

¹ gediDB: A toolbox for processing and providing Global Ecosystem Dynamics Investigation (GEDI) L2A-B and L4A-C data

⁴ **Simon Besnard**  ¹, **Felix Dombrowski**², and **Amelia Holcomb**  ³

⁵ 1 GFZ Helmholtz Centre Potsdam, Potsdam, Germany ² University of Potsdam, Potsdam, Germany ³
⁶ Department of Computer Science, University of Cambridge, Cambridge, UK ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

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

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a

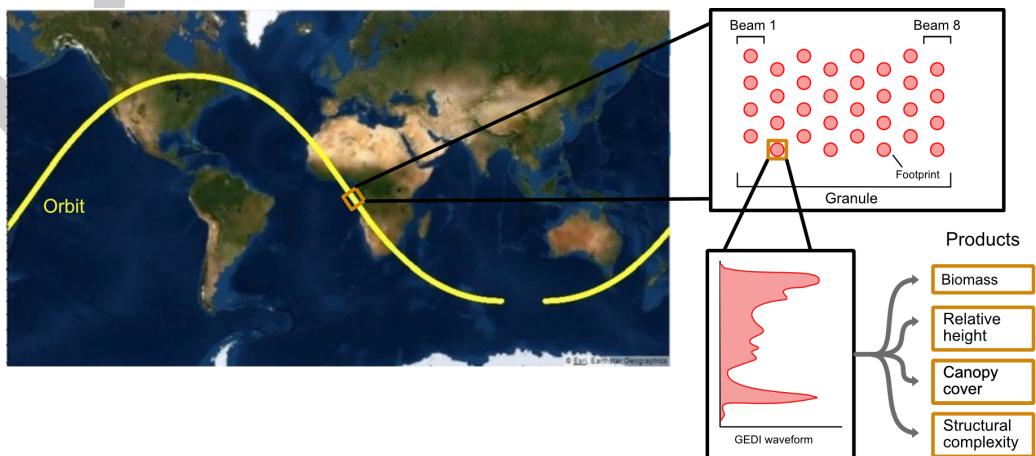
Creative Commons Attribution 4.0 International License ([CC BY 4.0](#))

⁷ Abstract

⁸ The Global Ecosystem Dynamics Investigation (GEDI) mission provides spaceborne LiDAR
⁹ observations that are essential for characterising Earth's forest structure and carbon dynamics.
¹⁰ However, GEDI datasets are distributed as complex HDF5 granules, which pose significant
¹¹ challenges for efficient, large-scale data processing and analysis. To overcome these hurdles,
¹² we developed gediDB, an open-source Python standardized Application Programming Interface
¹³ (API) that streamlines both the processing and querying of GEDI Level 2A–B and Level 4A–C
¹⁴ datasets. Built on the optimised multidimensional array database TileDB, gediDB enables
¹⁵ operational-scale processing, rapid spatial and temporal queries, and reproducible LiDAR-based
¹⁶ analyses of forest biomass, carbon stocks, and structural change.

Statement of Need

High-volume LiDAR datasets from the Global Ecosystem Dynamics Investigation (GEDI) mission ([R. Dubayah et al., 2020](#)) (Fig. 1) have become a key resource for quantifying forest dynamics, estimating biomass, and analysing carbon cycling. The open availability of GEDI's spaceborne LiDAR data has created unprecedented opportunities to extend forest structural analyses from local or regional case studies to near-global scales. However, despite the richness of information contained in GEDI datasets, their practical usability remains challenging due to the complexity of raw HDF5 granules, a lack of scalable infrastructure for efficient data retrieval, and insufficient standardized tools for large-scale spatial and temporal subsetting.

