Seeding structured data by default in open source library systems

ESWC 2014

Dan Scott, Laurentian University

https://coffeecode.net

+DanScott

#eswc2014Scott

Libraries and the Semantic Web

- Nardi and O'Day (1996) classified information seeking behaviour of library patrons as:
  - monitoring searches
  - planned searches
  - exploratory searches
- Lots of overlap with TBL's vision of the Semantic Web
- Q&A over LOD workshop (Freitas, Unger)

Traditional library systems

- Organizational systems for indexing and locating resources
- Inventory back end, with a catalogue front end
- Typically built with **MA**chine **R**eadable **C**ataloguing (*MARC*) as the metadata container

Libraries as early technology adopters

- Telnet access to catalogues
- Z39.50 protocol for sharing records
- SIP protocol for circulation transactions
- OpenURL protocol for resolving article requests
- COinS microformat for embedding citations in HTML
- unAPI for offering different metadata representations

Themes

- *Mission*: to satisfy a broad range of user queries
- Machine-readable metadata
- Culture of sharing and openness
- Technology adoption and innovation
- ...
- So traditional library systems are all over the Semantic Web, right?

No.

Libraries at an impasse

Strings, not things

MAchine **R**eadable **C**ataloging (MARC)

- Binary format designed for record sharing and tape storage in the 1960s
- Mixture of position-dependent data in "fixed fields" and fields with variable-length subfields
- Combined with cataloguing rules to create conventions like: `245 $a $b $c = Title`

```
02199cam a22004698i 4500
001 123456
005 20131223124722.0
008 130924s2013    nyu      b    001 0 eng
020    $a 9780804139571 (pbk).
100 1  $a Burgundy, Ron
245 10 $a Let me off at the top! $b my classy life and other musings
264  1 $a New York, NY : $b Crown Archetype, $c [2013].
300    $a 223 pages, 16 unnumbered pages of plates : $b illustrations ; $c 22
```

Boutique innovation

the use of HTML5 Microdata and schema.org by Google, Bing and Yahoo, and the use of RDFa by Facebook are [...] good reminders that the library software development community is best served by paying attention to mainstream solutions, as they become available, even if they eclipse homegrown stopgap solutions

Summers (2011)

GoodReads microdata

Proprietary library software

- Vendors are generally unmotivated to invest in enhancements
- Barriers to customizing for customers:
  - Non-persistent URIs(!)
  - Restricted (or non-existent!) APIs or raw data access
  - Non-standard templating systems
  - Walled garden for customizations
  - Lack of skilled resources to apply to the problem

Linked data library institutions

- Some library institutions *have* successfully implemented linked data models:
  - Swedish Union Catalog
  - German National Library
  - Bibliothèque nationale de France
  - OCLC - international co-operative

Library adoption of schema.org (2012)

- Ronallo (2013) analyzed American academic libraries' presence in the 2012 Common Crawl
- *Results*: fewer than 10,000 schema.org instances across American academic libraries

Change the things you can

- Teach open source library systems
- ... to express structured data in RDFa
- ... by default
- ... using the schema.org vocabulary

Open source library systems

- Native to the web: persistent URIs!
- Rapid, iterative community development
- Built on standard components == transferrable knowledge for contributors

Proof of concept: target implementations

Multiple systems to ensure general applicability of approach

- Koha
  - GPL since 1999
  - 2,500 - 8,000 live instances
- Evergreen
  - GPL since 2006
  - ~1,250 live instances
- VuFind
  - GPL since 2007
  - ~150 live instances
- 10,000 - 1,000,000 records per instance

schema.org vocabulary development

- W3C WebSchemas group
- W3C Schema.org Bibliographic Extension community group (SchemaBibEx)
  - Open source library systems as reference implementation

Approach

1. Map MARC21 record types to schema.org types
2. Map MARC21 record elements to schema.org properties
3. Link resources to the described object
4. Describe libraries themselves with linked data

1. Map MARC types to schema.org types

- Highest confidence in `schema:Book`, `schema:Map`, `schema:MusicAlbum`
- Fallback to `schema:CreativeWork`

- *Note*: Complexity and over-generalization of MARC21 hampers mapping efforts
- ... is a "Projected medium - Videorecording - Electronic" a `schema:Movie` or `schema:TVSeries`?

