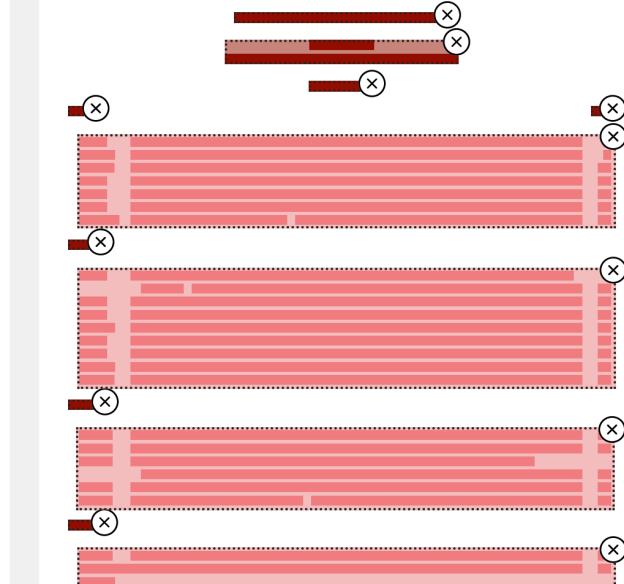
| Item | Page Number |
|-------------------------------|-------------|
| Introductory pages | 1 |
| Group Financial Highlights | 2 |
| Divisional Highlights | 3 |
| About AGL | 4 |
| Our Integrated Strategy | 5 |
| Chairman's Report | 6 |
| Managing Director's Report | 8 |
| Sustainable Business Strategy | 10 |
| Economic | 12 |
| Customers | 14 |
| Community | 16 |
| People and Safety | 18 |
| Climate Change | 20 |
| Environment | 22 |
| Retail Energy | 24 |
| Merchant Energy | 26 |
| Investments | 29 |
| Upstream Gas | 30 |
| Leadership Team | 32 |
| Corporate Governance | 33 |
| Reporting Contents | 42 |

⁸⁶ Page 30b083cac5d434c3aa429de8579e21d9f32d1e232401b5596a76f1a628cd9f4a

8.10.3 With a Page Number Column: Usually Table

Some tables of content have a separate column for page numbers. Then they often look more like tables than lists, and we annotate them as such. Some will be only 1 table, others have intermediate headers that look like starting a new Table, as in the following example.⁸⁷

| LOCKHEED MARTIN CORPORATION | |
|---|----|
| FORM 10-K | |
| For the Fiscal Year Ended December 31, 2011 | |
| CONTENTS | |
| Part I | |
| Item 1 Business | 3 |
| Item 1A Risk Factors | 8 |
| Item 1B Unresolved Staff Comments | 16 |
| Item 2 Properties | 16 |
| Item 3 Legal Proceedings | 16 |
| Item 4 Mine Safety Disclosures | 17 |
| Item 4(a) Executive Officers of the Registrant | 17 |
| Part II | |
| Item 5 Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities | 19 |
| Item 6 Selected Financial Data | 21 |
| Item 7 Management's Discussion and Analysis of Financial Condition and Results of Operations | 22 |
| Item 7A Quantitative and Qualitative Disclosures About Market Risk | 48 |
| Item 8 Financial Statements and Supplementary Data | 50 |
| Item 9 Changes in and Disagreements with Accountants on Accounting and Financial Disclosure | 83 |
| Item 9A Controls and Procedures | 83 |
| Item 9B Other Information | 85 |
| Part III | |
| Item 10 Directors, Executive Officers and Corporate Governance | 86 |
| Item 11 Executive Compensation | 86 |
| Item 12 Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters | 87 |
| Item 13 Certain Relationships and Related Transactions, and Director Independence | 88 |
| Item 14 Principal Accounting Fees and Services | 88 |
| Part IV | |
| Item 15 Exhibits and Financial Statement Schedules | 89 |
| Signatures | 93 |
| Exhibits | |



