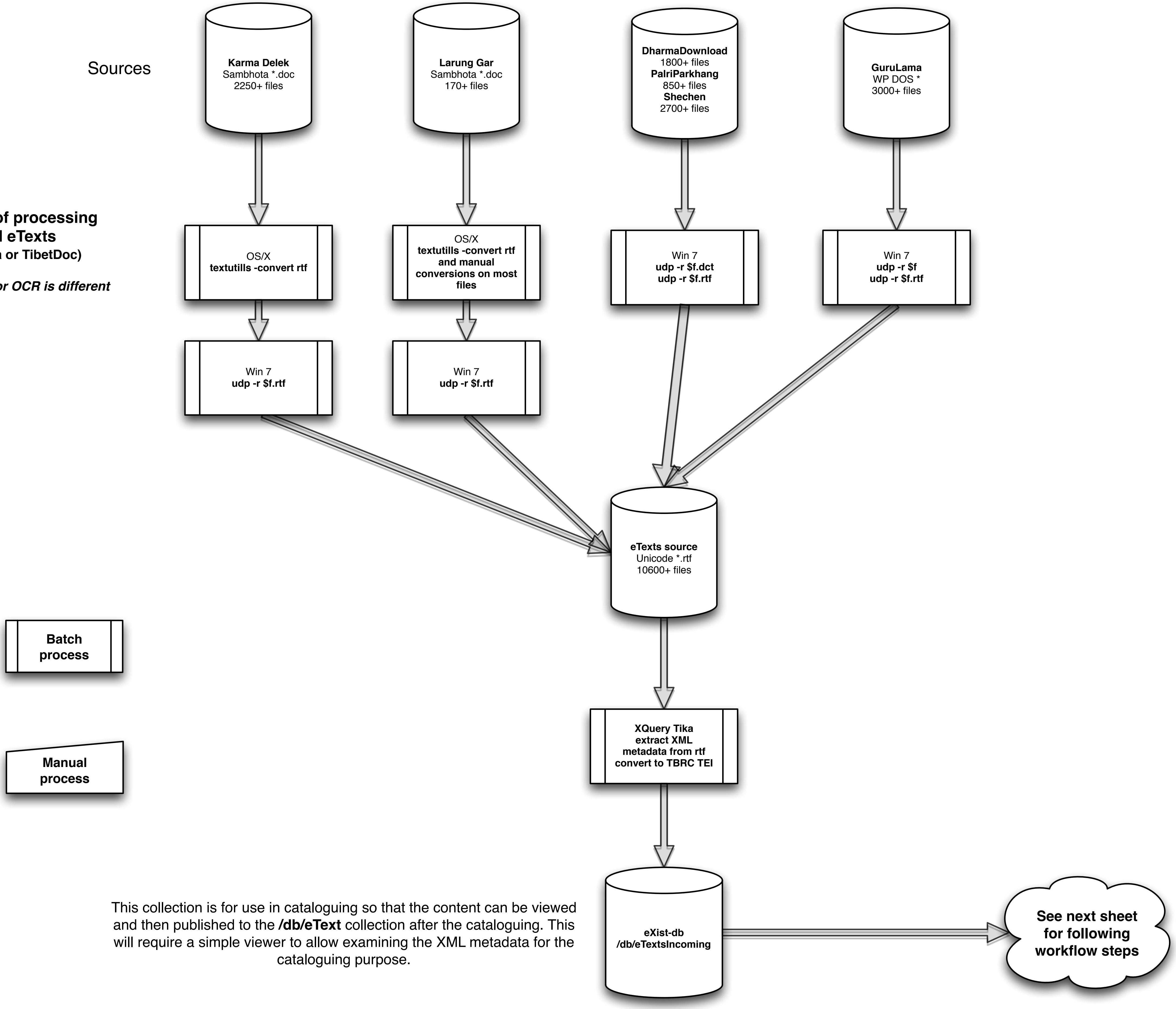
Overview of processing typed eTexts
(Sambhota or TibetDoc)