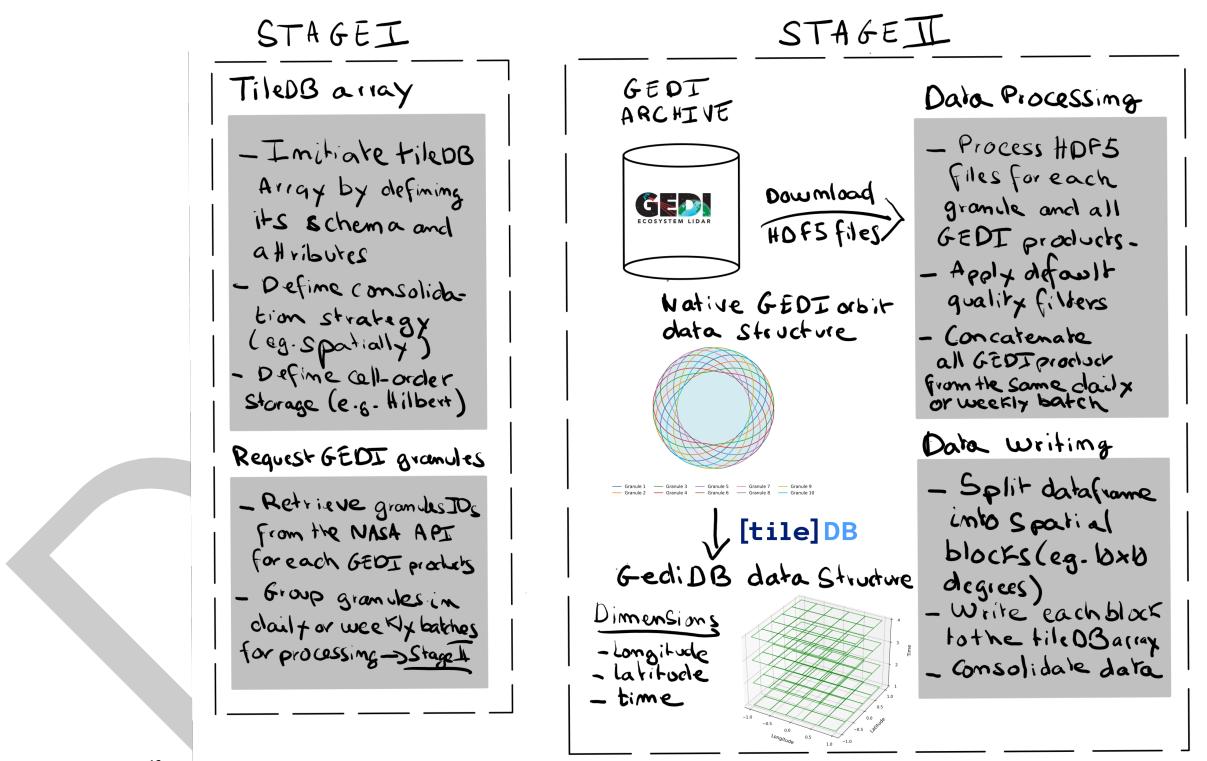


²⁶ ²⁷ Fig. 1: A schematic representation of the GEDI data structure. Credits: Amelia Holcomb's

28 *PhD dissertation*

29 Existing software tools for GEDI data analysis, such as the GEDI Subsetter provided by
 30 NASA's Multi-Mission Algorithm and Analysis Platform (MAAP) [GEDI Subsetter; Daniels
 31 et al. (2025)], primarily address small to moderate-scale data extraction scenarios. While
 32 suitable for interactive or limited spatial extents, these tools often struggle to efficiently support
 33 large-scale workflows, leading to computational bottlenecks and decreased efficiency when
 34 applied to extensive spatial and temporal analyses.

35 gediDB addresses these limitations by offering a robust and scalable framework that unifies
 36 access to GEDI Level 2A (R. Dubayah, Hofton, et al., 2021), Level 2B (R. Dubayah, Tang, et
 37 al., 2021), Level 4A (R. O. Dubayah et al., 2022) and Level 4C (De Conto et al., 2024) data
 38 via an Python (Van Rossum & Drake, 2009) standardized Application Programming Interface
 39 (API). Built on the TileDB storage engine, gediDB supports rapid querying of multidimensional
 40 arrays, allowing users to efficiently extract large data subsets by spatial extent, temporal
 41 range, and variable selection. It integrates seamlessly with Python's geospatial data ecosystem,
 42 including libraries such as xarray (Hoyer & Hamman, 2017) and geopandas (Jordahl et al.,
 43 2020), and integrates into reproducible workflows that can scale to high-performance computing
 44 environments and cloud platforms. By leveraging TileDB's advanced spatial indexing, gediDB
 45 substantially simplifies the processing and querying of GEDI data (see Fig. 2).