Here is another example. In both examples, we use “Section-header” and not “Caption” for intermediate larger headings without page numbers.⁸⁸

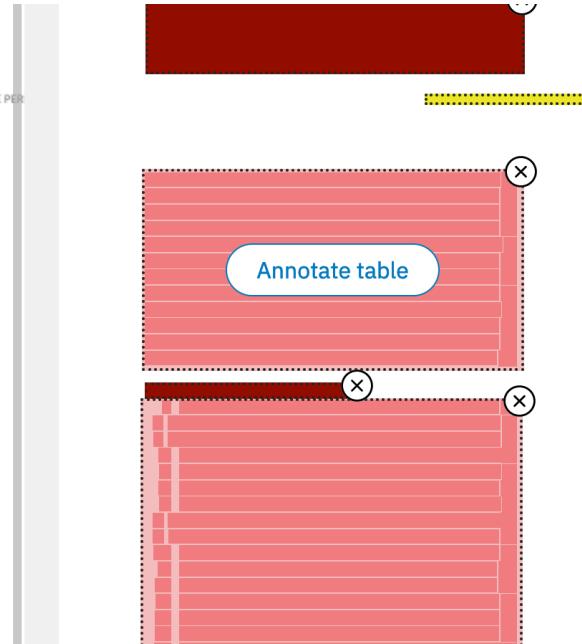
TABLE OF CONTENTS

FAIRFAX MEDIA LIMITED AND CONTROLLED ENTITIES FOR THE PERIOD ENDED 31 DECEMBER 2011

| | |
|--|-----------|
| Board of Directors | 22 |
| Directors' Report | 24 |
| Auditor's Independence Declaration | 28 |
| Remuneration Report | 29 |
| Remuneration Report (Audited) | 31 |
| Corporate Governance | 45 |
| Management Discussion and Analysis Report | 54 |
| Consolidated Income Statement | 56 |
| Consolidated Statement of Comprehensive Income | 57 |
| Consolidated Balance Sheet | 58 |
| Consolidated Cash Flow Statement | 59 |
| Consolidated Statement of Changes in Equity | 60 |

Notes to the Financial Statements

| | |
|---|-----------|
| 1. Summary of significant accounting policies | 62 |
| 2. Revenues | 72 |
| 3. Expenses | 73 |
| 4. Significant items | 74 |
| 5. Discontinued operations | 75 |
| 6. Income tax expense | 76 |
| 7. Dividends paid and proposed | 77 |
| 8. Receivables | 78 |
| 9. Inventories | 79 |
| 10. Assets and liabilities held for sale | 79 |
| 11. Other financial assets | 80 |
| 12. Investments accounted for using the equity method | 80 |
| 13. Available for sale investments | 82 |
| 14. Intangible assets | 82 |
| 15. Property, plant and equipment | 85 |



⁸⁷ Page ec14c6391233ee882b3a3a7e3ae345b0d64a9e846edcacc89d00d5e24332b4e8

⁸⁸ Page 06e4030408cc103a91dc9d5036d2018f54b81cba39adfafa1d9cd811bc302a86

In the next example, we decided that the smaller headings like “1 Discovering Grammar” were not so large and distanced to justify breaking the table up.⁸⁹

| Contents | |
|--|-----------|
| List of illustrations | ix |
| Preface | xi |
| PART ONE: THE BASICS | |
| 1 Discovering Grammar | 3 |
| Identifying nouns | 3 |
| Using capital letters | 5 |
| Replacing nouns with pronouns | 7 |
| Knowing the articles | 10 |
| Understanding verbs | 10 |
| Revising the points | 17 |
| Practising what you've learnt | 17 |
| 2 Expanding your Knowledge | 19 |
| Making words 'agree' | 19 |
| Introducing clauses | 21 |
| Handling phrases | 25 |
| Using adjectives to colour your writing | 25 |
| Employing adverbs | 27 |
| Using prepositions | 28 |
| Revising the points | 29 |
| Practising what you've learnt | 30 |
| 3 Polishing up your Punctuation | 32 |
| Knowing when to stop | 32 |
| Using commas correctly | 33 |
| Making use of the semi-colon, the colon and the dash | 37 |

The diagram illustrates a table of contents from a book. The table is set against a light gray background and features a vertical gray bar on the left side. The content is organized into sections and subsections with corresponding page numbers. A red rectangular box highlights the 'Table' section. A blue callout bubble with the text 'Annotate table' points to this red box. Three red 'X' marks are positioned at the top right corner of the table area.

