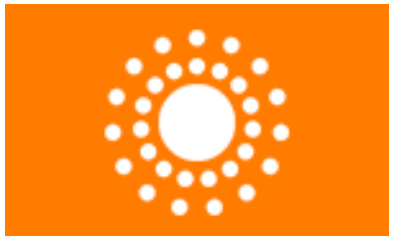
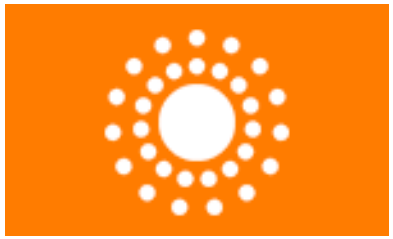


DCMI Kernel Metadata Community Meeting

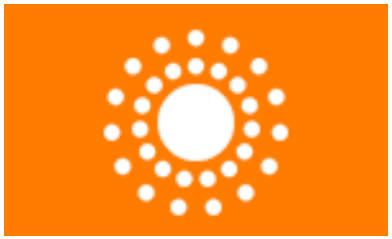
30 August 2007 – DC2007 Singapore

Kernel Agenda



- Current status and summary
- Group review of new Kernel spec, applicable to
  - Identifier support (e.g., ARK)
  - The HTTP URL Mapping Protocol (THUMP)
  - Making use of TEMPER dates, ANVL syntax, and ERC (Electronic Resource Citation) object descriptions
- Kernel Application Profile (KAP) progress report
  - Task group
  - Schedule, feedback
- Wrap up discussion and completing the KAP

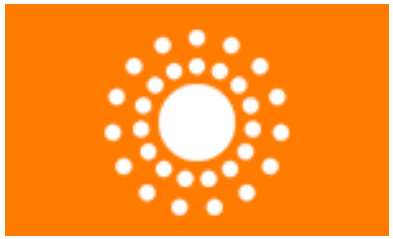
Kernel Community



- Established as Working Group in Oct 2002
- Became a DCMI community in Dec 2006
- Current mailing list: 61 subscribers
- Charter: to provide a forum…
  - a forum for those interested in very lightweight representations of Dublin Core and other metadata
  - to provide feedback to the task group creating the Kernel Application Profile
- Origin: DC 2001, Tokyo paper

```
erc: Kunze, John A. | A Metadata Kernel for Electronic Permanence
     | 20011106 | http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Kunze/
```

Dublin Kernel Results



- Supports ARK identifiers at California Digital Library
  - Every ARK identifies itself when you append a '?'
- Kernel metadata specification in Draft 2
  - http://dot.ucop.edu/home/jak/erc2.html in collaboration with Adrian Turner (CDL)
  - Specification used in teaching graduate school metadata classes (J. Greenberg, W. Moen)
- Task group drafting Kernel Application Profile
- Support in two open-source search engines
  - Amberfish and Isite2
- Perl module for production metadata available

Kernel supports ARK ids

An ARK identifies itself when you append a '?'

- Append '??' to get a support commitment also

```
http://ark.cdlib.org/ark:/13030/tf0v19n804?
```

When given to a browser, this returns

```
erc:
who:   (:unav) unavailable
what:  "Pack Shinto Temple Property for Moving -- Fumiko Miyoshi,
   18-year-old daughter of the priest of a Japanese Shinto temple
   in a Southern California defense area from which all Japanese
   are being evicted, was helped February 19 by Jimmy Okumura as
   she started packing some of the temple property in preparation
   for moving.  She is wrapping a koto, Japanese harp."--caption
   on photograph
when:  (:unav) unavailable
where: http://ark.cdlib.org/ark:/13030/tf0v19n804
```

Object Surrogates

Surrogates

- Time-honored way to avoid direct handling of objects
- Usually much smaller (eg, catalog card is smaller than a book)
- Often unencumbered and in a language you understand\
- More *uniform* (for easier processing) than objects

Reminder: What is metadata for?

- Metadata is a surrogate-based tool to help us find, use, and manage information *objects*, *resources*, or ***stuff***.

Simple metadata: pros and cons

Dublin Core metadata tried to be simple

- Goal: “specification shouldn’t register on a bathroom scale”
- Goal achieved

But DC needs application profiles to be useful; in fact what’s needed across the board are

- definition of *record*
- concept of minimal object description
- layout rules for author names and dates
- *meta-metadata*, eg, provenance, commitment statements

Simple Metadata: Dublin Core

15 elements thought to apply to almost any object – discovery as goal

Content	*Intellectual Property*	*Instantiation*
Coverage	Contributor	Date
Description	Creator	Format
Type	Publisher	Identifier
Relation	Rights	Language
Source		
Subject		
Title		

Despite DCMI efforts to correct known problems, the simplest protocol with the simplest metadata – OAI – reports an overall 36% failure rate, 77% due to metadata/encoding and protocol errors.

Simple Dublin Core metadata

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF PUBLIC "-//DUBLIN CORE//DCMES DTD
   2002/07/31//EN"
    "http://dublincore.org/documents/2002/07/31/dcmes-
   xml/dcmes-xml-dtd.dtd">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
   ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description
   rdf:about=”http://www.nap.edu/books/0309064996/html/”>
    <dc:title>The Digital Dilemma</dc:title>
    <dc:creator>National Research Council</dc:creator>
    <dc:date>2000-06-22</dc:date>
  </rdf:Description>
</rdf:RDF>
```

Same record with DCMI Kernel

Here's the same information, still machine-readable, as an Electronic Resource Citation (ERC) with Kernel metadata:

