OpenITI release 2022.2.7

The KITAB team has released a new version (2022.2.7) of the [OpenITI](https://github.com/OpenITI) corpus at [Zenodo](https://zenodo.org/record/7687795). The release is open access. It is our seventh release (second release in 2022). You can also access the release through our [GitHub repository](https://github.com/OpenITI/RELEASE).

This release features 6,858 unique books written by 2,854 authors. For some of these books, multiple versions are in the corpus (representing different print editions and/or digitizations); including all these versions, the corpus now comprises 11,296 text files (containing a total of 2,262,096,158 words).

Major additions to the corpus:

* The KITAB team have OCRed 75 new texts using the improved [OCR models for Arabic script](https://github.com/OpenITI/AOCP_print_models) developed by the [OpenITI-AOCP project](https://openiti.org/projects/OpenITI%20AOCP%20Phase%20One.html). For OCR’ing these texts, the team uses the command-line OCR software [Kraken](https://kraken.re/master/index.html) (these texts have an ID that starts with `Kraken`, followed by a date), sometimes through the graphical user interface of the open-source OCR platform [eScriptorium](https://escriptorium.openiti.org/) (these texts have an ID that starts with `EScr`). Most of these OCR’ed texts have not been post-corrected and still contain a significant number of transcription errors (in the metadata files accompanying these texts, you’ll find an “UNCORRECTED\_OCR” tag).   
  ***Note: These texts should be used only for types of analysis that are not affected by small transcription errors*** (for example, our text reuse algorithm [passim](https://github.com/dasmiq/passim) has no problem with small transcription errors).
* The [Library of Arabic Literature](https://www.libraryofarabicliterature.org/) has generously provided 13 texts (IDs start with `LAL`); more texts from this wonderful collection will be added in future corpus releases. Some of these texts are entirely new additions to the corpus, such as [al-Quḍāʿī’s *Kitāb al-Shihāb*](https://github.com/OpenITI/0475AH/tree/master/data/0454Qudaci/0454Qudaci.Shihab), Ḥannā Diyāb’s [*Book of Travels*](https://github.com/OpenITI/1200AH/tree/master/data/1200HannaDiyab/1200HannaDiyab.Siyaha) and the anonymous story collection [*A Hundred and One Nights*](https://github.com/OpenITI/1000AH/tree/master/data/1000Anonymous/1000Anonymous.MiatLaylaWaLayla).

A full list of the new additions can be found in the `ids.csv` file in the [release notes](https://zenodo.org/record/7687795/files/release_notes.zip?download=1) accompanying the corpus release.

In addition to the new texts, annotation work has been carried out: structural annotation (section headers) was added to 50 texts, and 62 previously annotated texts were vetted by a second annotator. For a full list of the new annotation work, see the `annotation\_update.csv` file in the [release notes](https://zenodo.org/record/7687795/files/release_notes.zip?download=1) accompanying the corpus release. Currently, the corpus contains 552 books that have been annotated and vetted (extension: `.mARkdown`), and 647 books that have gone through one round of annotation (extension: `.completed`).

Finally, some URIs were updated to correct mistakes. For an overview of these URI changes as well as the statistics on the corpus, please see the release notes and the corresponding csv files in the [release notes](https://zenodo.org/record/7687795/files/release_notes.zip?download=1) and the [GitHub repository](https://github.com/OpenITI/RELEASE/tree/master/release_notes).

To cite this version please include the following information (the bibliographical export is available at the [publication page](https://zenodo.org/record/7687795)):

Nigst, Lorenz, Romanov, Maxim, Savant, Sarah Bowen, Seydi, Masoumeh, & Verkinderen, Peter. (2023). OpenITI: a Machine-Readable Corpus of Islamicate Texts (2022.2.7) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7687795