48 **Core functionalities**

49 Extensive documentation and user tutorials for gediDB are available at <https://gedidb.readthedocs.io>. These resources provide comprehensive setup instructions, configuration
 50 guidance, and workflow examples. Users have immediate access to a globally processed GEDI
 51 dataset, eliminating the need for local downloads, as detailed in the [database documentation](#).

53 **Data processing framework**

54 The gediDB package is structured around two core modules designed to streamline GEDI data
55 processing and retrieval:

- 56 **GEDIProcessor**: Systematically ingests raw GEDI granules and transforms them into
57 structured TileDB arrays (Fig. 3). Key steps include data filtering, standardisation, and
58 efficient spatio-temporal chunking to ensure optimal query performance.
- 59 **GEDIProvider**: Enables rapid, flexible access to GEDI data using spatial bounding
60 boxes, temporal filters, and user-selected variables. Results are provided in formats fully
61 compatible with Python's geospatial libraries such as `xarray` and `pandas` ([Reback et al.,
62 2020](#)).

63 **Configurable and reproducible workflows**

64 gediDB uses customisable configuration files to define TileDB schemas and data retrieval
65 parameters. This facilitates reproducibility and adaptability across diverse research scenarios
66 and computing environments.

67 **Robust data downloading**

68 The API interfaces directly with NASA's Common Metadata Repository (CMR) to facilitate
69 reliable data acquisition. It incorporates comprehensive retry logic and robust error handling to
70 mitigate issues related to network interruptions or data inconsistencies.

71 **High-performance data storage**

72 GEDI data is efficiently stored using structured TileDB sparse arrays optimised for rapid
73 spatial and temporal queries. The array structure is specifically designed to handle large-scale,
74 multi-dimensional data seamlessly (Fig. 3).

75 **Parallel processing capabilities**

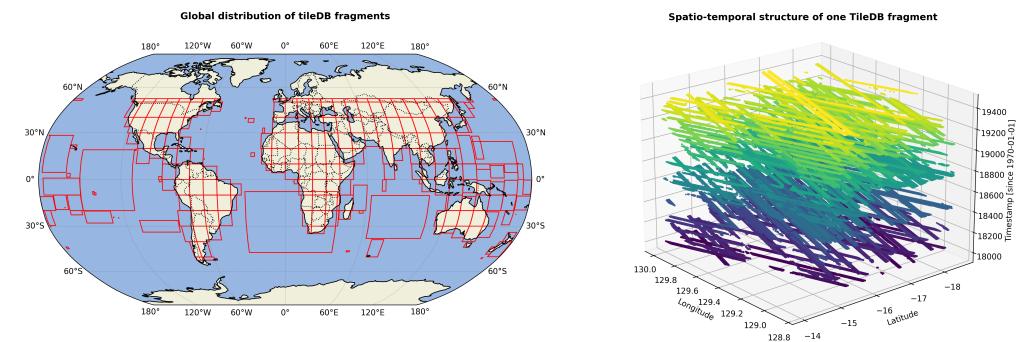
76 Parallelised operations are fully supported, including data downloading, processing, and
77 storage. gediDB leverages libraries such as Dask ([Rocklin, 2015](#)) and Python's built-in
78 `concurrent.futures` to maximise performance on large-scale workflows and high-performance
79 computing infrastructures.

80 **Advanced querying functionality**

81 gediDB offers sophisticated querying methods, enabling spatial and temporal subsetting through
82 bounding-box, time-range, and nearest-neighbour queries. The system efficiently handles both
83 scalar and profile-type variables.

84 **Comprehensive metadata management**

85 Rich metadata is systematically captured and managed, including data provenance, variable
86 units, descriptions, and product version details. Metadata is embedded directly within the
87 TileDB structure to facilitate clear data documentation and reproducibility.