2. Map MARC21 record elements to schema.org properties

- Respect `schema` range and embedded entities (`schema:Person` for `schema:author` rather than just literals)
- Focus on `schema:CreativeWork` properties, as more specific types inherit without adding many properties
- *Implementation*: `XPath` against an XML serialization of MARC, with regular expressions where necessary

3. Link resources to the described object

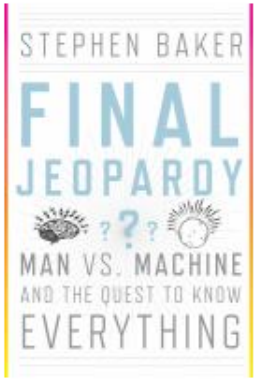
- Adopt the GoodRelations agent-promise-object model

4. Describe libraries themselves with linked data

- Use `schema:Library` to supply address, operating hours, contact information
- *Note*: Currently only implemented by Evergreen

Final Jeopardy : man vs. machine and the quest to know everything / Stephen Baker.

Baker, Stephen, 1955- (Author).



Book

Place Hold

Add to my list

Print / Email

Record details

- **ISBN:** 9780547483160 (hardback)
- **ISBN:** 0547483163 (hardback)
- **Physical Description:** 268 p. ; 22 cm.
- **Publisher:** Boston : Houghton Mifflin Harcourt, 2011.

Content descriptions

Summary, etc.: "Researchers at IBM have launched a billion-dollar project to develop a machine that can compete in the quiz show Jeopardy--and win."-- Provided by publisher.

Available copies

- 15 copies at Merrimack Valley Library Consortium. (Show)
- 1 copy at Amesbury Public Library.

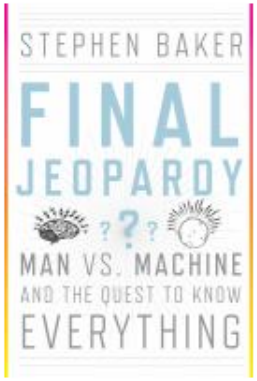
Current holds

0 current holds with 16 total copies.

Location	Call Number / Copy Notes	Collection	Status	Due Date
Amesbury Public Library	006.3 BAK (Text)	Stacks	Available	-

Final Jeopardy : man vs. machine and the quest to know everything / Stephen Baker.

Baker, Stephen, 1955- (Author).



Book

<#schemarecord> a schema:Book, schema:Product;
schema:name "Final Jeopardy : ...."@en-us ;

Place Hold

Add to my list

Print / Email

Record details

- **ISBN:** 9780547483160 (hardback)
- **ISBN:** 0547483163 (hardback)
- **Physical Description:** 268 p. ; 22 cm.
- **Publisher:** Boston : Houghton Mifflin Harcourt, 2011.

Content descriptions

Summary, etc.: "Researchers at IBM have launched a billion-dollar project to develop a machine that can compete in the quiz show Jeopardy--and win."-- Provided by publisher.

Available copies

- 15 copies at Merrimack Valley Library Consortium. (Show)
- 1 copy at Amesbury Public Library.

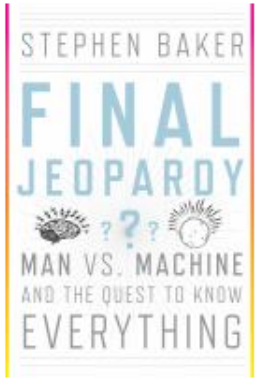
Current holds

0 current holds with 16 total copies.

Location	Call Number / Copy Notes	Collection	Status	Due Date
Amesbury Public Library	006.3 BAK (Text)	Stacks	Available	-

Final Jeopardy : man vs. machine and the quest to know everything / Stephen Baker.

Baker, Stephen, 1955- (Author).



Book


```
<#schemarecord> a schema:Book, schema:Product ;
   schema:author <#schemacontrib1> ;

<#schemacontrib1> a  schema:Person ;
   schema:birthDate "1955"@en-us ;
   schema:description "Author"@en-us ;
   schema:name "Baker, Stephen,"@en-us .
```


Place Hold

Add to my list

Print / Email

Record details

- **ISBN:** 9780547483160 (hardback)
- **ISBN:** 0547483163 (hardback)
- **Physical Description:** 268 p. ; 22 cm.
- **Publisher:** Boston : Houghton Mifflin Harcourt, 2011.

