




Inscriptis - A Python-based HTML to text conversion library optimized for knowledge extraction from the Web

Albert Weichselbraun¹

¹ Swiss Institute for Information Science, University of Applied Sciences of the Grisons, Pulvermühlestrasse 57, Chur, Switzerland

DOI: [10.21105/joss.03487](https://doi.org/10.21105/joss.03487)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Pending Editor](#) 

Submitted: 12 July 2021

Published: 12 July 2021

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Inscriptis provides a library, command line client and Web service for converting HTML to plain text.

Its development has been triggered by the need to obtain accurate text representations for knowledge extraction tasks that preserve the spatial alignment of text without drawing upon heavyweight, browser-based solutions such as [Selenium](#). In contrast to existing software packages such as [HTML2text](#), [jusText](#) and [Lynx](#), Inscriptis

1. provides a layout-aware conversion of HTML that more closely resembles the rendering obtained from standard Web browsers and, therefore, better preserves the spatial arrangement of text elements. Inscriptis excels in terms of conversion quality, since it correctly converts complex HTML constructs such as nested tables and also interprets a subset of HTML (e.g., `align`, `valign`) and CSS (e.g., `display`, `white-space`, `margin-top`, `vertical-align`, etc.) attributes that determine the text alignment.
2. supports annotation rules, i.e., user-provided mappings that allow for annotating the extracted text based on structural and semantic information encoded in HTML tags and attributes used for controlling structure and layout in the original HTML document.

These unique features ensure that downstream knowledge extraction components can operate on accurate text representations, and may even use information on the semantics and structure of the original HTML document, if annotation support has been enabled.

Statement of need

Research in a growing number of scientific disciplines relies upon Web content. [Li et al. \(2014\)](#), for instance, studied the impact of company-specific News coverage on stock prices, in medicine and pharmacovigilance social media listening plays an important role in gathering insights into patient needs and the monitoring of adverse drug effects ([Convertino et al., 2018](#)), and communication sciences analyze media coverage to obtain information on the perception and framing of issues as well as on the rise and fall of topics within News and social media ([Scharl et al., 2017](#); [Weichselbraun et al., 2021](#)).

Computer science focuses on analyzing content by applying knowledge extraction techniques such as entity recognition ([Fu et al., 2021](#)) to automatically identify entities (e.g., persons, organizations, locations, products, etc.) within text documents, entity linking ([Ding et al., 2021](#)) to link these entities to knowledge bases such as Wikidata and DBpedia, and sentiment

37 analysis to automatically assess sentiment polarity (i.e., positive versus negative coverage)
38 and emotions expressed towards these entities (Wang et al., 2020).

39 Most knowledge extraction methods operate on text and, therefore, require an accurate con-
40 version of HTML content which also preserves the spatial alignment between text elements.
41 This is particularly true for methods drawing upon algorithms which directly or indirectly
42 leverage information on the proximity between terms, such as word embeddings (Mikolov et
43 al., 2013; Pennington et al., 2014), language models (Reis et al., 2021), sentiment analysis
44 which often also considers the distance between target and sentiment terms, and automatic
45 keyword and phrase extraction techniques.

46 Despite this need from within the research community, many standard HTML to text con-
47 version techniques are not layout aware, yielding text representations that fail to preserve the
48 text's spatial properties, as illustrated below:

Climate [edit]

Chur has an oceanic climate in spite of its inland position. Summers are warm and

Climate data for Chur (1981-2010)

Month	Jan	Feb	Mar	Apr	May
Average high °C (°F)	4.8 (40.6)	6.4 (43.5)	11.2 (52.2)	15.1 (59.2)	20.0 (68.0)
Daily mean °C (°F)	0.7 (33.3)	1.8 (35.2)	5.9 (42.6)	9.7 (49.5)	14.3 (57.7)
Average low °C (°F)	-2.6 (27.3)	-2.0 (28.4)	1.6 (34.9)	4.6 (40.3)	8.9 (48.0)
Average precipitation mm (inches)	51 (2.0)	47 (1.9)	55 (2.2)	49 (1.9)	71 (2.8)
Average snowfall cm (inches)	34.0 (13.4)	24.7 (9.7)	10.3 (4.1)	1.5 (0.6)	0.4 (0.2)
Average precipitation days (≥ 1.0 mm)	7.3	6.6	8.1	7.5	9.9
Average snowy days (≥ 1.0 cm)	4.8	3.9	2.5	0.4	0.1
Average relative humidity (%)	73	70	65	63	64
Mean monthly sunshine hours	97	112	139	147	169