⁸⁹ Page b8b4379e65d264e08d5d6d9dc3613c4ac7bc11f6e8b33f9acc799e3fea52c1e6

8.10.4 Page Number Column, but Deeply Indented: List-items

In the following example, we hesitated, but decided to treat it as a multi-level list, because the hanging paragraphs at various indentation levels optically dominate over the column with the page numbers.⁹⁰

| Contents | |
|---|----|
| I. Introduction | 1 |
| II. Inelastic atom - molecule collisions | 2 |
| A. Vibrational and rotational relaxation | 2 |
| 1. Collisions at cold and ultracold temperatures | 2 |
| 2. Shape resonances in molecular collisions | 6 |
| 3. Feshbach resonances in molecular collisions | 6 |
| B. Quasi-resonant transitions | 7 |
| C. Atom - molecular ion collisions | 8 |
| III. Chemical Reactions at ultracold temperatures | 9 |
| A. Tunneling dominated reactions | 9 |
| 1. Reactions at zero temperature | 10 |
| 2. Feshbach resonances in reactive scattering | 12 |
| B. Barrierless reactions | 14 |
| 1. Collision systems of three alkali metal atoms | 14 |
| 2. Role of PES in determining ultracold reactions | 16 |
| 3. Relaxation of vibrationally excited alkali metal dimers | 18 |
| 4. Reactions of heteronuclear and isotopically substituted alkali-metal dimer systems | 19 |
| IV. Inelastic molecule - molecule collisions | 20 |
| A. Molecules in the ground vibrational state | 20 |
| B. vibrationally inelastic transitions | 22 |

You will need to use your judgement between “Table” and “List-item” or “Text”, and in case of tables how many tables to make.

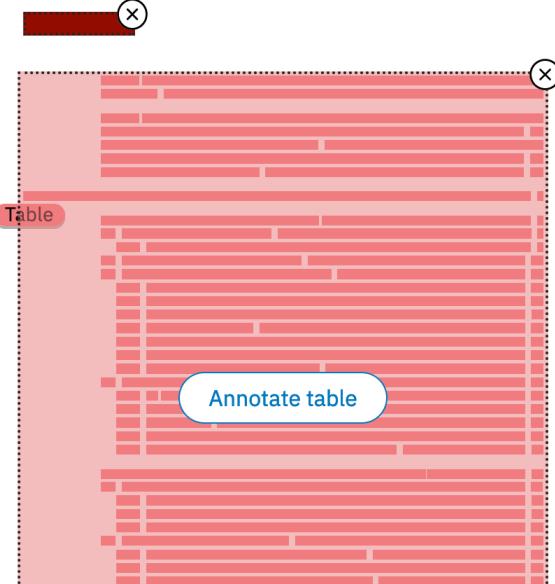
⁹⁰ Page 99a7c31511b8a31f7d065b2f56617f3fc58b8abf13bb9674b45dbfcebe9af3

8.10.5 Special Case Manuals / Redbooks

In the rather homogeneous “Redbooks” collection in the “Manuals” category, even the “Parts” have page numbers. Thus we consider everything as one “Table”.⁹¹

