

ADL Search Buckets Description

<!-- --> <!-- -->

Table of contents

1 ADL Search Buckets.....	2
1.1 ADL Search Buckets Summary.....	2
2 Search Bucket Detailed Description.....	5
2.1 adl:geographic locations.....	5
2.2 adl:dates.....	6
2.3 adl:types.....	6
2.4 adl:formats.....	6
2.5 adl:title.....	7
2.6 adl:originators.....	7
2.7 adl:assigned terms.....	7
2.8 adl:subject-related text.....	7
2.9 adl:identifiers.....	8

1. ADL Search Buckets

As a distributed digital library for geographically referenced information, the Alexandria Digital Library has two primary goals:

- Make it easy for new collections to be added to the library.
- Allow users of the library to issue a single query against multiple heterogeneous collections.

While the first goal requires ADL to accept existing, collection-specific metadata schemata, the second goal requires all collections to appear to the user to support identical searchable metadata.

ADL has resolved this conflict by developing search buckets, a framework for creating homogeneous views of heterogeneous metadata.

When setting up your ADL collection, you must provide mappings between the search bucket and your own specific metadata (see "Adding Collections"). We've kept the buckets few and simple to make it easier to map the search buckets to your metadata.

1.1. ADL Search Buckets Summary

The middleware's query language is a generic, XML-encoded language that supports boolean combinations of typed constraints against abstract metadata fields, or "search buckets." Its syntax and semantics are defined by the DTD [ADL-query.dtd](#) (`../dtds/ADL-query.dtd`).

ADL has defined a set of nine standard search buckets. These buckets are defined conventionally, not architecturally. Nevertheless, clients can assume that all collections support all of the following buckets. The ADL search buckets are listed below, a more detailed description follows::

1.1.1. Geographic locations

Internal name		adl:geographic-locations
Type		spatial
Operators		contains, is-contained-in, overlaps
Content		The item's spatial footprint, i.e., an approximation of the subset

ADL Search Buckets Description

		of the Earth's surface to which the item is relevant, expressed as any of several types of geometric regions defined in WGS84 latitude/longitude coordinates.
--	--	---

1.1.2. Dates

Internal name		adl:dates
Type		temporal
Operators		contains, is-contained-in, overlaps
Content		The item's temporal footprint, i.e., the range of calendar dates to which the item is relevant.

1.1.3. Types

Internal name		adl:types
Type		hierarchical
Operators		is-a
Content		Terms drawn from the ADL Object Type Thesaurus identifying the meaning or content of the item.

1.1.4. Formats

Internal name		adl:formats
Type		hierarchical
Operators		is-a
Content		Terms drawn from the ADL Object Format Thesaurus identifying the form or representation of the item.

1.1.5. Titles

Internal name		adl:titles
Type		textual
Operators		contains-all-words, contains-any-words, contains-phrase
Content		The item's title. This bucket is a subset of the "Subject-related text" bucket.

1.1.6. Originators

Internal name		adl:originators
Type		textual
Operators		contains-all-words, contains-any-words, contains-phrase
Content		Names of entities related to the origination of the item (authors, publishers, distributors, etc.).

1.1.7. Assigned Terms

Internal name		adl:assigned-terms
Type		textual
Operators		contains-all-words, contains-any-words, contains-phrase
Content		Subject-related terms from controlled vocabularies. This bucket is a subset of the "Subject-related text" bucket.

1.1.8. Subject Related Text

Internal name		adl:subject-related-text
Type		textual

ADL Search Buckets Description

Operators		contains-all-words, contains-any-words, contains-phrase
Content		Text indicative of the subject of the item, not necessarily from controlled vocabularies. This bucket is a superset of the "Titles" and "Assigned terms" buckets.

1.1.9. Identifiers

Internal name		adl:identifiers
Type		identification
Operators		matches
Content		Item names and codes that serve as unique identifiers.

2. Search Bucket Detailed Description

2.1. adl:geographic locations

The adl:geographic locations bucket contains the footprints of library objects. Each object in an ADL collection should be associated with at least one footprint (region on the surface of the Earth). The adl:geographic location bucket contains the geographic coordinates (latitude and longitude) of the vertices of the geographic polygons bounding the object's footprint. Note that a single object may have multiple disjoint footprints.

ADL uses the adl:geographic locations bucket to support its geographic search capability. Three query operations are defined on the bucket, contains (the footprint in the bucket's wholly contained in a query footprint), within (the footprint is wholly within the query footprint), and overlaps (the footprint in the bucket at least partially overlaps a query footprint). Note that contains is a proper subset of overlaps.

In practice, the footprints assigned to the adl:geographic location bucket usually fall into one of two special cases: points (single vertices) and "bounding boxes" (rectangles). Points are often the only footprint information available for relatively small features (e.g., cities on a world map). Bounding boxes have two advantages. If a feature does not have a polygon associated with it, but is associated with more than one point, then the bounding box may be trivially computed. Moreover, even if polygonal footprints are available, a bounding box may

be preferable to search on, for reasons of index efficiency.

2.2. adl:dates

The adl:dates bucket may contain any dates or date ranges (beginning and ending date) associated with the object; such as its date of creation or period of validity. Date ranges are specified by pairs of beginning and ending dates, formatted as SQL-92 TIMESTAMPS.

