

Get your docs in a row* with Docling

* reliably organized way

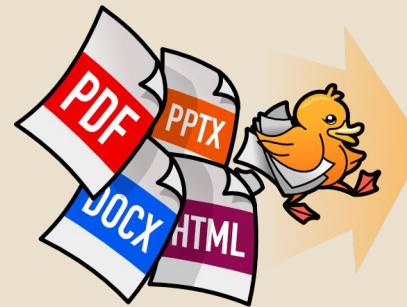


Carol Chen, Red Hat

@cybette@mastodon.org.uk | [@cybette:matrix.org](https://matrix.to/#/@cybette:matrix.org) | linkedin.com/in/cybette

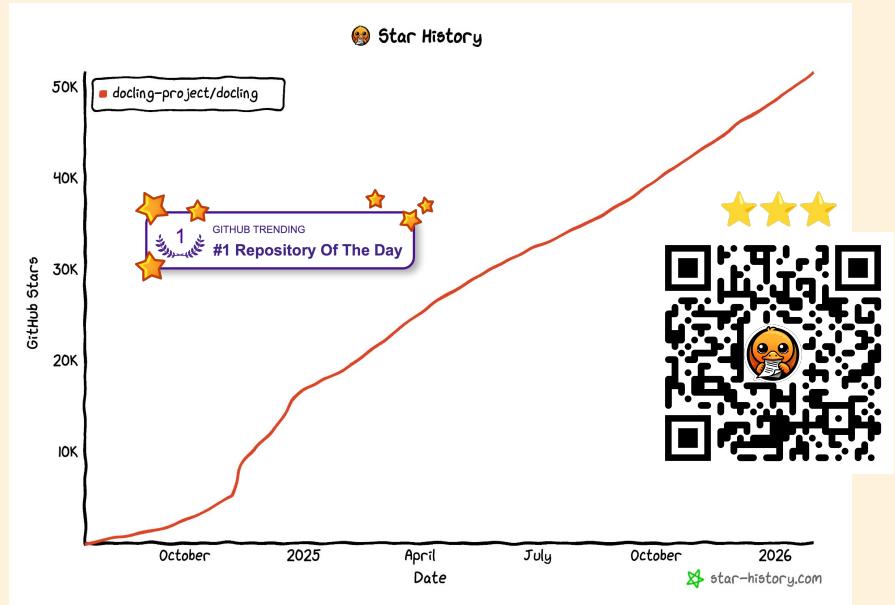
Introducing Docling

-  Parsing of multiple document formats incl. PDF, DOCX, XLSX, HTML, images, and more
-  Advanced PDF understanding incl. page layout, reading order, table structure, code, formulas, image classification, ...
-  Unified, expressive DoclingDocument representation format
-  Various export formats (Markdown, HTML, JSON)
-  Local execution for sensitive data and air-gapped environments
-  Many plug-and-play ecosystem integrations
-  Extensive OCR support for scanned PDFs and images
-  Support of Visual Language Models
-  Support for Audio with Automatic Speech Recognition model
-  Simple and convenient CLI



Docling project

- Community adoption
- ⭐ 51k+ GitHub stars
- 🐍 2.4M+ downloads last month from [PyPI](#)



 /docling-project



The quick brown fox

You know the story...
or do you?

(image generated by Gemini)



Vegetative electron microscopy

<https://www.freethink.com/artificial-intelligence/vegetative-electron-microscopy>

gurovdigital 15 h

lol, over 20 scientific papers now feature the nonsensical term 'vegetative electron microscopy'.

all because an AI misinterpreted a 1959 article, merging 'vegetative' and 'electron microscopy' from separate columns.

enzyme present in spores, used from spore coats of *B. cereus*, a composition similar to the cell wall, from the results of Norris of Leeds University (1959). He treated spores with preparation of lytic enzymes and examined the electron microscopy. No exosporium was obtained.

In the vegetative cell, a sporangium. It is by no means known what happens to the vegetative pore is released. In *Clostridium* it appears that at least part of the membrane is an outer membrane and the opinion of some is the opinion of some all of the sporulating cell which consists of an outer

that the exosporium of *B. cereus* is composed of the enzymes from *B. cereus*. The enzyme did not attack the exospore, but it did attack the cell wall, from the results of Norris of Leeds University (1959). He treated spores with preparation of lytic enzymes and examined the electron microscopy. No exosporium was obtained.