Contents

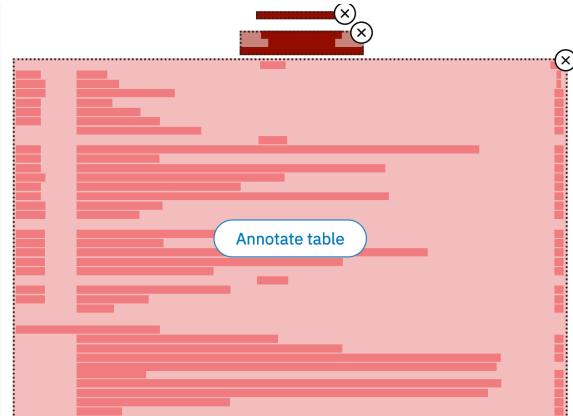
| | |
|---|------|
| Notices | xii |
| Trademarks | xii |
| Preface | xiii |
| Authors | xiv |
| Now you can become a published author, too! | xv |
| Comments welcome | xvi |
| Stay connected to IBM Redbooks | xvi |
| Part 1. Concepts and architecture | |
| Chapter 1. Introduction to the IBM DS8800 | 1 |
| 1.1 Introduction to the IBM DS8800 | 3 |
| 1.1.1 At-a-glance features of the DS8800 | 4 |
| 1.1.2 DS8800 controller options and frames | 7 |
| 1.2 DS8800 architecture and functions overview | 11 |
| 1.3.1 Overall architecture and components | 12 |
| 1.3.2 Storage capacity | 16 |
| 1.3.3 Supported environments | 16 |
| 1.3.4 Configuration flexibility | 17 |
| 1.3.5 Copy Services functions | 19 |
| 1.3.6 Service and setup | 19 |
| 1.3.7 IBM Certified Secure Data Overwrite | 20 |
| 1.4 Performance features | 20 |
| 1.4.1 16 Gbps host adapters | 20 |
| 1.4.2 Sophisticated caching algorithms | 20 |
| 1.4.3 Flash storage | 21 |
| 1.4.4 Performance for IBM Z systems | 22 |
| 1.4.5 Performance enhancements for IBM Power Systems | 23 |
| Chapter 2. IBM DS8800 hardware components and architecture | 25 |
| 2.1 Flash drive terminology of the DS8800 | 26 |
| 2.1.1 Storage system | 26 |
| 2.1.2 Management Console | 27 |
| 2.1.3 Central processor complex | 27 |
| 2.2 DS8800 configurations and models | 28 |
| 2.2.1 DS8888F Analytic Class all-flash configuration | 29 |
| 2.2.2 DS8886F Enterprise Class all-flash configuration | 30 |
| 2.2.3 DS8884F Business Class all-flash configuration | 31 |



8.10.6 Special Case Financial Reports / SEC Filings

SEC filings whose table of contents has page numbers usually have the following format:

| | |
|--|----|
| THE DIXIE GROUP, INC. | |
| Index to Annual Report | |
| Form 10-K for the Year Ended December 26, 2020 | |
| PART I | |
| Item 1. Business | 4 |
| Item 1A. Risk Factors | 5 |
| Item 1B. Unresolved Staff Comments | 11 |
| Item 2. Properties | 11 |
| Item 3. Legal Proceedings | 11 |
| Item 4. Mine Safety Disclosures | 12 |
| Executive Officers of the Registrant | 13 |
| PART II | |
| Item 5. Market for the Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities | 14 |
| Item 6. Selected Financial Data | 17 |
| Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations | 18 |
| Item 7A. Quantitative and Qualitative Disclosures About Market Risk | 23 |
| Item 8. Financial Statements and Supplementary Data | 23 |
| Item 9. Changes in and Disagreements with Accountants on Accounting and Financial Disclosure | 28 |
| Item 9A. Controls and Procedures | 26 |
| Item 9B. Other Information | 26 |
| PART III | |
| Item 10. Directors, Executive Officers and Corporate Governance | 27 |
| Item 11. Executive Compensation | 27 |
| Item 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters | 27 |
| Item 13. Certain Relationships and Related Transactions, and Director Independence | 27 |
| Item 14. Principal Accounting Fees and Services | 27 |
| PART IV | |
| Item 15. Exhibits and Financial Statement Schedules | 28 |
| Item 16. Form 10-K Summary | 28 |
| Signatures | 28 |
| CONSOLIDATED FINANCIAL STATEMENTS | |
| Report of Independent Registered Public Accounting Firm | 33 |
| Consolidated Balance Sheets - December 26, 2020 and December 28, 2019 | 34 |
| Consolidated Statements of Operations - Years ended December 26, 2020, December 28, 2019, and December 29, 2018 | 35 |
| Consolidated Statements of Comprehensive Income (Loss) - Years ended December 26, 2020, December 28, 2019, and December 29, 2018 | 36 |
| Consolidated Statements of Cash Flows - Years ended December 26, 2020, December 28, 2019, and December 29, 2018 | 37 |
| Consolidated Statement of Stockholders' Equity - December 26, 2020, December 28, 2019, and December 29, 2018 | 38 |
| Notes to Consolidated Financial Statements | 38 |
| Exhibit Index | 58 |