the workflow for OCR is different



Sources

Karma Delek
Sambhota *.doc
2250+ files

Larung Gar
Sambhota *.doc
170+ files

DharmaDownload
1800+ files
PalriParkhang
850+ files
Shechen
2700+ files

GuruLama
WP DOS *
3000+ files

OS/X
textutills -convert rtf

OS/X
textutills -convert rtf and manual conversions on most files

Win 7
udp -r $f.dct
udp -r $f.rtf

Win 7
udp -r $f
udp -r $f.rtf

Win 7
udp -r $f.rtf

Win 7
udp -r $f.rtf

eTexts source
Unicode *.rtf
10600+ files

XQuery Tika extract XML metadata from rtf convert to TBRC TEI

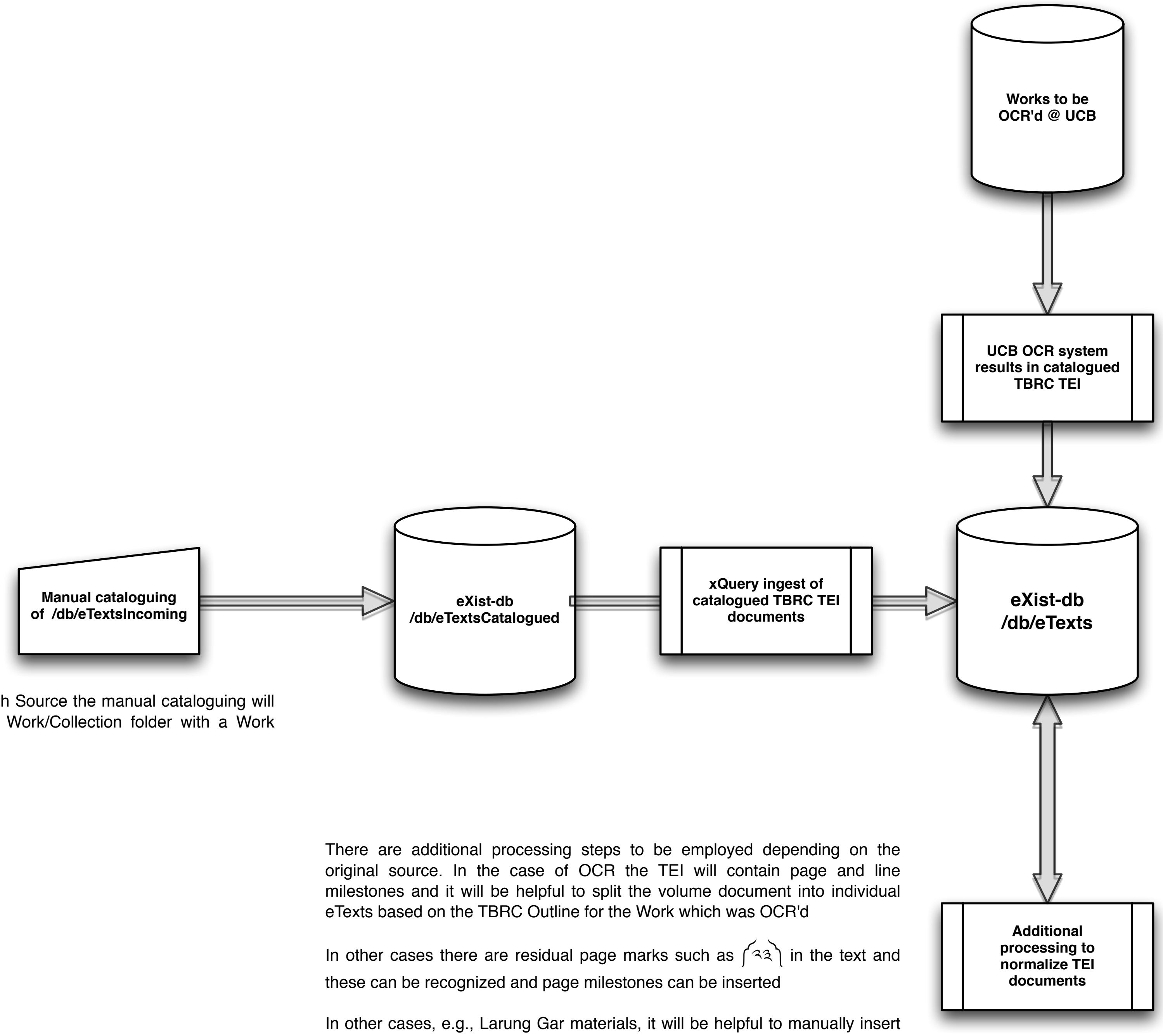
eXist-db
/db/eTextsIncoming

See next sheet for following workflow steps

Batch process

Manual process

This collection is for use in cataloguing so that the content can be viewed and then published to the **/db/eText** collection after the cataloguing. This will require a simple viewer to allow examining the XML metadata for the cataloguing purpose.