It was not known whether the enzyme present in spores, or another enzyme, was responsible for lysis of the sporangial walls. When thick suspensions of living cells of *B. cereus* were

...

15 h ago · 692 · 12 · 43 · 7

Date syrup (as one of the agricultural wastes) was used to produce bacterial cellulose using Gluconostobacter xylinus. Fourier transform infrared spectroscopy (FTIR), vegetative electron microscopy, and X-ray diffraction were used to determine the structure of bacterial cellulose, cellulose fibers, and crystallinity of the samples (Moosavi and Yousefi, 2011). After 14 days of incubation at 28 °C, the highest yield of cellulose



Silver and gold nanoparticles for antimicrobial purposes against multi-drug resistance bacteria [HTML] m

N Rabiee, S Ahmadi, O Akhavan, R Luque - Materials, 2022 - mdpi.com

... Dead bacteria have been observed by imaging and elemental analysis using transmission electron microscopes (TEM), vegetative electron microscopy, and EDX (X-Ray Probe Microscopy).

☆ Cited by 112 Related articles

Study of CNT@Fe3O4 effects on Aeromonas hydrophila and Yersinia ruckeri bacteria isolated from fish. M Alishti, KR Tavabe, A Mirvaghefi - Journal of ... 2019 - search.ebscohost.com

-carbon nanotubes synthesized by spectroscopic and microscopic techniques including X-ray diffraction spectrum, shaking sample magnetometer and vegetative electron microscopy ...

Cited by 1 Related articles

Green synthesis of silver nanoparticles via Ganoderma lucidum fungus extract and its antibacterial effects on Klebsiella pneumonia isolates from urinary tract ... M Jamschidian-Mojaver, M Amiri - Alborz University ..., 2021 -

Vegetative electron microscopy was used to measure the organic compounds in the sample. IR analysis was also performed to investigate possible organic compounds that ...

Cited by 1 Related articles

METALLOGRAPHIC STUDIES OF IRAN'S IRON AGE: CASE STUDY BRONZE PIECES FROM JEYRĀN TEPE, OZBAKİ B SODAEI, H RAHNEMA - researchgate.net

This study is a report of the results of metallographic study of 5 bronze pieces found in Jeyrān Tepe dating back to the Iron

Vegetative cell wall

Output via Docling

were incubated with an extract from spores disintegrated at pH 7.0. Peptide was released which established that the coats contained substrate for the lytic enzyme present in spores. Peptide was also released from spore coats of *B. megaterium* by the action of the enzyme from *B. cereus* spores. The lytic enzyme did not attack intact resting spores.

The spore develops in the vegetative cell, which thus becomes a sporangium. It is by no means certain what happens to the vegetative cell wall when the spore is released. In *Clostridium* species it appears that at least part of this structure is retained as an outer membrane around the spore. It is the opinion of some workers that the wall of the sporulating cell forms the exosporium which exists as an outer

acteristic type. It was concluded that at least part of the sporangial wall was dissolved away to allow release of the spore. It appears likely that the exosporium of *B. cereus* does not have a composition similar to that of the vegetative cell wall, from the results obtained by Dr. J. R.

The spore develops in the vegetative cell, which thus becomes a sporangium. It is by no means certain what happens to the vegetative cell wall when the spore is released. In *Clostridium* species it appears that at least part of this structure is retained as an outer membrane around the spore. It is the opinion of some workers that the wall of the sporulating cell forms the exosporium which exists as an outer coat around spores of several *Bacillus* species. Spores of several varieties of *B. cereus* had exospria whereas these structures appeared to be absent from spores of *B. megaterium* and *B. subtilis*. It seems, however, that in *Bacillus* species at least, the greater part of the vegetative cell wall is dissolved away before the developed spore is released. If this is true, then soluble components containing the characteristic constituents should appear in the medium during spore release. Culture filtrates from *B. cereus* organisms at various stages of growth and sporulation were hydrolyzed and the hydrolyzates analyzed for amino sugars and diaminopimelic acid (28). Results showed that a large increase in the concentration of these substances in the culture filtrate occurred during spore release (table 2); they were found to be present in a nondialyzable peptide of the characteristic type. It was concluded that at least part of the sporangial wall was dissolved away to allow release of the spore. It appears likely that the exosporium of *B. cereus* does not have a composition similar to that of the vegetative cell wall, from the results obtained by Dr. J. R. Norris of Leeds University (personal communication). He treated spores with a highly active preparation of lytic enzyme from *B. cereus* spores and examined the effect by means of electron microscopy. No evidence of lysis of the exosporium was obtained.