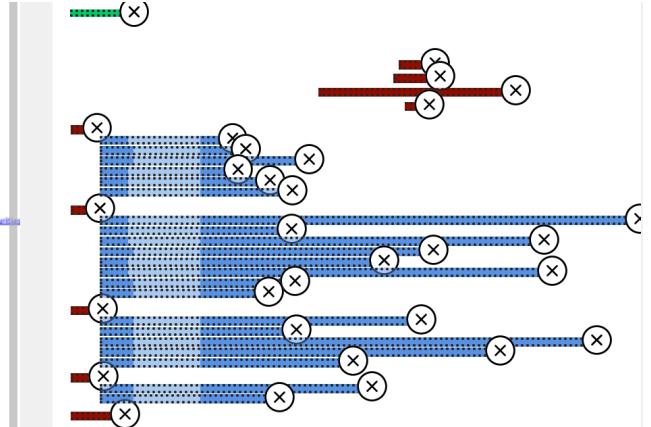
Here we made 1 Table from everything inside the blue-and-white striping. In tables where that is not present, intermediate headers might become Section-headers between several smaller tables.

Some SEC filings have a table of contents without page numbers, which we label like this:⁹²

⁹¹ Page fc2f07d017ac559f4fc0096bb11783b3839a3c68dd74aa29b1722720d6a1b48a

⁹² Page afc44bff6318df33a3125b55556464fe87becd302af9a4d00afc7e8a85f1de8e

Table of Contents

| | |
|--|--|
| AZZ INC. FORM 10-K For the Fiscal Year Ended February 28, 2021 INDEX |  |
| PART I | |
| Item 1. Business | (x) |
| Item 1A. Risk Factors | (x) |
| Item 2. Properties | (x) |
| Item 3. Legal Proceedings | (x) |
| Item 4. Mine Safety Disclosures | (x) |
| PART II | |
| Item 5. Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities | (x) |
| Item 6. Selected Financial Data | (x) |
| Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operation | (x) |
| Item 7A. Quantitative and Qualitative Disclosures About Market Risk | (x) |
| Item 8. Financial Statements and Supplementary Data | (x) |
| Item 9. Controls and Procedures | (x) |
| Item 9A. Other Information | (x) |
| PART III | |
| Item 10. Directors, Executive Officers and Corporate Governance | (x) |
| Item 11. Executive Compensation | (x) |
| Item 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters | (x) |
| Item 13. Certain Relationships and Related Transactions, and Director Independence | (x) |
| Item 14. Principal Accounting Fees and Services | (x) |
| PART IV | |
| Item 15. Exhibits and Financial Statement Schedules | (x) |
| Item 16. Form 10-K Summary | (x) |
| SIGNATURES | |

8.11 Index Pages

Long documents, in particular manuals, may have an index near the end, i.e., an alphabetical list of keywords with the pages numbers where these words can be found. The general rules should be applied as much as possible, similar to how we decided for tables of content.