88

89 Fig. 3: Illustration of the global GEDI data storage schema using TileDB arrays.

90 Performance benchmarks

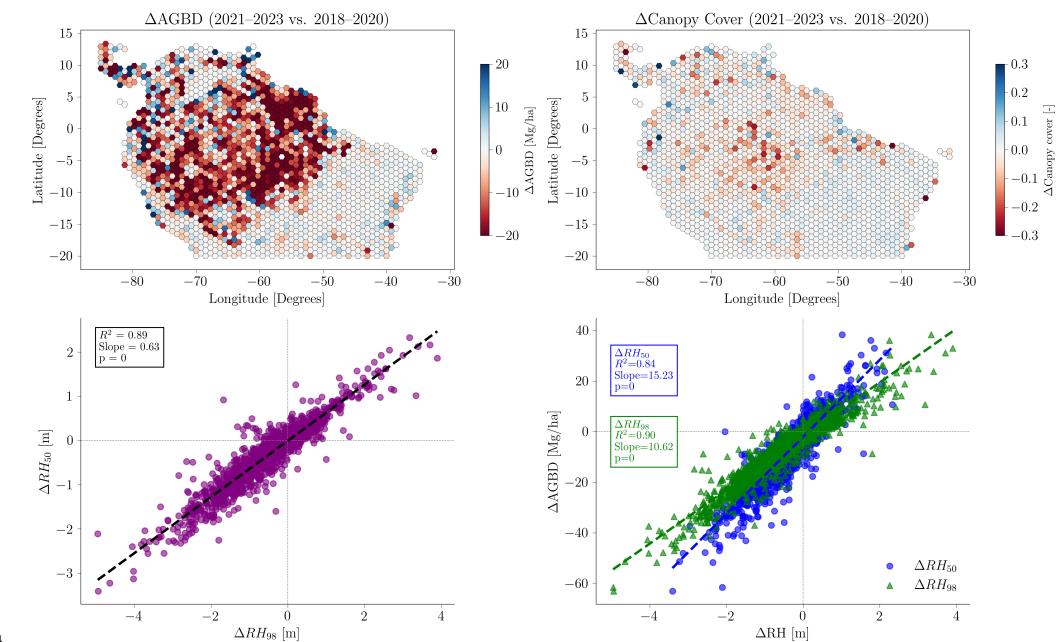
91 The efficiency of gediDB was evaluated under realistic research scenarios. The table below
92 summarises query times across different spatial and temporal extents:

Scenario	Spatial extent	Time range	Variables queried	Query time (seconds)
Local-scale query	$1^\circ \times 1^\circ$ bounding box	1 month	relative height metrics, canopy cover	1.8
Regional-scale query	$10^\circ \times 10^\circ$ bounding box	6 months	relative height metrics, biomass, plant area index	17.9
Continental-scale query	Amazon Basin	1 year	canopy cover, biomass	28.9

93 Benchmarks were conducted on a Linux server equipped with dual Intel® Xeon® E5-2643
94 v4 CPUs (12 physical cores, 24 threads total), 503 GB RAM, and a combination of NVMe
95 SSD (240 GB) and HDD storage (16.4 TB total). Queries were executed from NVMe-backed
96 storage to ensure high I/O performance. Compared to workflows based on direct HDF5 access,
97 gediDB provides a significant speedup and streamlined user experience.

98 Example use cases

99 An illustrative use case involved the analysis of aboveground biomass and canopy cover
100 dynamics across the Amazon Basin (Fig. 4). Leveraging gediDB, variables representing
101 aboveground biomass, canopy cover, and vertical canopy structure (i.e., relative height (RH)
102 metrics) were efficiently extracted over large spatial extents and multiple years. The data
103 were aggregated within a $1^\circ \times 1^\circ$ hexagonal grid framework, enabling spatiotemporal analysis
104 of forest structure change. Integration with Python's geospatial libraries, such as geopandas
105 and xarray, allowed for a fully reproducible workflow from data extraction to visualization.
106 To explore structural drivers of biomass change, a scatter plot analysis compared changes in
107 upper and lower canopy height metrics ($\Delta RH98$ and $\Delta RH50$) with $\Delta AGBD$. This illustrated
108 the nuanced relationships between vertical canopy dynamics and biomass accumulation or loss
109 across forested ecosystems. The greater slope observed between changes in median canopy
110 height ($\Delta RH50$) and aboveground biomass density ($\Delta AGBD$), compared to changes in upper
111 canopy height ($\Delta RH98$) (lower right panel), indicates that biomass dynamics are more closely
112 linked to widespread structural adjustments across the lower and mid-canopy layers rather than
113 to changes restricted to the tallest forest emergents.