Content descriptions

Summary, etc.: "Researchers at IBM have launched a billion-dollar project to develop a machine that can compete in the quiz show Jeopardy--and win."-- Provided by publisher.

Available copies

- 15 copies at Merrimack Valley Library Consortium. (Show)
- 1 copy at Amesbury Public Library.

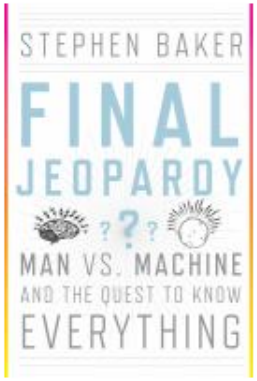
Current holds

0 current holds with 16 total copies.

Location	Call Number / Copy Notes	Collection	Status	Due Date
Amesbury Public Library	006.3 BAK (Text)	Stacks	Available	-

Final Jeopardy : man vs. machine and the quest to know everything / Stephen Baker.

Baker, Stephen, 1955- (Author).



Book

<#schemarecord> a schema:Book, schema:Product;
schema:isbn "0547483163"@en-us,
"9780547483160"@en-us ;

Place Hold

Add to my list

Print / Email

Record details

- **ISBN:** 9780547483160 (hardback)
- **ISBN:** 0547483163 (hardback)
- **Physical Description:** 268 p. ; 22 cm.
- **Publisher:** Boston : Houghton Mifflin Harcourt, 2011.

Content descriptions

Summary, etc.: "Researchers at IBM have launched a billion-dollar project to develop a machine that can compete in the quiz show Jeopardy--and win."-- Provided by publisher.

Available copies

- 15 copies at Merrimack Valley Library Consortium. (Show)
- 1 copy at Amesbury Public Library.

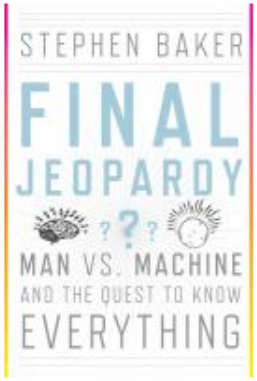
Current holds

0 current holds with 16 total copies.

Location	Call Number / Copy Notes	Collection	Status	Due Date
Amesbury Public Library	006.3 BAK (Text)	Stacks	Available	-

Final Jeopardy : man vs. machine and the quest to know everything / Stephen Baker.

Baker, Stephen, 1955- (Author).



Book

<#schemarecord> a schema:Book, schema:Product;
schema:publisher [a schema:Organization ;
schema:location "Boston :"@en-us ;
schema:name "Houghton Mifflin Harcourt,"@en-us]

Place Hold

Add to my list

Print / Email

Record details

- **ISBN:** 9780547483160 (hardback)
- **ISBN:** 0547483163 (hardback)
- **Physical Description:** 268 p. ; 22 cm.
- **Publisher:** Boston : Houghton Mifflin Harcourt, 2011.

Content descriptions

Summary, etc.: "Researchers at IBM have launched a billion-dollar project to develop a machine that can compete in the quiz show Jeopardy--and win."-- Provided by publisher.

Available copies

- 15 copies at Merrimack Valley Library Consortium. (Show)
- 1 copy at Amesbury Public Library.

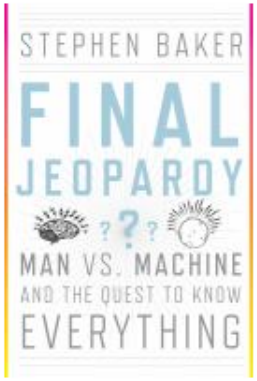
Current holds

0 current holds with 16 total copies.

Location	Call Number / Copy Notes	Collection	Status	Due Date
Amesbury Public Library	006.3 BAK (Text)	Stacks	Available	-

Final Jeopardy : man vs. machine and the quest to know everything / Stephen Baker.

Baker, Stephen, 1955- (Author).



Book

<#schemarecord> a schema:Book, schema:Product;
schema:datePublished "2011."@en-us ;

Place Hold

Add to my list

Print / Email

Record details

- **ISBN:** 9780547483160 (hardback)
- **ISBN:** 0547483163 (hardback)
- **Physical Description:** 268 p. ; 22 cm.
- **Publisher:** Boston : Houghton Mifflin Harcourt, 2011.

