OpenITI, ver. 2022.1.6
—Release Notes—

Corpus Metadata

The current release metadata is available in the `OpenITI_metadata_2022-1-6.csv` and `OpenITI_metadata_2022-1-6_merged.csv` (merged[1] version) files.

Folder Structure

- `data`: main data folder with *Author* > *Book* > *Versions* structure;
- `metadata`
    - `OpenITI_metadata_2022-1-6.csv`: metadata file, with a row for each text in the corpus, including a row for each part of the multi-part books in the corpus;
    - `OpenITI_metadata_2022-1-6_merged.csv`: metadata file, with one row for each multi-part book in the corpus;
- `release_notes`
    - `OpenITI_release-notes_2022-1-6.pdf`: these release notes;
    - `release-notes_files_2022-1-6.zip`: csv files including the provided list of new changes in the current release (see the description of each file in this release notes).

[1] The corpus contains a number of texts that are too big for GitHub and had to be split into multiple files (currently only two versions of the book *Biḥār al-anwār* (`1111Majlisi.BiharAnwar`)). The metadata file contains statistics on each part of this split text. We also provide a separate metadata file in which the statistics for the separate parts of those books that have been merged. Since the merged metadata for these split files does not refer to an existing file, the `local_path` field for these virtual texts will be "NA".

Corpus Statistics

Category	Stats
Number of unique titles	6,785
Number of authors	2,843
Number of book titles (all versions/editions)	11,195
Number of uncorrected OCRed texts[2]	246

Length of texts (all books)

	Number of words	Number of pages (300 w/p)
Total	2,251,035,992	7,503,454
Min.	47	1
1st Qu.	8,502	29
Median	41,022	137
Mean	201,076	671
3rd Qu.	149,797	500
Max.	11,912,693	39,709

Length of texts (unique books)

	Number of words	Number of pages (300 w/p)
Total	1,080,907,744	3,603,026
Min.	48	1
1st Qu.	7,638	26
Median	33,211	111
Mean	159,309	532

[2] The uncorrected OCRed texts are specified by the "UNCORRECTED_OCR" tag in the "tags" column of the metadata files.

3rd Qu.	121,789	406
Max.	11,912,693	39,709

Annotation statistics

Number of texts with extension `.mARkdown`	479
Number of texts with extension `.completed`	646
Number of texts with extension `.inProgress`	10

Book Ids

The list of the new book ids in this version is available in the `ids.csv` file. It includes the newly added book ids and modified ids. The URI includes the information of the new book (i.e., date, author, and book title).

Modified URIs

List of modified URIs in the current release is available in `modified_uris.csv`. Changes typically affect such fields as year, author, and title. These changes are applied to the entire metadata (book IDs remain unchanged).

Annotation Update

The list of texts that have been structurally annotated or where the annotation has changed (can be tracked by the file extensions) since our previous release (version 2021.2.5) is provided in `annotation_update.csv`. This file shows URIs of texts together with their current extension, which is a part of the `local_path` in the metadata file.

For more information on the OpenITI mARkdown and the extensions please see here.

Credits

Current contributors (*alphabetically*):

- Sohail Merchant (*metadata app*)
- Lorenz Nigst (*corpus management; structural annotation*)
- Maxim Romanov (*OpenITI co-PI; EIS1600 Project PI; conceptual development; mARkdown*)
- Sarah Bowen Savant (*OpenITI co-PI; KITAB Project PI*)
- Masoumeh Seydi (*technical development*)
- Peter Verkinderen (*technical development; preparing new texts for the corpus; structural annotation*)

- Mathew Barber (*structural annotation*)
- Gowaart Van Den Bossche (*structural annotation*)
- Hamidreza Hakimi (*structural annotation*)
- Simon Loynes (*structural annotation*)
- Aslisho Qurboniev (*structural annotation*)

Past contributors:

- Maroussia Bednarkiewicz (*structural annotation*)
- Christoph Gümmer (*structural annotation*)
- Jonas Köpsel (*structural annotation*)
- Cornelis [Eric] Van Lit (*structural annotation*)
- Cornelia Neubauer (*structural annotation*)
- Leonie Nückell (*structural annotation*)
- Fatemeh Shams (*structural annotation*)