

The ADEPT Bucket Framework



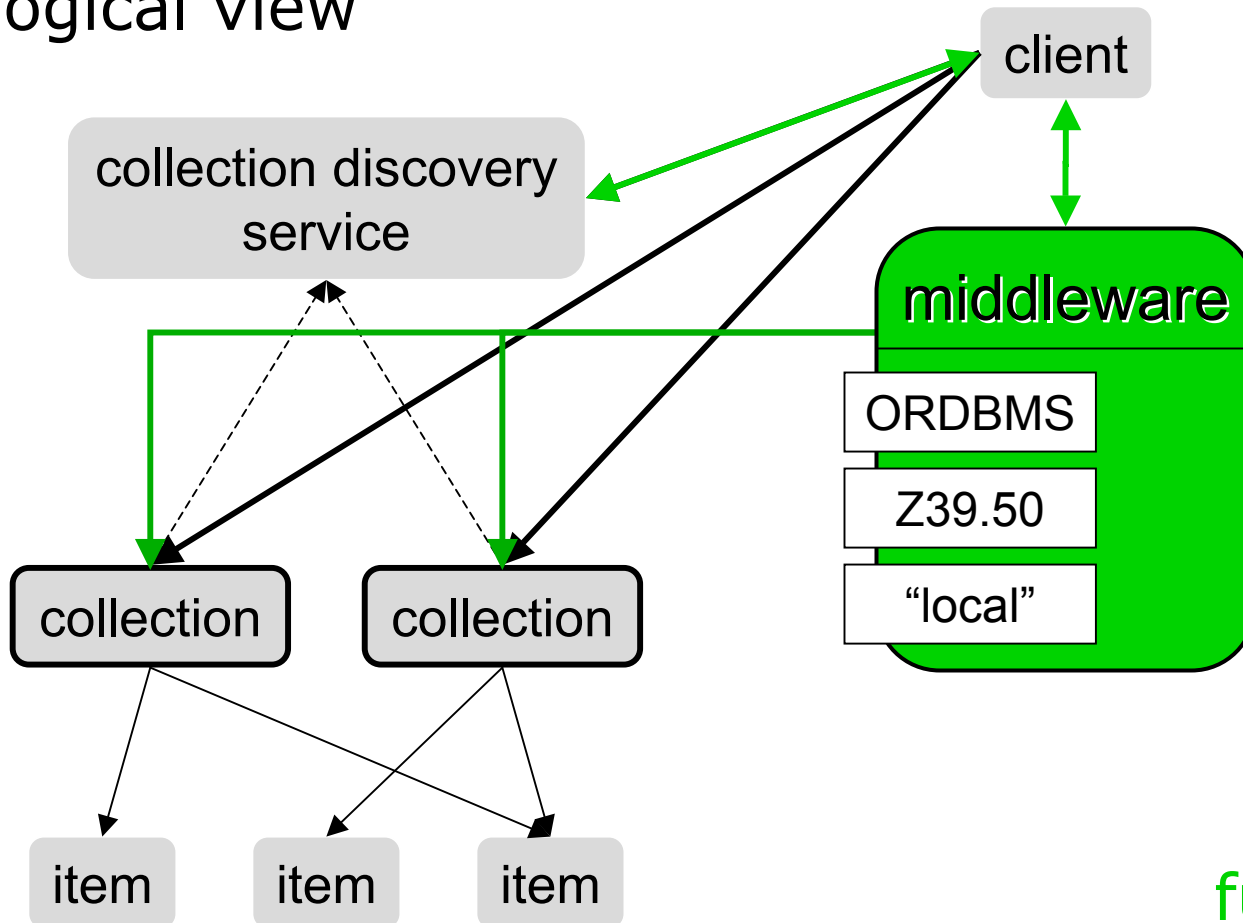
ADEPT's interoperability problem

- Distributed, heterogeneous collections
 - locally, autonomously created and managed
- Minimal requirements on collection providers
 - allow use of native metadata
- Provide uniform client services
 - common high-level interface across collections
 - structured means of discovering and exploiting (possibly collection-specific) lower-level interfaces
- Assumptions
 - items have metadata
 - items have sufficient, "good" metadata
 - i.e., this is a metadata interoperability problem