Works to be OCR'd @ UCB
UCB OCR system results in catalogued TBRC TEI
Manual cataloguing of /db/eTextsIncoming
eXist-db /db/eTextsCatalogued
xQuery ingest of catalogued TBRC TEI documents
eXist-db /db/eTexts
Additional processing to normalize TEI documents

This collection is the target to be searched to provide match highlights with context for users of the website. The metadata is organized as paragraphs, chunks or pages depending on how the input was prepared. Each eText is linked to the input rtf and Sambhota .doc; TibetDoc .dct file or other source document.

The original files and Unicode rtf files derived for use as input during initial TiKa extraction are stored in the filesystem with a directory layout that mirrors the scans in the image archive (see the following sheet.)

For each Source the manual cataloguing will tag the Work/Collection folder with a Work RID or

There are additional processing steps to be employed depending on the original source. In the case of OCR the TEI will contain page and line milestones and it will be helpful to split the volume document into individual eTexts based on the TBRC Outline for the Work which was OCR'd

In other cases there are residual page marks such as ༢༣ in the text and these can be recognized and page milestones can be inserted

In other cases, e.g., Larung Gar materials, it will be helpful to manually insert text break milestones and then run a script to split the volume into separate eTexts.

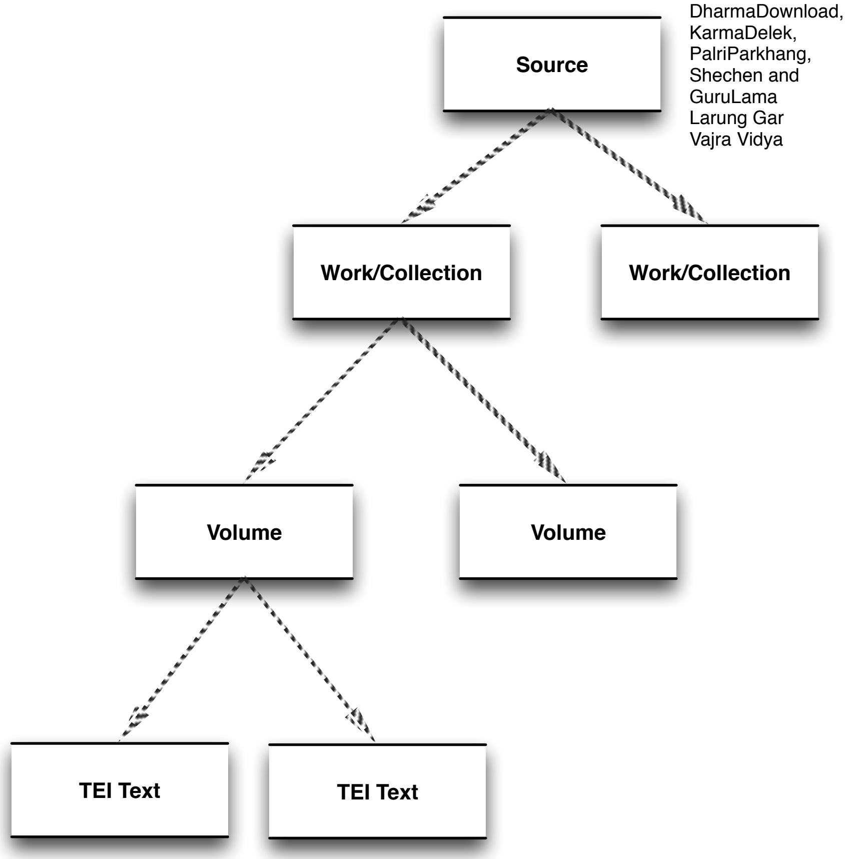
Organization of eText TEI documents in eXist database

This is the structure of the **/db/eTextsIncoming**, **/db/eTextsCatalogued** and **/db/eTexts**.

For *incoming* the naming of Work, Volume and Text will be based on the original names in the materials as TBRC receives them.

