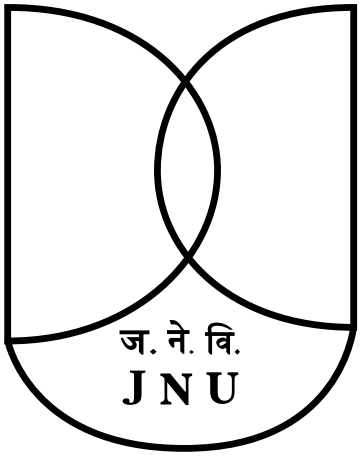

Tiered Architecture Description

JNU Data Depot

Professor Andrew Lynn, Principal Investigator

The essential aim of the JNU Data Depot is to store content for non-consumptive research as defined in the Terms of Use and Initial Concept Memorandum of the JNU Data Depot.

The JNU Data Depot is structured to have three tiers:

- **Tier 0: Raw Data Document Data Store.** The content stored in this tier will contain documents from various sources, but essentially available through a URL over a public internet connection. As this tier may contain content which may include copyrighted material, strict compliance on its use with the terms of non-consumptive research are to be followed. The content will be stored and retrieved only on the basis of a unique internal ID, which is to be linked to its publicly known identifier - such as the Document Object Identifier (DOI) or source URL. It is expected that only algorithms associated with bulk content extraction will be permitted on this tier.
- **TIer 1: Preprocessed Extracted data.** This tier will contain text and image content extracted from the raw data document tier using different protocols. They may include multiple protocols applied on the same original data source, with different levels of details required for NLP such as text extraction, section tagging, sentence tagging, word tokenisation and POS tagging. Besides the source file internal ID, files containing extracted content will also be named with a code specifying the protocol used for content extraction. As this tier may also contain content which may include copyrighted material, compliance with the terms and conditions required for non-consumptive research are to be followed.
- **Tier 2: Public Domain Data.** This tier will include public domain data sets along with mirrors of open source software repositories. This may also include extracted feature datasets derived from data in the Tier 1 of the data depot that are deemed to be independent of copyright after expert review. This tier will be freely accessible through the internet.