114
115 *Fig. 4: Visualisation of changes in aboveground biomass density (AGBD) (top left panel) and*
116 *canopy cover (top right panel) between 2018–2020 and 2021–2023, aggregated to a $1^\circ \times 1^\circ$*
117 *hexagonal grid over the Amazon Basin. The bottom left panel shows the relationship between*
118 *changes in ΔRH_{50} and ΔRH_{98} , with each point representing a hexagon. The bottom right*
119 *panel shows the relationship between changes in canopy height metrics (ΔRH_{50} and ΔRH_{98})*
120 *and $\Delta AGBD$, with each point representing a hexagon. This highlights how vertical canopy*
121 *structure dynamics relate to biomass change across the region.*

122 Beyond regional change assessments, gediDB supports advanced analyses such as biome-
123 level comparisons of forest structural profiles, precise retrieval of GEDI data near field plots
124 for calibration and validation, and the production of spatially gridded datasets at diverse
125 resolutions. Recent studies have leveraged GEDI data to map canopy height across global
126 forested ecosystems (Pauls et al., 2024), to assess forest disturbances (Holcomb et al., 2024),
127 and to characterise forest degradation dynamics (Bourgoin et al., 2024). These applications
128 demonstrate the potential of GEDI data to inform ecological monitoring and policy development.
129 By streamlining data access, subsetting, and integration into scalable workflows, gediDB can
130 significantly enhance the efficiency and reproducibility of such large-scale analyses—supporting
131 efforts like global canopy height mapping, disturbance detection, and forest degradation
132 monitoring.

133 Community impact and future development

134 gediDB fosters an open and collaborative research environment by actively encouraging
135 community-driven development through its [GitHub repository](#). Its open-source nature promotes
136 transparency, reproducibility, and long-term accessibility, benefiting a wide range of scientific
137 applications in LiDAR research and environmental analysis.

138 Planned future developments for gediDB include:

- 139 ■ Maintaining compatibility with new GEDI data releases and product updates
- 140 ■ Enhancing performance and flexibility in querying profile variables
- 141 ■ Supporting direct reading of native HDF5 files from AWS S3 buckets
- 142 ■ Expanding tutorials and documentation to reach a broader user base
- 143 ■ Improving the testing framework to ensure greater reliability and maintainability

144 Feedback, feature requests, and code contributions from users and developers are warmly

¹⁴⁵ welcomed. Through sustained community engagement, gediDB aims to continually evolve as a
¹⁴⁶ robust and adaptable tool for forest remote sensing and ecosystem monitoring.

¹⁴⁷ Conclusion

¹⁴⁸ gediDB substantially significantly improves the practical usability of GEDI LiDAR datasets by
¹⁴⁹ addressing key challenges related to data complexity, scalability, and reproducibility. Leveraging
¹⁵⁰ TileDB's optimised multidimensional array storage, it enables efficient data management, fast
¹⁵¹ querying, and seamless integration into diverse geospatial analysis workflows. This empowers
¹⁵² the systematic exploration of forest dynamics and carbon cycling at unprecedented spatial
¹⁵³ and temporal scales. Through its open-source, community-driven design, gediDB fosters
¹⁵⁴ collaborative progress in remote sensing, ecology, and environmental science, supporting the
¹⁵⁵ evolving needs of the Earth observation community.

¹⁵⁶ Acknowledgements

¹⁵⁷ The development of gediDB was supported by the European Union through the FORWARDS
¹⁵⁸ and OpenEarthMonitor projects. We would also like to acknowledge the R2D2 Workshop
¹⁵⁹ (March 2024, Potsdam) for providing the opportunity to meet and discuss GEDI data processing.