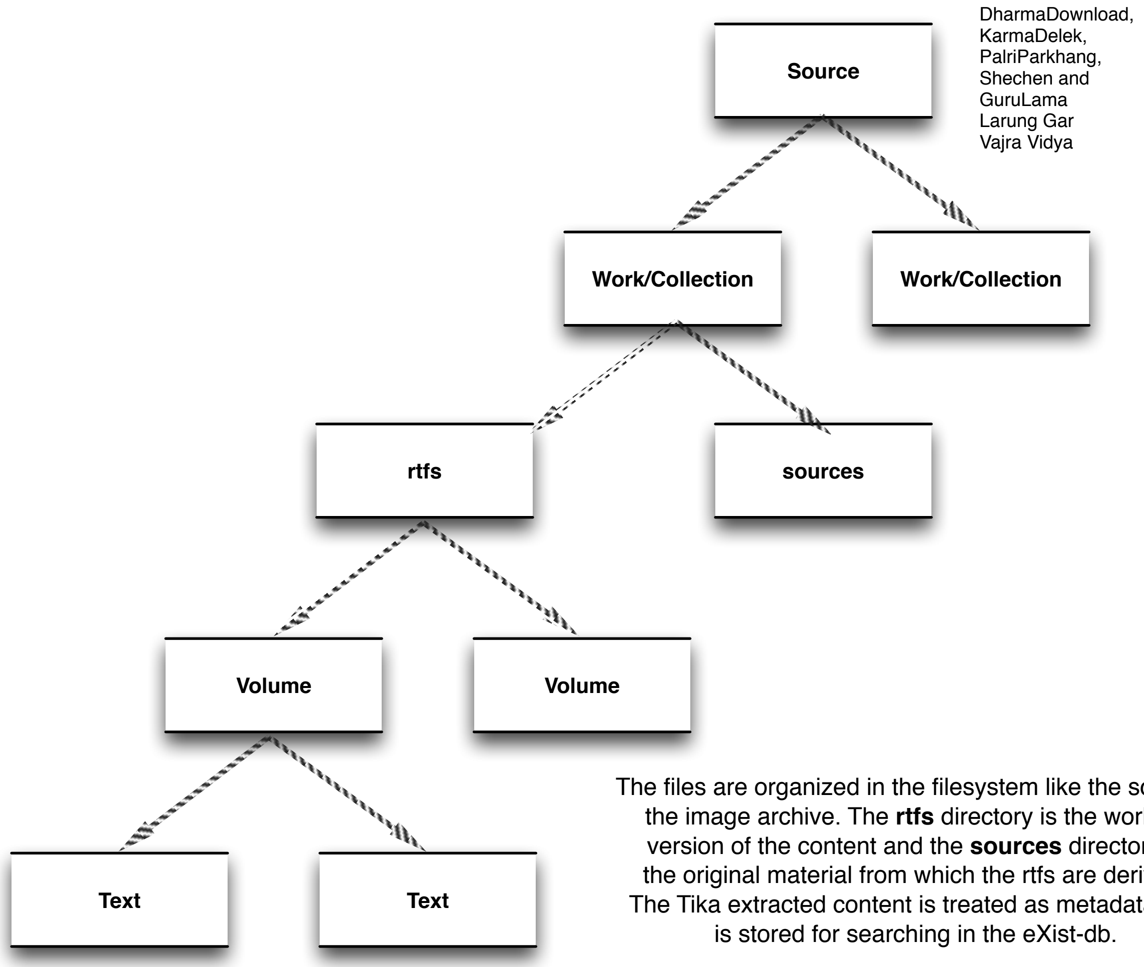
After manual *cataloguing* of keyed in materials the naming will have a Work RID appended to the Work name or in some cases there will be Work RIDs appended to "volume" names. These materials are then uploaded to **/db/eTextsCatalogued** where they will be programmatically ingested to move the items to their final locations in **/db/eTexts**.

In the case of OCR materials these are by definition already catalogued and can be uploaded directly to /db/eTexts with a simple ocr-ingest xQuery. The essential work being to name the Work with the **UT** prefix replacing the **W** of the RID of the work that has been OCR'd and naming the volume collections as **UTxxx-vvv**. The OCR'd volumes are currently each a single TEI document and will be named **UTxxx-vvv-001.xml**



Each source archive is organized as follows:

The source archive contains the original keyed in materials received by TBRC for inclusion in the eText repository. Each eText in /db/eTexts will contain metadata that links to the files in the source archive.



The files are organized in the filesystem like the scans in the image archive. The **rtfs** directory is the working version of the content and the **sources** directory is the original material from which the rtfs are derived. The Tika extracted content is treated as metadata that is stored for searching in the eXist-db.