Source: MeteoSwiss[19]

Climate[edit]

Chur has an oceanic climate in spite of its inland position. During this month there was precipitation for an average of 1

Climate data for Chur (1981-2010)

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Average high °C (°F)	4.8	6.4	11.2	15.1	20.0								
Daily mean °C (°F)	0.7	1.8	5.9	9.7	14.3								
Average low °C (°F)	-2.6	-2.0	1.6	4.6	8.9								
Average precipitation mm (inches)	51	47	55	49	71								
Average snowfall cm (inches)	34.0	24.7	10.3	1.5	0.4								
Average precipitation days (≥ 1.0 mm)	7.3	6.6	8.1	7.5	9.9								
Average snowy days (≥ 1.0 cm)	4.8	3.9	2.5	0.4	0.1								
Average relative humidity (%)	73	70	65	63	64								
Mean monthly sunshine hours	97	112	139	147	169								

Daily mean °C (°F) 0.7

Figure 1: Text representation of a table from Wikipedia computed by Inscriptis (left) and lynx (right). Lynx fails to correctly interpret the table and, therefore, does not properly align the temperature values.

49 Inscriptis is not only able to correctly render such pages but also offers the option to
50 preserve parts of the original HTML document's semantics (e.g., information on headings,
51 emphasized text, tables, etc.) by complementing the extracted text with annotations obtained
52 from the document. Figure 2 provides an example of annotations extracted from a Wikipedia
53 page. These annotations can be useful for

- 54 ■ providing downstream knowledge extraction components with additional information
55 that may be leveraged to improve their respective performance. Text summarization
56 techniques, for instance, can put a stronger emphasis on paragraphs that contain bold
57 and italic text, and sentiment analysis may consider this information in addition to
58 textual clues such as uppercase text.
- 59 ■ assisting manual document annotation processes (e.g., for qualitative analysis or gold
60 standard creation). Inscriptis supports multiple export formats such as XML, anno-
61 tated HTML and the JSONL format that is used by the open source annotation tool
62 [doccano](#)¹. Support for further annotation formats can be easily added by implementing
63 custom annotation post-processors.
- 64 ■ enabling the use of Inscriptis for tasks such as content extraction (i.e., extract task-
65 specific relevant content from a Web page) which rely on information on the HTML
66 document's structure.

¹Please note that doccano currently does not support overlapping annotations and, therefore, cannot import files containing overlapping annotations.

Annotations extracted from the Wikipedia entry for Chur that have been exported to HTML using the `--postprocessor html` command line option.

Politics [edit] [link](#)

Coat of arms [edit] [link](#)

Blazon: Argent, a city gate gules with three merlons, within which a capricorn rampant sable, langued and viriled of the

Administrative divisions [edit] [link](#)

Government [edit] [link](#)

The City Council (**Stadtrat**) constitutes the executive government of the City of Chur and operates as a collegiate author

As of 2017, Chur's City Council is made up of one representative of the FDP ([FDP.The Liberals](#), who is also the mayor), or Stadtrat of Chur[20]

City Councillor	Party	Head of Department (Leitung, since) of	elected since
(Stadtrat/ Stadträtin)			
Urs Marti ^(CC 1)	FDP	Departement 1 (2013)	2012
Tom Leibundgut	FLV	Departement 3 (2013)	2012
Patrik Degiacomi	SP	Departement 2 (2017)	2016

Figure 2: Annotations extracted from the Wikipedia entry for Chur that have been exported to HTML using the `--postprocessor html` command line option.

In conclusion, Inscriptis provides knowledge extraction components with high quality text representations of HTML documents. Since its first public release in March 2016, Inscriptis has been downloaded over 125,000 times from the Python Package Index (PyPI)², has proven its capabilities in national and European research projects, and has been integrated into commercial products such as the [webLyzard Web Intelligence and Visual Analytics Platform](#).