8.11.1 Redbook Indices

Specifically for the rather homogenous collection “Redbooks” in the “Manuals” category, we decided to use “Text” clusters where the keywords are close together (following Section 3.1.7), in spite of occasional indentation, as in the following example.⁹³

⁹³ Page 24193e0c5935b64ccaa6e578aefe3b161ed64522b1f03a81ba594e8a0c9e574f

Index

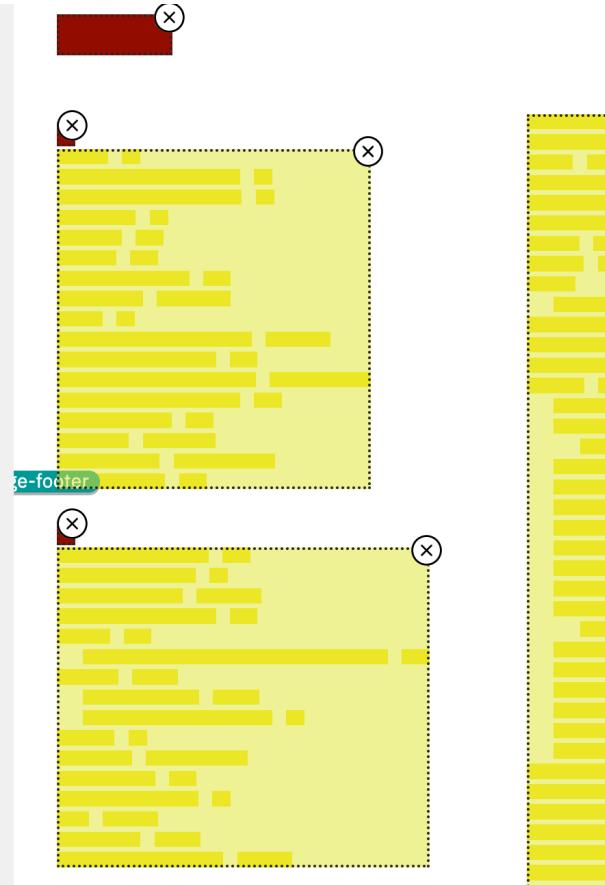
A

access 59
 active-active architecture 20
 Advanced Copy Services 72
 alias string 62
 analyses 674
 analysis 740
 application testing 437
 architecture 9, 20, 674
 arrays 28
 asynchronous notifications 464, 466
 asynchronous remote 525
 asynchronous remote copy 522, 524, 526
 asynchronous replication 543
 asynchronously 524
 audit logs 726–727
 authentication 334, 756, 764
 auxiliary VDisk 525

B

back-end application 777
 background copies 17
 background copy 523, 532
 background copy rate 461
 backup 436
 of data with minimal impact on production 458
 backups 16, 41
 application data 23, 38
 system configuration data 23
 balance 64
 bandwidth 37, 45, 64, 77
 bitmap space 250
 block-level protocol 61
 boot 78, 682
 bottlenecks 80, 82
 business requirements 80, 740

Challenge-I
 channel ext
 CHAP 61
 CHAP auth
 CHAP secr
 chat sessio
 chunks 65
 CIMOM 23
 cluster
 IP addr
 cluster (SVI)
 cluster part
 clustered s
 clusters 22
 backups
 configur
 data
 configur
 error log
 extents
 I/O grou
 image M
 Image M
 IP addr
 MDisks
 data
 node da
 node pa
 partners
 public k
 shutdow
 unmanag
 colliding wr
 command-l
 762, 766, 7
 commands
 compressed
 compressio



8.12 Final Pages

Long documents often have special final pages. However, among the free-form documents, we can only give specific guidance for the almost homogeneous Redbooks collection.

8.12.1 Redbook Final Pages

The last page of many redbooks looks like this, with our labeling standard:⁹⁴



⁹⁴ Page fddfb5d4501ab1be90dd7ec2a1b12f3db63a9431a478c99a58af6bf5741d559