# Using NSDL Metadata

The NSDL publishes the metadata it aggregates at an OAI provider which is open to the public. Information about the OAI provider and how to query it can be found on the Services and Tools wiki.

All metadata available at our OAI provider has been carefully normalized to ensure that the data is standardized and of the hightest quality. The process by which we do this is known as normalization, and is explained below. Thanks to the work that has been done to develop sound, complete metadata framewords, extensive indexing, and careful normalization, NSDL's metadata is considered by many to be unique in its high quality and considerable quantity.

### The Normalization Process

The NSDl harvests a wide variety of metadata in the nsdl_dc and (in some cases) oai_dc formats. These formats have many fields which do not have controlled vocabularies, so it is extremely difficult to get useful search results from so many different (and often misspelled or misused) terms. The normalization is what allows users to receive all the possible relevant search results. Otherwise they'd only get the metadata records which use the exact same spelling, capitalization, plural vs. singular, etc. of their search term. It also allows downstream users of our OAI provider to harvest reliable, consistent metadata.

After the NSDL harvests records from external collection builders, the DDS (our Lucene index API which builds the NSDL's xml repository) checks the original metadata term against xml thesauruses we maintain for fields important to user searches (like subject, education level, type, etc.). The DDS stores the xml record in the repository with the original metadata term replaced by the thesaurus's term (this is especially useful to correct spelling errors in the original metadata). The subject field is exceptional, though, in that both the original term and the thesaurus's term are added to the record.

Figuring out which original metadata terms to replace with which normalized terms is a time-consuming activity. It is somewhat automated, but depends largely on human judgement for each term, since the choices are often higher-level than what machine-learning would be capable of. Of course, the entire process can be side-stepped by using a carefully-constructed controlled vocabulary to constrict valid xml records to certain terms. Our search interface could then be built around those specific terms--this is basically what we did with our newest metadata format, LAR (for Learning Application Readiness). It allows us to provide more innovative and advanced searching options for our users because of the constricted vocabulary.