Recovering structured content from PDF

with low level PDF parsers

KDD '22, August 14–18, 2022, Washington, DC, USA Birgit Pfitzmann, Christoph Auer, Michele Dolli, Ahmed S. Nassar, and Peter Staar

Table 1: DocLayNet dataset overview. Along with the frequency of each class label, we present the relative occurrence (as % of row “Total”) in the train, test and validation sets. The inter-annotator agreement is computed as the mAP@0.5-0.95 metric between pairwise annotations from the triple-annotated pages, from which we obtain accuracy ranges.

class label	Count	% of Total									
		Train	Test	Val	All	Pn	Man	Rd	Law	Pat	Ten
Caption	22524	1.77	2.32	84.89	40.61	86.92	95.49	n/a	n/a	n/a	n/a
Footnote	6318	2.04	0.31	0.58	83.91	n/a	10	62.88	85.94	n/a	62.97
Formal	25027	2.25	1.90	0.30	84.89	40.61	86.92	95.49	62.88	85.94	69.71
List-item	185660	17.19	13.34	15.82	87.88	74-97	96-92	97-97	81-95	75-85	93-95
Page-header	70878	6.51	5.58	4.00	93.94	88-90	95-96	100	92-97	100	96-98
Page-footer	58022	3.16	0.70	0.70	85.82	84-86	94-96	97-99	80-85	89-91	94-96
Text	45976	4.21	2.78	3.51	83.71	56-59	82-84	89-92	80-95	89-92	89-95
Section-header	142884	12.66	15.77	12.85	83.84	76-81	96-92	94-95	87-95	69-71	78-86
Table	34733	3.20	2.27	3.40	77.81	75-80	83-86	89-99	58-80	79-84	70-85
Figure	510377	4.01	4.01	4.01	83.94	40-43	84-87	89-92	80-83	87-90	87-92
Title	5071	0.67	0.30	0.50	60-72	24-63	50-63	94-100	82-86	68-79	24-56
Total	1107470	9.4123	99816	66531	82-83	71-74	79-81	89-94	86-91	71-78	48-85

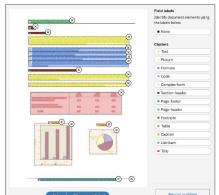


Figure 3: Corpus Computation Service annotation user interface. The interface shows a document page with several text segments highlighted by red boxes. The annotation boxes can be drawn by dragging a rectangle over each segment with the respective label from the palette on the right.

We distributed the annotation workload and performed continuous quality controls. Phase one and two required a small team of experts only. For phases three and four, a group of 40 dedicated annotators were assembled and supervised.

Please note that the preprint version of our inclusion criteria for documents were described in Section 3. A large effort went into ensuring that all documents are free to use. The data sources

KDD '22, August 14–18, 2022, Washington, DC, USA Birgit Pfitzmann, Christoph Auer, Michele Dolli, Ahmed S. Nassar, and Peter Staar

Table 1: DocLayNet dataset overview. Along with the frequency of each class label, we present the relative occurrence (as % of row “Total”) in the train, test and validation sets. The inter-annotator agreement is computed as the mAP@0.5-0.95 metric between pairwise annotations from the triple-annotated pages, from which we obtain accuracy ranges.

% of Total

triple inter-annotator mAP @ 0.5-0.95 (%)

[...]

Count

22524

6318

25027

185660

70878

58022

45976

142884

34733

510377

5071

1107470

[...]

! Tables not understood

! Image content missing

! Line wraps not understood

[...]