Mentions

The following research projects use Inscriptis within their knowledge extraction pipelines:

- [CareerCoach](#): “Automatic Knowledge Extraction and Recommender Systems for Personalized Re- and Upskilling suggestions” funded by Innosuisse.
- [Job Cockpit](#): “Web analytics, data enrichment and predictive analysis for improved recruitment and career management processes” funded by Innosuisse
- [EPOCH project](#) funded by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility and Technology (BMK) via the ICT of the Future Program.
- [MedMon](#): “Monitoring of Internet Resources for Pharmaceutical Research and Development” funded by Innosuisse.
- [ReTV project](#) funded by the European Union's Horizon 2020 Research and Innovation Programme.

Acknowledgements

Work on Inscriptis has been conducted within the MedMon, Job Cockpit and CareerCoach projects funded by Innosuisse.

References

- Convertino, I., Ferraro, S., Blandizzi, C., & Tuccori, M. (2018). The usefulness of listening social media for pharmacovigilance purposes: A systematic review. *Expert Opinion on Drug Safety*, 17(11), 1081–1093. <https://doi.org/10.1080/14740338.2018.1531847>

²Source: <https://pepy.tech/project/inscriptis>

- 91 Ding, W., Chaudhri, V. K., Chittar, N., & Konakanchi, K. (2021). JEL: Applying End-to-
92 End Neural Entity Linking in JPMorgan Chase. *Proceedings of the AAAI Conference*
93 *on Artificial Intelligence*, 35(17), 15301–15308. [https://ojs.aaai.org/index.php/AAAI/](https://ojs.aaai.org/index.php/AAAI/article/view/17796)
94 [article/view/17796](https://ojs.aaai.org/index.php/AAAI/article/view/17796)
- 95 Fu, J., Huang, X., & Liu, P. (2021). SpanNER: Named Entity Re-/Recognition as Span
96 Prediction. *arXiv:2106.00641 [cs]*. <http://arxiv.org/abs/2106.00641>
- 97 Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and
98 public mood on stock movements. *Information Sciences*, 278, 826–840. [https://doi.org/](https://doi.org/10.1016/j.ins.2014.03.096)
99 [10.1016/j.ins.2014.03.096](https://doi.org/10.1016/j.ins.2014.03.096)
- 100 Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed
101 Representations of Words and Phrases and their Compositionality. *Advances in*
102 *Neural Information Processing Systems 26: 27th Annual Conference on Neural In-*
103 *formation Processing Systems 2013. Proceedings of a Meeting Held December 5-8,*
104 *2013, Lake Tahoe, Nevada, United States*, 3111–3119. [http://papers.nips.cc/paper/](http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality)
105 [5021-distributed-representations-of-words-and-phrases-and-their-compositionality](http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality)
- 106 Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Represent-
107 ation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*
108 *Processing (EMNLP)*, 1532–1543. <https://doi.org/10.3115/v1/D14-1162>
- 109 Reis, E. S. D., Costa, C. A. D., Silveira, D. E. D., Bavaresco, R. S., Righi, R. D. R., Barbosa,
110 J. L. V., Antunes, R. S., Gomes, M. M., & Federizzi, G. (2021). Transformers aftermath:
111 Current research and rising trends. *Communications of the ACM*, 64(4), 154–163. [https:](https://doi.org/10.1145/3430937)
112 [//doi.org/10.1145/3430937](https://doi.org/10.1145/3430937)
- 113 Scharl, A., Herring, D., Rafelsberger, W., Hubmann-Haidvogel, A., Kamolov, R., Fischl, D.,
114 Föls, M., & Weichselbraun, A. (2017). Semantic Systems and Visual Tools to Support
115 Environmental Communication. *IEEE Systems Journal*, 11(2), 762–771. [https://doi.org/](https://doi.org/10.1109/JSYST.2015.2466439)
116 [10.1109/JSYST.2015.2466439](https://doi.org/10.1109/JSYST.2015.2466439)
- 117 Wang, Z., Ho, S.-B., & Cambria, E. (2020). A review of emotion sensing: Categorization
118 models and algorithms. *Multimedia Tools and Applications*. [https://doi.org/10.1007/](https://doi.org/10.1007/s11042-019-08328-z)
119 [s11042-019-08328-z](https://doi.org/10.1007/s11042-019-08328-z)
- 120 Weichselbraun, A., Kuntschik, P., Fancolino, V., Saner, M., & Wyss, V. (2021). Adapting
121 Data-Driven Research to the Fields of Social Sciences and the Humanities. *Future Internet*,
122 13(3). <https://doi.org/10.3390/fi13030059>