Content descriptions

Summary, etc.: "Researchers at IBM have launched a billion-dollar project to develop a machine that can compete in the quiz show Jeopardy--and win."-- Provided by publisher.

Available copies

- 15 copies at Merrimack Valley Library Consortium. (Show)
- 1 copy at Amesbury Public Library.

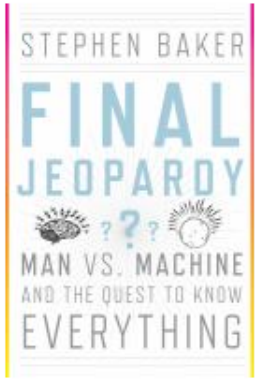
Current holds

0 current holds with 16 total copies.

Location	Call Number / Copy Notes	Collection	Status	Due Date
Amesbury Public Library	006.3 BAK (Text)	Stacks	Available	-

Final Jeopardy : man vs. machine and the quest to know everything / Stephen Baker.

Baker, Stephen, 1955- (Author).



Book

<#schemarecord> a schema:Book, schema:Product;
[a schema:Offer ;
schema:availability schema:InStock ;
schema:availableAtOrFrom "Stacks"@en-us ;
schema:businessFunction <http://purl.org/goodrelations/v1#LeaseOut> ;
schema:itemOffered <#schemarecord> ;
schema:seller <http://example.org/library/MAM> ;
schema:sku "006.3 BAK"@en-us],

Place Hold

Add to my list

Print / Email

Record details

- **ISBN:** 9780547483160 (hardback)
- **ISBN:** 0547483163 (hardback)
- **Physical Description:** 268 p. ; 22 cm.
- **Publisher:** Boston : Houghton Mifflin Harcourt, 2011.

Content descriptions

Summary, etc.: "Researchers at IBM have launched a billion-dollar project to develop a machine that can compete in the quiz show Jeopardy--and win."-- Provided by publisher.

Available copies

- 15 copies at Merrimack Valley Library Consortium. (Show)
- 1 copy at Amesbury Public Library.

Current holds

0 current holds with 16 total copies.

Location	Call Number / Copy Notes	Collection	Status	Due Date
Amesbury Public Library	006.3 BAK (Text)	Stacks	Available	-

Library structured data

```
[] a schema:Library ;
    schema:address _:N636dfc60d58348a6abfe137cedde470a ;
    schema:branchOf <http://example.org/library/AMESBURY> ;
    schema:location _:N636dfc60d58348a6abfe137cedde470a ;
    schema:name "Example Public Library"@en-us ;
    schema:openingHoursSpecification [ a schema:OpeningHoursSpecification ;
            schema:closes "8:00 PM"@en-us ;
            schema:dayOfWeek <http://purl.org/goodrelations/v1#Monday> ;
            schema:opens "10:00 AM"@en-us ], ... ;
    schema:telephone <tel:555-555-5555> ;
    schema:url <http://example.org/> .

_:N636dfc60d58348a6abfe137cedde470a a schema:PostalAddress ;
    schema:addressCountry "US"@en-us ;
    schema:addressLocality "Example"@en-us ;
    schema:addressRegion "MA"@en-us ;
    schema:contactType "Mailing address"@en-us ;
    schema:postalCode "90210"@en-us ;
```

Results

- Successfully implemented approach in all three target systems
  - Currently expresses 10 schema.org types and 37 properties
  - Code contributions were reviewed and accepted by upstream projects for release
- schema.org was able to express most common library resources (`Periodicals` excepted)
- Proof-of-concept union catalogue using Google Custom Search Engine
  - 5 minutes to combine arbitrary library systems
- ~5,000 libraries worldwide will publish linked data, with no effort, as they upgrade to the latest release

Conclusion

- Contributions:
  - *In-use*: two most prominent open source library systems, and one discovery system, now express structured data
  - Patterns and code for publishing structured data in library systems
  - GoodRelations scope now includes non-commercial
  - Proposed extensions to schema.org vocabulary to support Periodicals and multi-volume CreativeWorks
- schema.org structured data offers a useful, lightweight transitional approach for libraries
- Standard Web technologies such as RDFa and sitemaps offer viable alternatives to legacy library protocols and standards