! Multi-column often breaks order

undesired
page headers

✓ Very fast and cheap

✗ Incomplete

✗ Loss of structure

✗ Noisy

→ Unfit for most use cases



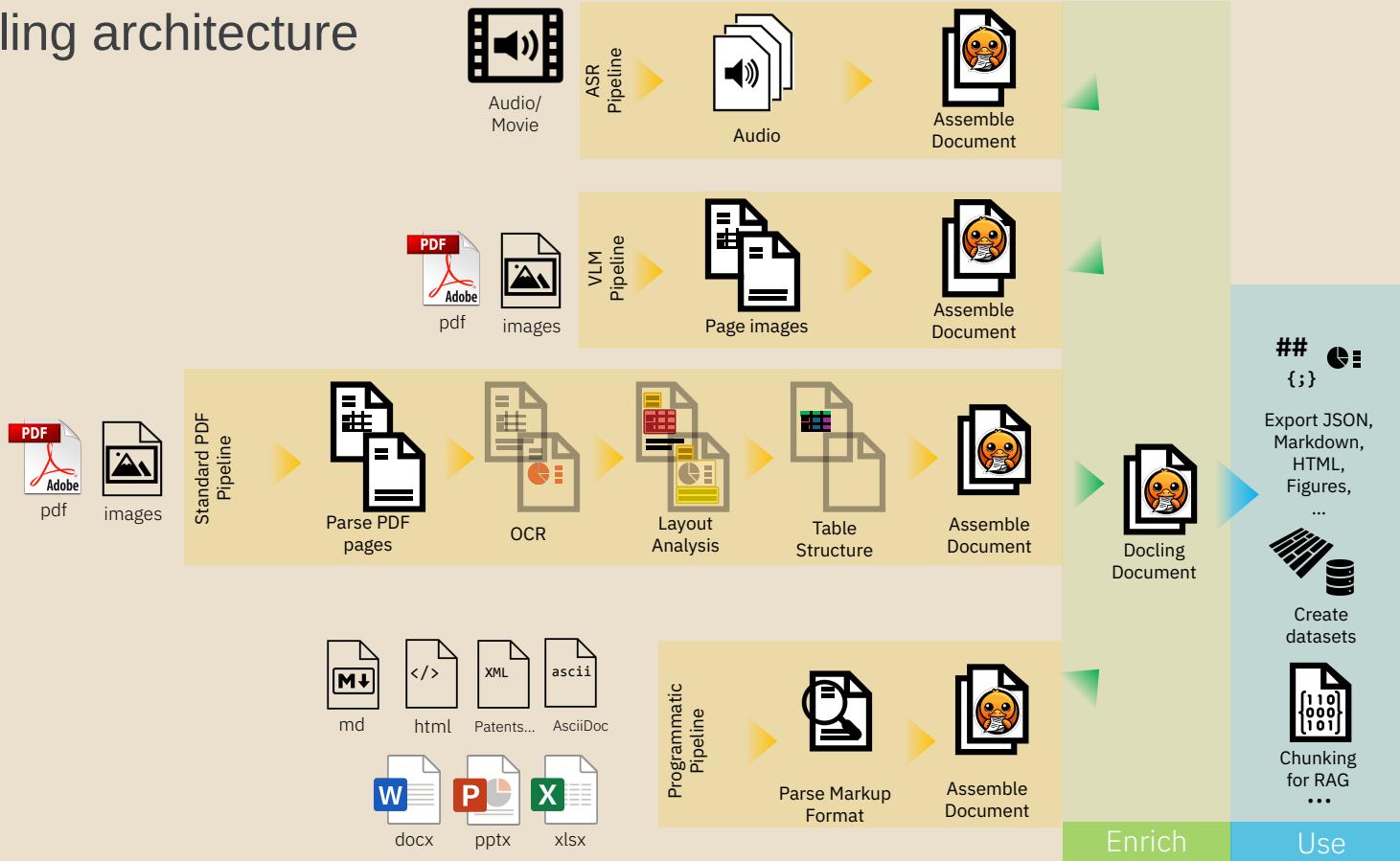
include publication repositories such as arXiv³, government o’ces, company websites as well as data directory services for nancial reports and patents. Scanned documents were excluded wherever possible because they can be rotated or skewed. This would not allow us to perform annotation with rectangular bounding-boxes and therefore complicate the annotation process.

Preparation work included uploading and parsing the sourced PDFs for the DocLayNet dataset (CCS) [22], a cloud native platform which provides a visual annotation interface and allows for document inspection and analysis. The annotation interface is based on a shared document view where users can switch between the different document categories was achieved by selective subselection of pages with certain desired properties. For example, we chose to always select the entire page for a document and bias the remaining page selection to those with images or tables. This was achieved by leveraging pre-trained object detection models [11, 12], which help us estimate how many images and tables a given page contains.

