BEST Event Interchange Model

1. Purpose

This document describes a schema for marking up log event information. The schema is intended to be used as a standard means of representing event information, with the scope of content being described defined by APSR's Benchmark Statistics project (BEST). The Event Interchange Model was previously referred to as the Standard Statistics Specification.

2. Approach

The context of the development of this schema is the BEST project. The original objective for the schema was to act as a standard means for repositories to expose repository metadata pertinent to broader scale statistical analysis. However on closer scrutiny there are two classes of repository metadata that may be useful to an aggregator service: descriptive metadata and event metadata. Harvesting descriptive metadata allows an aggregator to answer questions such as how items are distributed across subjects and identifying gaps in the national coverage of subject areas. Event metadata allows an aggregator service to provide information about events of interest such as record downloads over time (number, where accessed from).

From an aggregation point of view, gathering descriptive metadata from repositories is a problem already solved via OAI-PMH harvesting. Repositories shipped with Data Providers by default will typically be able to expose descriptive metadata about their holdings in Dublin Core and often it is trivial to support other crosswalks (e.g. MODS or a specific DC Profile). Gathering event metadata can again be solved through the OAI-PMH model but requires a schema for representing this information. This document and accompanying schema addresses this latter issue and also proposes guidelines for providing DC metadata in a BEST context.

Across the descriptive metadata and event metadata, there are naturally points of overlap. In an event context, information that is typically captured in a repository's log files will be ID based. For example, User ID or IP address for a referrer, item ID for the referent. To make sense of this information the IDs captured in event logs need to be resolved to the descriptive metadata associated with the IDs. Any schema needs to take this into account and in an aggregation context there are means to handle this.

There is also exclusion if only event-based metadata is harvested. Only those repository holdings involved in an event of interest will be included in any event schema. The proposed schema assumes a dual harvesting model would fill in the gaps, i.e. where an aggregator has scope beyond event information or wants to provide value-add through more detailed metadata, it will harvest both descriptive metadata and event metadata and match on some record identifier. To this end, in providing for a dual-harvest model, the idea of a minimum metadata set is introduced. The intent of the minimum set is to define the metadata required in order that some description of the agents involved in an event can be known immediately after harvesting an event record.

There are opportunities available to services using the schema if a URI to the full record is also included, as well as to any services known by the providing repository. For example, if an item has a Scopus or Thomson ISI ID, a URI could be included so users of the data may value-add in their applications.

The remainder of this document covers a BEST service profile for descriptive metadata with examples, and includes a proposed event interchange schema. The proposal is to use MODS, however a DC version that may be used for testing and pilot services is also included. The hierarchical structure of

MODS may be more expressive for future descriptive metadata requirements, however DC is usually available with OAI-PMH Data Provider software and so may be more convenient for implementing demonstration services.

3. Descriptive Metadata Profiles

The following table takes the metadata fields identified by the BEST reference group and indicates how they must be marked up using two common descriptive metadata schemas, DC and MODS. It is suggested that the markup options suggested be strictly adhered to in order that the specification begin tightly then be loosened or evolved over time as use cases requiring changes are encountered. These profiles should be named and published in a “Contributor's guidelines”-type document.

Metadata Field	DC	MODS	Notes	Usage*
**Item ID	identifier	identifier	Must be globally unique	M, NR
Title	title	titleInfo/title		M, R
Author	creator	name/namePart		O, R
Abstract	description	abstract		O, R
Access Policy	rights	accessPolicy	Values must be “Open Access” or “Restricted”	M, NR
RFCD	subject	subject/topic	Values must be RFCD codes, MODS authority=”rfcd”	O, R
Item Format	format	physicalDescription /internetMediaType	Values must be pronom format ids, else mime-type	M, R
Item Type	type	genre	***Values must be sourced from MACAR resource type vocabulary	M, R
Item URI	source	location/url	URL resolvable to item metadata and content package	M, NR
Repository ID	n/a	n/a	To be provided as part of OAI-PMH “identify” request. In the absence of a persistent identifier, the home page url must be used	n/a

*M=Mandatory; O=Optional; R=Repeatable; NR=Non-Repeatable

** Other identifiers may add value, in which case the XML profiles below will need addressing as DC is not rich enough to support this. Perhaps only MODS should be supported, else order must be enforced (primary identifier listed first in DC for example)

*** At the time of writing there is discussion regarding whether the MACAR resource type provides the granularity required for research assessments; it may be a different authority list will be required.

Where the above specification is not followed, processing systems are free to discard extraneous fields and/or those fields not meeting the specification. Where the specification cannot be met and may require improvements, a feedback process must be established to evolve the above mappings.