The adl:dates range bucket may be thought of as the temporal analog of the adl:geographic location bucket (that is, date ranges are "time footprints"). Searches against the adl:dates range bucket may likewise specify overlap or containment, although the mechanism is somewhat different. Instead of specifying a target "time footprint," a date range search is specified in terms of "on or before" a beginning date, and/or "on or after" an ending date.

2.3. adl:types

The adl:types bucket contains terms that precisely specify an object's form or content. In general, a type is the target of an "is-a" relationship; i.e., a type can be applied to an object if the assertion "object is a type" makes sense. Valid terms for the type bucket are drawn from an ADL-controlled hierarchical vocabulary. The single query operation defined on the type bucket is thus an exact match against nodes or leaves in the type hierarchy.

The adl:types and adl:format buckets are the only buckets for which we currently mandate specific sets of valid terms. This is a difficult decision, for it means that we (or somebody) must serve as the controlling authority for these term lists. In the particular cases of the adl:types and adl:format buckets, the need for program-level interoperability is strong enough, and the scope of the term spaces constrained enough, to make controlled vocabularies both useful and feasible. In the specific case of the adl:types bucket, for example, we are growing the type schemes for catalogs and gazetteers by merging existing domain-specific vocabularies.

2.4. adl:formats

The adl:formats bucket contains terms, drawn from an ADL-controlled vocabulary, that indicate the mechanisms by which a copy of the object can be delivered to a library user. In general, a format is the target of an "is-available-as" relationship; i.e., a format can be applied to an object if the assertion "object is available as format" makes sense. The distinction between type and format can be subtle (such as "aerial photograph" is a type, while "9x9 inch monochrome film positive" is a format), which is a further argument in favor of their controlled vocabularies.

Where possible, the controlled vocabulary for the adl:formats bucket exploits standard

ADL Search Buckets Description

terminology, such as MIME type specifiers for digital data. These have the advantage of being directly intelligible to programs accessing ADL holdings.

Since digital data can often be trivially and losslessly translated from one format to another, it can make sense to assign multiple formats to a single digital object. On the other hand, an analog object (such as a paper map) is generally cast into a single format upon creation, and can only be re-formatted imperfectly and with substantial effort (such as photographic enlargement.)

2.5. adl:title

The adl:title bucket contains the name by which an individual item in the collection is known. Item titles are composed of formatting instructions and names of fields which uniquely describe the item. Example: "Digital Raster Graphic, of ' || mapname || ', ' || state."

2.6. adl:originators

The adl:originators bucket contains any text describing people or institutions responsible for the object's creation or dissemination (that is, any entity that could be an object's source). In addition to authors and publishers, this could include funding agencies, curators, technical support personnel, etc.

The level of aggregation in the adl:originators bucket tends to complicate some kinds of searches, since the bucket doesn't include semantics for personal or corporate names, nor does it require a particular kind of normalization when names are entered. As a result, queries for specific personal names will necessarily be approximate.

2.7. adl:assigned terms

The adl:assigned terms bucket is a proper subset of the adl:subject-related text bucket. Text should only be placed in the adl:assigned terms bucket if it was specifically marked as significant by the object's original catalogers (that is, the creators of the object's detailed metadata). If the assigned terms are selected from a controlled vocabulary (such as a thesaurus or a subject heading list), then a link can be made from this bucket to that vocabulary. We encourage this practice, but do not assume it in the semantics of the bucket.

2.8. adl:subject-related text

The adl:subject-related text bucket contains text gleaned or derived from the object's detailed (that is, collection-specific) metadata, that indicates any topics or themes associated with the object.

Topical text is the most loosely specified of the buckets. It is an aggregation of words and phrases, from any underlying attribute, that indicate the object's topic or theme (such as title, abstract, subject, etc.) (As a limiting case, the adl:subject-related text bucket could be populated with textual representations of all an object's metadata, although we don't recommend this.)

The adl:subject-related text, adl:assigned terms, adl:originators, and adl:identifiers buckets are all searchable with the standard ADL text operators:

any word: true if the target bucket contains any of the query words, anywhere and in any order.

all words: true only if the target bucket contains all of the query words, anywhere and in any order.

exact phrase: true only if the target bucket contains the query phrase as a proper substring. This set of operators was selected as a reasonable trade-off between query power and ease of efficient implementation on a variety of platforms.

2.9. adl:identifiers

The adl:identifiers bucket may contain any labels for the object that are defined with respect to an externally controlled vocabulary or production rule (such as ISBNs, technical report numbers, satellite image identifiers, etc.). These kinds of labels are useful for finding "known objects" for which the user knows the particular identifier. Digital data are often assigned identifiers as a function of their creation; published items often have ISBNs or ISSNs; "gray literature" often has unique technical report numbers.

In effect, the adl:identifiers bucket provides a "trap door" into alternative cataloging schemes. A nomenclature that is well-defined elsewhere can be accessed through the adl:identifiers bucket, and often provides the simplest and most precise means of locating an object, assuming one is familiar with the nomenclature. (Of course, there is always a risk of false hits from multiple nomenclatures that use similar encoding rules; for example, an arbitrary 10-digit numeric identifier could be mistaken for an ISBN code).

We assume that identifiers are normalized upon entry into the buckets (such as removing dashes from ISBNs), and that the search terms are likewise normalized prior to issuing a query. The adl:identifiers bucket has no knowledge of the structure or semantics of specific identifier types.