Phase 2: Label selection and guidance. We reviewed the collected annotations and identiﬁed the most common structural features they contained. This was achieved by grouping layout elements and lead to the deinition of 11 distinct class labels. These 11 class labels are *Caption*, *Footnote*, *Formula*, *List-item*, *Page-header*, *Page-footer*, *Text*, *Table*, *Figure*, *Section-header*, and *Title*. Critical factors that were considered for the choice of these class labels were (1) the overall occurrence of the label, (2) the specicity of the label, (3) the overall coverage of the label, (4) the context from previous or next page and (5) overall coverage of the page. Specicity ensures that the choice of label is not ambiguous, while overall occurrence and coverage ensure that the labels can be annotated. We refrained from class labels that are very specic to a document category, such as *Abstract* in the *Scientific Articles* category. We also avoided labels that have a close semantic relation to the semantics of the text, labels such as *Acknowledgments* as seen in DocBank, are often only distinguishable by discriminating on

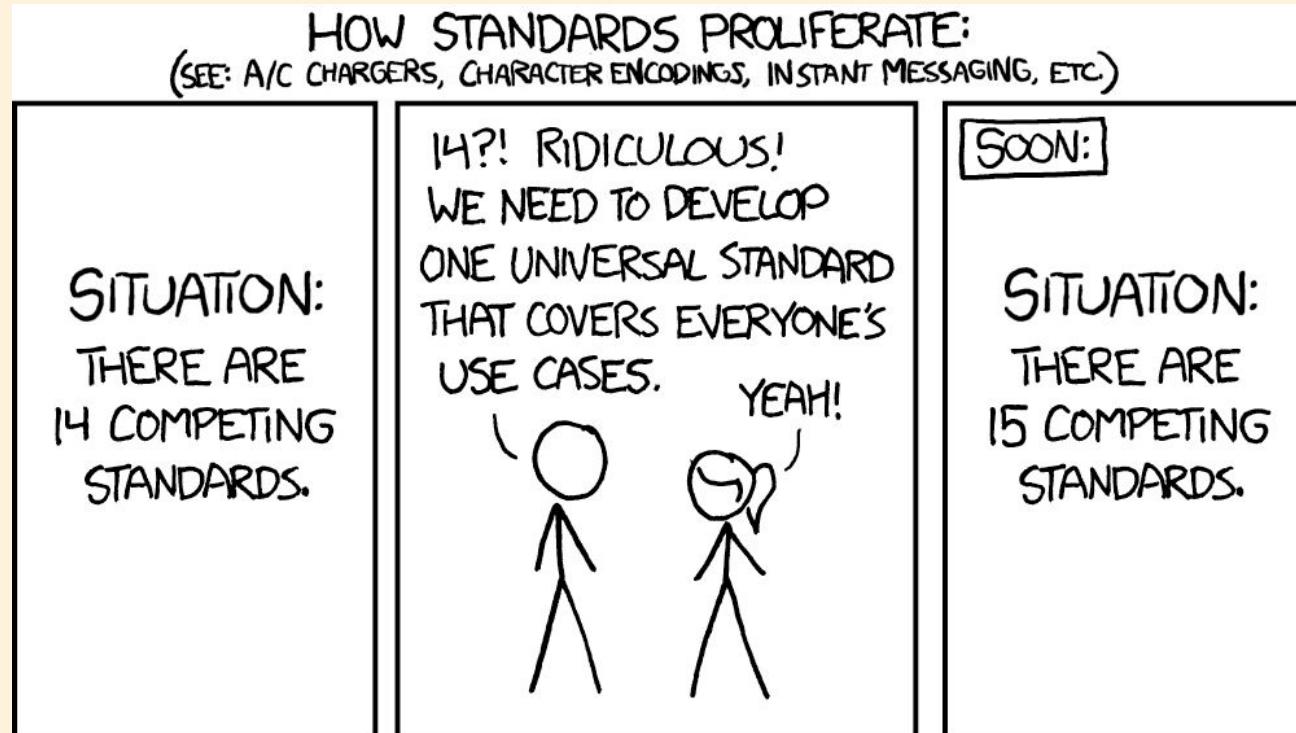
³<https://arxiv.org>

Docling architecture



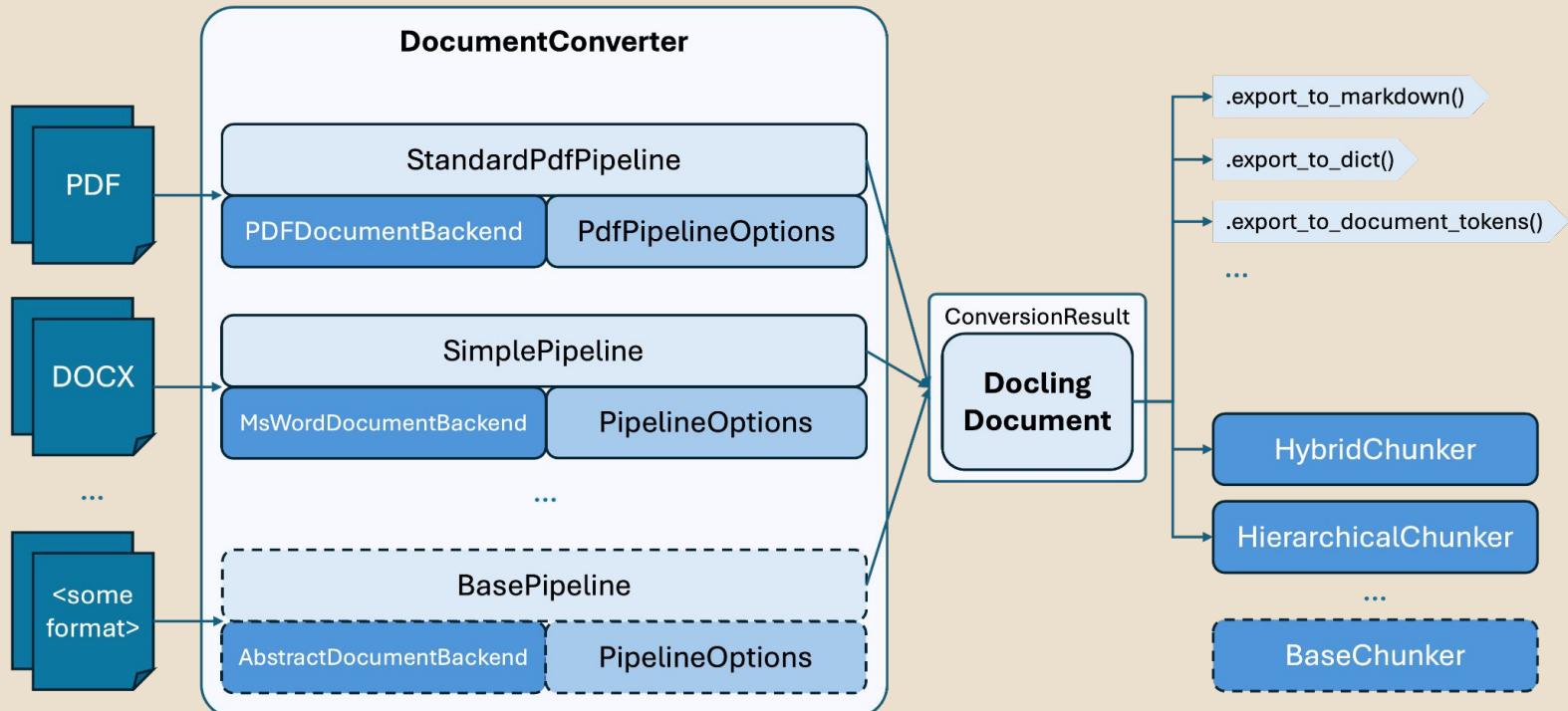
Obligatory xkcd

<https://xkcd.com/927/>



Docling architecture

<https://docling-project.github.io/docling/concepts/architecture/>



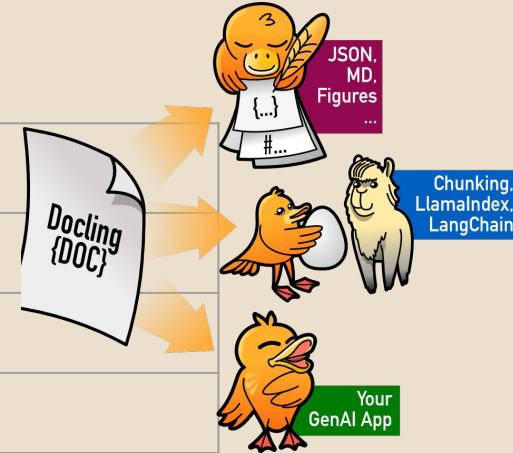
Supported input formats



Format	Description
PDF	
DOCX, XLSX, PPTX	Default formats in MS Office 2007+, based on Office Open XML
Markdown	
AsciiDoc	Human-readable, plain-text markup language for structured technical content
HTML, XHTML	
CSV	
PNG, JPEG, TIFF, BMP, WEBP	Image formats
WebVTT	Web Video Text Tracks format for displaying timed text