¹⁶⁰ References

- ¹⁶¹ Bourgoin, C., Ceccherini, G., Girardello, M., Vancutsem, C., Avitabile, V., Beck, P., Beuchle,
¹⁶² R., Blanc, L., Duveiller, G., Migliavacca, M., & others. (2024). Human degradation of
¹⁶³ tropical moist forests is greater than previously estimated. *Nature*, 631(8021), 570–576.
- ¹⁶⁴ Daniels, C., French, J., Adhikari, S., Bhusal, A., Mandel, A. I., & Kirkland, S. (2025). MAAP-
¹⁶⁵ project/gedi-subsetter: 0.10.0 (Version 0.10.0). Zenodo. <https://doi.org/10.5281/zenodo.15122227>
- ¹⁶⁶ De Conto, T., Armston, J., & Dubayah, R. O. (2024). *GEDI L4C Footprint Level Waveform
Structural Complexity Index, Version 2*. ORNL DAAC, Oak Ridge, Tennessee, USA.
<https://doi.org/10.3334/ORNLDAAAC/2338>
- ¹⁶⁷ Dubayah, R. O., Armston, J., Kellner, J. R., Duncanson, L., Healey, S. P., Patterson, P. L.,
¹⁶⁸ Hancock, S., Tang, H., Bruening, J. M., Hofton, M. A., Blair, J. B., & Luthcke, S. B.
¹⁶⁹ (2022). *GEDI L4A Footprint Level Aboveground Biomass Density, Version 2.1*. ORNL
DAAC, Oak Ridge, Tennessee, USA. <https://doi.org/10.3334/ORNLDAAAC/2056>
- ¹⁷⁰ Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M.,
¹⁷¹ Hurt, G., Kellner, J., Luthcke, S., & others. (2020). The global ecosystem dynamics
¹⁷² investigation: High-resolution laser ranging of the earth's forests and topography. *Science
of Remote Sensing*, 1, 100002.
- ¹⁷³ Dubayah, R., Hofton, M., Blair, J., Armston, J., Tang, H., & Luthcke, S. (2021). *GEDI L2A
Elevation and Height Metrics Data Global Footprint Level V002*. NASA EOSDIS Land
Processes Distributed Active Archive Center. https://doi.org/10.5067/GEDI/GEDI02_A_002
- ¹⁷⁴ Dubayah, R., Tang, H., Armston, J., Luthcke, S., Hofton, M., & Blair, J. (2021). *GEDI
L2B Canopy Cover and Vertical Profile Metrics Data Global Footprint Level V002*. NASA
EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/GEDI/GEDI02_B_002
- ¹⁷⁵ Holcomb, A., Burns, P., Keshav, S., & Coomes, D. A. (2024). Repeat GEDI footprints measure
¹⁷⁶ the effects of tropical forest disturbances. *Remote Sensing of Environment*, 308, 114174.

- 188 <https://doi.org/https://doi.org/10.1016/j.rse.2024.114174>
- 189 Hoyer, S., & Hamman, J. J. (2017). Xarray: N-d labeled arrays and datasets in python.
190 *Journal of Open Research Software*, 5(1), 10. <https://doi.org/10.5334/jors.148>
- 191 Jordahl, K., Bossche, J. V. den, Fleischmann, M., Wasserman, J., McBride, J., Gerard,
192 J., Tratner, J., Perry, M., Badaracco, A. G., Farmer, C., Hjelle, G. A., Snow, A. D.,
193 Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., maxalbert, Bilogur,
194 A., ... Leblanc, F. (2020). *Geopandas/geopandas: v0.8.1* (Version v0.8.1). Zenodo.
195 <https://doi.org/10.5281/zenodo.3946761>
- 196 Pauls, J., Zimmer, M., Kelly, U. M., Schwartz, M., Saatchi, S., Ciais, P., Pokutta, S., Brandt,
197 M., & Gieseke, F. (2024). *Estimating canopy height at scale*. <https://arxiv.org/abs/2406.01076>
- 199 Reback, J., McKinney, W., jbrockmendel, Van den Bossche, J., Augspurger, T., Cloud, P.,
200 Hawkins, S., Gfyoung, Sinhrks, Klein, A., Roeschke, M., & Tratner, W. (2020). Pandas-
201 dev/pandas: pandas. Zenodo. <https://doi.org/10.5281/zenodo.3509134>
- 202 Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling.
203 *Proceedings of the 14th Python in Science Conference*, 130–136. <https://doi.org/10.25080/majora-7b98e3ed-013>
- 205 Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.
206 ISBN: 1441412697

DRAFT