Interoperability architecture

logical view



- local access point
- standard services
- access control
- thin client support
- distributed search
- brokering of queries & results
- proxying of collections & items
- creation & organization of "local" collections

functional view



What is a bucket? (1/2)

- Strongly typed, abstract metadata category with defined search semantics to which source metadata is mapped
- Key properties
 - name
 - **Coverage date**
 - semantic definition
 - **The time period to which the item is relevant.**
 - data type (strictly observed)
 - **calendar date or range of calendar dates**
 - syntactic representation (strictly observed)
 - **ISO 8601**



What is a bucket? (2/2)

- Source metadata is mapped to buckets
 - buckets hold *not* just simple values
 - “**2001-09-08**”
 - but rather, explicit representations of mappings
 - (**FGDC, 1.3, “Time period of content”, “2001-09-08”**)
 - multiple values may be mapped per bucket
- Bucket definition includes search semantics
 - defines query terms
 - **ISO 8601 date range**
 - defines query operators
 - **contains, overlaps, is-contained-in**
 - semantics are slightly fuzzy in certain cases to accommodate multiple implementations



Item-level metadata mapping

- Each mapping is described by
 - source metadata standard
 - metadata field code and human-readable name
 - typed value
- Typical metadata value
 - DC.Creator = "U.S. Geological Survey; Photo Science, Inc."
- By contrast, a bucket value
 - Originator = {
 (FGDC, 1.1/8.1, "Citation/Originator",
 "U.S. Geological Survey"),
 (USGS DOQ, PRODUCER, "Producer",
 "Photo Science, Inc.")
 }



Collection-level aggregation

- Collection-level metadata describes
 - buckets supported by the collection
 - item-level metadata mappings
 - statistical overviews
 - **item counts**
 - **spatiotemporal coverage histograms**
- Example (de-XML-ized)
 - in collection *foo*, the *Originator* bucket is supported and the following item fields are mapped to it:
 - **(FGDC, 1.1/8.1, "Citation/Originator") [973 items]**
 - **(USGS DOQ, PRODUCER, "Producer") [973 items]**
 - **(DC, Creator, "Creator") [1249 items]**
 - **unknown [6 items]**



Searching collections

- Bucket-level
 - uniform across all collections
 - example
 - **search all collections for items whose *Originator* bucket contains the phrase “geological survey”**
- Field-level
 - collection-specific
 - but discovery and invocation mechanisms are uniform
 - functionally equivalent to searching the entire bucket plus additional constraint
 - example
 - **search collection *foo* for items whose *FGDC 1.1/8.1* field **within the *Originator*** bucket contains the phrase...**



Bucket types (1/7)

- 6 bucket types: **spatial, temporal, hierarchical, textual, qualified textual, numeric**
- Type captures the portion of the bucket definition that has functional implications
 - data type & syntactic representation
 - query terms
 - query operators
- Complete bucket definition
 - name
 - semantic definition
 - bucket type



Bucket types (2/7)

□ Spatial

- **data type:** any of several types of geometric regions defined in WGS84 latitude/longitude coordinates
- **syntax:** defined by ADEPT
- **query terms:** WGS84 box or polygon
- **operators:** contains, overlaps, is-contained-in
- **example query:**
 - **<spatial-constraint
bucket="geographic-location"
operator="overlaps">
 <box north="37.5" south="30.0" east="-110"
 west="-140"/>
</spatial-constraint>**



Bucket types (3/7)

□ Temporal

- **data type:** calendar date or range of calendar dates
- **syntax:** ISO 8601
- **query term:** range of calendar dates
- **operators:** contains, overlaps, is-contained-in
- **example query:**
 - **<temporal-constraint
bucket="coverage-date"
operator="contains"
from="1970-01-01" to="1979-12-31"/>**



Bucket types (4/7)