4. Example MODS and DC item records

```
<mods>
   <identifier>hdl:1885/123</identifier>
   <titleInfo>
      <title>My example item record</title>
   <titleInfo>
   <abstract>The abstract for my item</abstract>
   <genre>type from MACAR vocabulary<genre>
   <subject authority=”rfcd”>
      <topic>270101</topic>
   </subject>
   <physicalDescription>
      <internetMediaType>fmt/3</internetMediaType>
   </physicalDescription>
   <accessCondition>Open Access</accessCondition>
   <location>
      <url>http://dspace.uni.edu.au/handle/1885/123</url>
   </location>
   <name type=”personal”>
      <namePart type=”family”>Bloggs</namePart>
      <namePart type=”given”>Joe D</namePart>
   </name>
</mods:mods>

<dc>
   <identifier>hdl:1885/123</dc:identifier>
   <title>My example item record</dc:title>
   <description>The abstract for my item</dc:description>
   <type>type from MACAR vocabulary</dc:type>
   <subject>270101</dc:subject>
   <format>fmt/3</dc:format>
   <rights>Open Access</rights>
   <source>http://dspace.uni.edu.au/handle/1885/123</dc:source>
   <creator>Joe D Bloggs</dc:creator>
</dc>
```

5. Event Schema

A schema for describing repository events is described below. Two schemas in this area were examined as potential candidates, however were not adopted due to the lack of openURL infrastructure in place to support them[1] or a focus on XML representation of log files as opposed to high-level event description[2].

The aim of this schema is to provide a standard way of marking up events of interest from disparate repositories. The schema has been designed to be simple and tight in the first instance but can be loosened as more requirements are added. At this stage very little information is required about events or requests. Should this be required the addition of an eventInfo wrapper element along with profiles for different event types could be added (e.g. search terms, request parameters, etc).

Proposed Schema

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://apsr.edu.au/standards/event"
        xmlns:evt="http://apsr.edu.au/standards/event"
xmlns:oai="http://www.openarchives.org/OAI/2.0/" elementFormDefault="qualified"
attributeFormDefault="unqualified">

   <xsd:import namespace="http://www.openarchives.org/OAI/2.0/"
schemaLocation="http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"/>

   <xsd:annotation>
      <xsd:documentation xml:lang="en">
         Schema for log events. Scope is that defined by the APSR BEST project. The *Info elements
are wrapper elements for describing components/actors of the event. This should keep the schema
"neater" as it evolves.
      </xsd:documentation>
   </xsd:annotation>

   <xsd:element name="event">
      <xsd:complexType>
         <xsd:sequence>
            <xsd:element name="timestamp" type="oai:UTCdatetimeType" minOccurs="1"
maxOccurs="1">
               <xsd:annotation>
                  <xsd:documentation xml:lang="en">
                     The timestamp needs to be international. The UTCdatetimeType is defined and used
by OAI-PMH.
                  </xsd:documentation>
               </xsd:annotation>
            </xsd:element>
            <xsd:element name="requesterInfo" minOccurs="1" maxOccurs="1">
               <xsd:complexType>
                  <xsd:sequence>
                     <xsd:element name="ip" type="evt:ipType" minOccurs="1" maxOccurs="1"/>
                  </xsd:sequence>
               </xsd:complexType>
            </xsd:element>
            <xsd:element name="referrentInfo" minOccurs="1" maxOccurs="1">
               <xsd:complexType>
                  <xsd:sequence>
                     <xsd:element name="identifier" type="xsd:string" minOccurs="1" maxOccurs="1"/>
                  </xsd:sequence>
               </xsd:complexType>
            </xsd:element>
         </xsd:sequence>
         <xsd:attribute name="type" use="required">
            <xsd:simpleType>
               <xsd:restriction base="xsd:string">
                  <xsd:enumeration value="access"/>
                  <xsd:enumeration value="ingest"/>
```

```
                <xsd:enumeration value="retrieve"/>
                <xsd:enumeration value="view"/>
              </xsd:restriction>
            </xsd:simpleType>
          </xsd:attribute>
        </xsd:complexType>
    </xsd:element>


    <!-- Global Type Definitions -->
    <xsd:simpleType name="ipType">
      <xsd:restriction base="xsd:string">
        <xsd:pattern value="\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}"/>
      </xsd:restriction>
    </xsd:simpleType>

</xsd:schema>
```

Example retrieve event

```
<event type=”retrieve”>
    <timestamp>2007-12-07T01:58:46Z</timestamp>
    <requesterInfo>
      <ip>10.1.3.236</ip>
    </requesterInfo>
    <referrentInfo>
      <identifier>hdl:1885/123</identifier>
    </referrentInfo>
</event>
```

Note the lack of support for descriptive metadata. As previously mentioned it is envisaged a dual-harvesting approach will be implemented to make life easier for contributing repositories, and to avoid complexity in a single schema handling very different classes of metadata. The identifier above is expected to be the link between the descriptive metadata and event-based metadata.

6 . References

[1] Johan Bollen and Herbert Van de Sompel. An Architecture for the Aggregation and Analysis of Scholarly Usage Data

[2] Marcos Andre Gonclaves, Ming Luo, Rao Shen, Mir Farooq Ali and Edward A Fox. An XML Log Standard and Tool for Digital Library Logging Analysis