

BucketFramework

Table of contents

1 ADL Bucket Framework.....	2
-----------------------------	---

1. ADL Bucket Framework

By Greg Janée, James Frew, Linda L. Hill, Terence R. Smith {*gjane, frew, lhill, smithtr*}@alexandria.ucsb.edu

Third DELOS Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries (Darmstadt, Germany; September 8-9, 2001)

The Alexandria Digital Library Project, as part of its Digital Earth Prototype (ADEPT) effort, is developing a distributed digital library for georeferenced information. Key features of the library include a distributed, peer-to-peer architecture; services supporting federation and interoperation of collections and items; personalized “learning spaces” in which collections of pedagogical materials can be built, shared, organized, and incorporated into curricula; and explicit support for both very large and very small (i.e., personal, local) geospatial collections.

A key component of this architecture is a framework—the ADL “bucket” framework—for creating homogeneous views of heterogeneous metadata. The ADL bucket framework supports aggregation of semantically similar, strongly typed metadata fields into high-level “buckets” for the specific purposes of providing higher-level search and description functions. Notably, in mapping metadata fields to buckets, the mappings themselves are formally and explicitly represented by the system, and metadata semantics are preserved and carried throughout. The framework defines standard representations and bucket types, including sophisticated types such as the “geospatial” bucket type which supports description of geospatial regions and searching using spatial operations. The framework also defines a particular set of 10 buckets that have proven successful in capturing key searchable and descriptive characteristics of widely heterogeneous items, while having sufficient global applicability as to be easily populated. For example, the “Originator” bucket is defined to contain textual values (specifically, names) related to the origin of an item. One item may map the tuple (MARC 100 [Personal name], “Mark Twain”) to the Originator bucket, while another may map the tuple (FGDC 1.1/8.1 [Citation/Originator], “U.S. Geological Survey”) to the same bucket. In general, an item may map any number of typed (field, value) pairs to a bucket. The ultimate effect of such mappings is that clients of the library can choose to 1) operate at the high level of buckets (e.g., a client can simultaneously search multiple collections by Originator), or 2) operate at the native metadata level (e.g., search by and retrieve MARC 100 fields in a collection that supports such native metadata).

Several other approaches to metadata interoperability have been pursued in the past, and it is useful to compare the bucket framework to them. There have been comprehensive metadata standards that attempt to capture most of the nuances and breadth of a domain [MARC, FGDC]. There have been minimalist, high-level metadata standards [Dublin Core]. Other

BucketFramework

approaches have focused on automated translation of metadata [Stanford Infobus]. And still other approaches emphasize explicit representation of, and operation on, metadata semantics [RDF, Semantic Web].

The ADL bucket framework combines many aspects of these different approaches in a new and novel way. Like the Semantic Web, the bucket framework maintains representations of metadata semantics. But unlike the Semantic Web, the bucket framework provides a mechanism for aggregating metadata into higher-level, uniform abstractions. The bucket framework is notably similar to unqualified Dublin Core in that it provides a standard set of high-level fields. But it differs just as notably: to Dublin Core the bucket framework adds search-oriented fields, strong typing, and well-defined, rich search semantics. The bucket framework provides a refinement mechanism similar to qualified Dublin Core, but unlike the latter, the refined fields are not mandated by the system, but rather are simply discoverable in a standard way. Finally, the bucket framework is similar to automated metadata translation systems in that metadata is mapped from one form to another, but the bucket framework differs from the Stanford Infobus approach in that it is focused on the simpler problem of mapping arbitrary metadata to a common, higher-level form, as opposed to mapping arbitrary forms of metadata to one another.

The ADEPT logical view of the library is that of 1) many, distributed, collections of heterogeneous items, and 2) a central collection discovery service that clients can use to locate relevant collections. ADEPT employs a standard three-tier architecture in which clients connect to collections through a middleware server, which acts as a common access point and as a kind of broker. The middleware provides SDLIP-like search and retrieval services as well as services related to collection development. The ADL bucket framework is a common thread running throughout the tiers that provides a uniform view of the library. Specifically, at the lowest level individual items within collections map their native metadata to buckets. Collections accumulate and index the mapped metadata, thereby providing a search capability over their contents at both the bucket level and at the native metadata level. Collections also aggregate information and statistics about the metadata mappings employed by the items contained in the collection, and include such information in collection-level metadata for the benefit of clients and the collection discovery service. The middleware provides standard representations of arbitrary metadata and metadata mappings. The central collection discovery harvests collection-level metadata and builds a master index.

The ultimate effect of the ADL bucket framework is that clients of the library can, without relying on any a priori knowledge or out-of-band agreements, successfully search arbitrary collections at a uniform, high level. Furthermore, the types of search provided are far more capable than just text-based search. At the same time, the richness and semantics of the underlying native metadata are entirely preserved, and are discoverable and exploitable by clients in an entirely regular way.