Supported output formats

Format	Description
HTML	Both image embedding and referencing are supported
Markdown	
JSON	Lossless serialization of Docling Document
Text	Plain text, i.e. without Markdown markers
Doctags	Markup format for efficiently representing the full content and layout characteristics of a document



https://docling-project.github.io/docling/usage/supported_formats/

DoclingDocument

DoclingDocument is a unified document representation format which can express several features common to documents, such as:

- Text, Tables, Pictures, and more
- Document hierarchy with sections and groups
- Disambiguation between main body and headers, footers (furniture)
- Layout information (i.e. bounding boxes) for all items, if available
- Provenance information



It also brings a set of document construction APIs to build up a DoclingDocument from scratch.

https://docling-project.github.io/docling/concepts/docling_document/

Introducing Duckling

A modern, user-friendly graphical interface for Docling

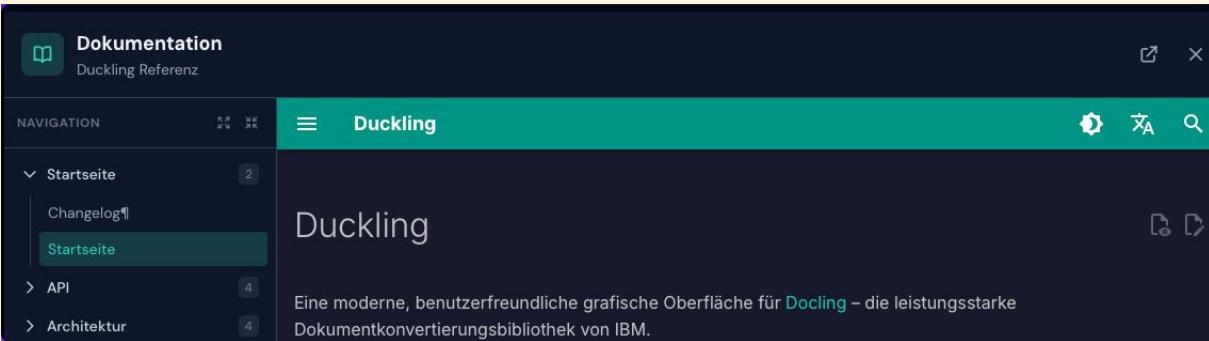
Key features:

- Drag and drop files for instant processing
- Multiple input & output formats
- Image and table extraction
- RAG-ready chunking
- Conversion history

<https://duckling-ui.org/>



Latest release with i18n



Demo (maybe)

Convert Any Document

Transform PDFs, Word documents, presentations, images, and more
into structured formats ready for AI processing.



Drag and drop your document here
or click to browse

Documents: [PDF](#) [DOCX](#) [PPTX](#) [XLSX](#) Web: [HTML](#) [Markdown](#)

Images: [PNG](#) [JPG](#) [TIFF](#) [WebP](#) Data: [XML](#) [AsciiDoc](#)

Maximum file size: 100MB per file



OCR
Extract text from images



Tables
Export to CSV



Images
Extract figures



RAG
Document chunks

Chunking for RAG

Naive Splitting

...Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. The results are shown in the table below.



Column A	Column B
Column A	Column B
...	...
...	...



Breaks context, splits tables, confuses retrieval.

Docling Hybrid Chunking

Header
...Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. The results are shown in the table below.



Column A	Column B
Table	

Respects headers, lists, and table boundaries.

Enriched metadata per chunk helps embedding models understand layout and context, reducing hallucination.

Dockling

- Website: <https://www.docling.ai/>
- Github: <https://github.com/docling-project>
- HF: <https://huggingface.co/docling-project>
- Discord: <https://www.docling.ai/discord>



Duckling

- Website: <https://duckling-ui.org/>
- Github: <https://github.com/davidgs/duckling>



Resources



Acknowledgements:

Docling team (IBM Research Zurich) -
Michele Dolfi, Peter Staar, Panos
Vagenas, Ming Zhao

Duckling author (Red Hat) -
David Simmons

Thank you!