```
erc:
who:   National Research Council
what:  The Digital Dilemma
when:  2000
where: http://books.nap.edu/html/digital%5Fdilemma
```

Motivators for the ERC

- Meet the need for a simple and manipulable record
- Direct human contact with metadata is inevitable
- Record should place minimal strain on people
- Succinct, transparent, trivially parseable (2 lines of Perl code)

Making it minimal: Kernel/ERC

Electronic Resource Citation (ERC) – back to basics

- ANVL/ERC record is element sequence n email header format:

  ⇒ label, colon, value
- Long values are continued on indented lines
- A blank line ends a record

Based on cross-domain kernel distilled from Dublin Core

- **who** – a responsible person or party
- **what** – a name or other human-oriented identifier
- **when** – a date important in the object's lifecycle
- **where** – a location or a machine-oriented identifier

The Kernel notion of “story”

The same record as before, in its most compact form:

```
erc: National Research Council
     | The Digital Dilemma | 2000
     | http://books.nap.edu/html/digital%5Fdilemma
```

Short or long form starts by telling the story of an *expression* of the resource, applying who-what-when-where questions to it.

- All 4 kernel elements are required by Electronic Resource Citation
- Absent values must be explained; 7 flavors of “empty”
- Element ordering is rigid in compact form (positional semantics)
- Arbitrary additional elements may occur after the 4 elements

Other story types are possible, e.g.,

- `About-erc, support-erc, meta-erc`

A 2-story ERC record

```
erc:
who:     Tomlinson, Richard
what:    Adjustable knock down chair
when:    (:unkn)
where:   http://espacenet.com/dips/bnsviewer%{
                ? CY=ec & LG=en & DB=EPD & PN=US5498054
                & ID=US+++5498054A1+I+   %}
support-who:     European Patent Office
support-what:    (:permuc) Permanent, Unchanging Content
#        Note to ops staff:  verify date.
support-when:    20010621
support-where:   http://ark.espacenet.com/ark:/23003/US5498054
```

Mapping Kernel to Dublin Core

Kernel Element	Equivalent DC Element
who	Creator/Contributor/Publisher
what	Title
when	Date
where	Identifier
how	Type (restricted, under construction)
about-who	Subject (personage)
about-what	Subject
about-when	Coverage (temporal)
about-where	Coverage (spatial)

Kernel special values

Controlled element values have the form, "(:*ccode*)"

- e.g., missing: (:unkn) Anonymous, (:unas) Unassigned
- e.g., general: (:791) Bee Stings

Natural word order recovery keyed off of initial comma

```
who:, van Gogh, Vincent
who:, Howell, III, PhD, 1922-1987, Thurston
who:, Mao Tse Tung
what:, Health and Human Services, United States Government
Department of, The,
```

and their equivalents in natural word order:

```
Vincent van Gogh
Thurston Howell, III, PhD, 1922-1987
Mao Tse Tung
The United States Government Department of Health and
  Human Services
```

ERC dates and expansion blocks

ERC value with an “expansion” block – “%{“ and “%}”

```
where: http://foo.bar.org/node%{
        ?db= foo
        &start = 1
        &end = 5
        &buf = 2
        &query = foo + bar + zaf
    %}
```

is equivalent to the correct and intact URL,

```
where:
 http://foo.bar.org/node?db=foo&start=1&end=5&buf=2&query=foo+bar+zaf
```

Dates are in TEMPER format

`1996-2000`	(range of four years)
`1952, 1957, 1969`	(list of three years)
`1952, 1958-1967, 1985`	(mixed list of dates & ranges)
`20001229-20001231`	(range of three days)
`BCE0551~`	(circa the birth of Confucius)

Kernel/ERC summary

Kernel provides cheap, general-purpose metadata vocabulary, ERC provides requirement of 4 elements

- Kernel metadata is designed to be a low-barrier way to support orderly management of collections
- Might help resource discovery and description too
- Succinct, trivial to parse, extensible yet predictable in the kernel elements

See http://dublincore.org/groups/kernel/ for more

Thinking tiny: THUMP

The HTTP URL Mapping Protocol (THUMP)

- A set of URL-based conventions for retrieving information and conducting searches
- Can be used for focused retrievals or for broad database searches
- Based on commands put in the query string after '?'

```
http://example.foo.com/?in(books)find(war and peace)show(full)
```

Broad searching in THUMP

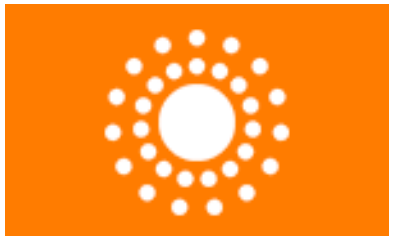
General form of broad query

```
Key ? in(DB) find(QUERY) list(RANGE) show(ELEMS) as(FORMAT)
```

Many details to be worked out; watch for

```
http://www.ietf.org/internet-drafts/draft-kunze-thump-01.txt
```

DCMI Kernel Draft 2

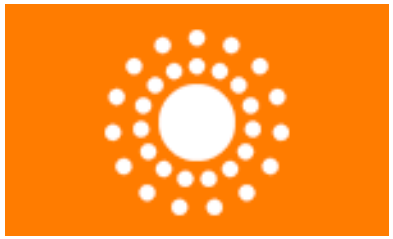


http://dot.ucop.edu/home/jak/erc2.html

With Adrian Turner, CDL

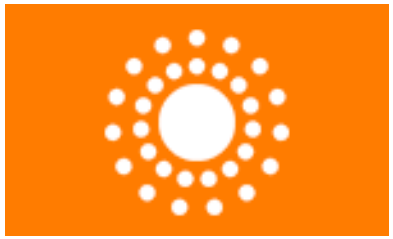
Specification used in teaching graduate school metadata classes (J. Greenberg, W. Moen)

KAP Draft



http://dublincore.org/kernelwiki/KernelApplicationProfileDraft

Kernel Wrapup



- Wrap up discussion and completing the KAP