□ Hierarchical

- **data type:** term drawn from a controlled vocabulary (thesaurus, etc.)
- one-to-one relationship between hierarchical buckets and vocabularies
- **query term:** vocabulary term
- **operator:** is-a
- **example query:**
 - **<hierarchical-constraint
bucket="feature-type"
operator="is-a"
vocabulary="ADL Feature Type Thesaurus"
term="populated place"/>**



Bucket types (5/7)

□ Textual

- **data type:** text
- **query term:** text
- **operators:** contains-all-words, contains-any-words, contains-phrase
- **example query:**
 - **<textual-constraint
bucket="subject-related-text"
operator="contains-all-words"
text="orthophotograph"/>**



Bucket types (6/7)

□ Qualified textual

- **data type:** text with optional associated namespace
- **query term:** same
- **query operator:** matches
- **example query:**

- **<qualified-textual-constraint
bucket="identifier"
operator="matches"
text="90-70002-34-5"
namespace="ISBN"/>**



Bucket types (7/7)

□ Numeric

- **data type:** real number
- **query term:** real number
- **query operators:** standard relational operators
- **example query:**
 - **<numeric-constraint
bucket="minimum-feature-size"
operator="less-than"
value="1.0"
unit="meters"/>**



Bucket types vs. buckets

- Bucket *types* are defined architecturally
- Buckets in use are defined by collections and items
 - need standard buckets, defined conventionally, to support cross-collection uniformity
- **ADL core buckets**
 - simple; universal; easily & broadly populated; useful
- Bucket descriptions in the following slides:
 - bucket type
 - semantic definition
 - effective treatment of multiple values in searching
 - comparison to Dublin Core



ADL core buckets (1/6)

- Subject-related text
 - Title
 - Assigned term
- Originator
- Geographic location
- Coverage date
- Object type
- Feature type
- Format
- Identifier



ADL core buckets (2/6)

□ Subject-related text

- **type:** textual
- **description:** text indicative of the subject of the item, not necessarily from controlled vocabularies
- superset of *Title* and *Assigned term*
- **multiple values:** concatenated
- **compare:** DC.Subject

□ Title

- **type:** textual
- **description:** the item's title
- subset of *Subject-related text*
- **multiple values:** concatenated
- **compare:** DC.Title



ADL core buckets (3/6)

□ Assigned term

- **type:** textual
- **description:** subject-related terms from controlled vocabularies
- subset of *Subject-related text*
- **multiple values:** concatenated
- **compare:** qualified DC.Subject

□ Originator

- **type:** textual
- **description:** names of entities related to the origination of the item
- **multiple values:** concatenated
- **compare:** DC.Creator + DC.Publisher



ADL core buckets (4/6)

- **Geographic location**
 - **type:** spatial
 - **description:** the subset of the Earth's surface to which the item is relevant
 - **multiple values:** unioned
 - **compare:** DC.Coverage.Spatial
- **Coverage date**
 - **type:** temporal
 - **description:** the calendar dates to which the item is relevant
 - **multiple values:** unioned
 - **compare:** DC.Coverage.Temporal



ADL core buckets (5/6)

- **Object type**
 - **type:** hierarchical
 - **vocabulary:** ADL Object Type Thesaurus (image, map, thesis, sound recording, etc.)
 - **multiple values:** unioned
 - **compare:** DC.Type

- **Feature type**
 - **type:** hierarchical
 - **vocabulary:** ADL Feature Type Thesaurus (river, mountain, park, city, etc.)
 - **multiple values:** unioned
 - **compare:** none



ADL core buckets (6/6)

□ Format

- **type:** hierarchical
- **vocabulary:** ADL Object Format Thesaurus (loosely based on MIME)
- **multiple values:** unioned
- **compare:** DC.Format

□ Identifier

- **type:** qualified textual
- **description:** names and codes that function as unique identifiers
- **multiple values:** treated separately
- **compare:** DC.Identifier



Summary

- A *bucket* is a strongly typed, abstract metadata category with defined search semantics to which source metadata is mapped
- Supports discovery/search across distributed, heterogeneous collections that use metadata structures of their choosing
- Uses high-level search buckets for cross-collection searching and supports “drill-down” searching to the item-level metadata elements