

Semantic MARC.

or, How Charles Ammi Cutter was using RDF in 1867



Rob Styles & Nadeem Shabir

Pre-Publication DRAFT, February 2008, submitted to <http://events.linkeddata.org/ldow2008/>

SEMANTIC MARC, MARC21 AND THE SEMANTIC WEB

Rob Styles
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
rob.styles@talis.com

Danny Ayers
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
danny.ayers@talis.com

Nadeem Shabir
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
nadeem.shabir@talis.com

ABSTRACT

The MARC standard for exchanging bibliographic data has been in use for several decades and is used by major libraries worldwide. This paper discusses the possibilities of representing the most prevalent form of MARC, MARC21, as RDF for the Semantic Web, and aims to understand the tradeoffs, if any, resulting from transforming the data. Critically our approach goes beyond a simple transformation of the MARC21 record content to develop rich semantic descriptions of the varied things which may be described using bibliographic records. We present an algorithmic approach for consistently generating URIs from critical data, discuss the algorithmic matching of author names and suggest how RDF generated from MARC records may be linked to other data sources on the Web.

Keywords

MARC, MARC21, RDF, Semantic Web, Data Conversion, Inferred Semantics.

1. INTRODUCTION

A great deal of data exists as strings of text in structured form within binary file formats. Imagine all the ID3 tags on MP3s or all the EXIF tags in jpeg images. A more complex variation is the bibliographic data created by the hard work of generations of librarians, going as far back as the purpose of this paper. The principles described here, though, are equally applicable to any form of data where humans are left to infer meaning from literal strings.

Copyright notice to go here.

Pre-Publication DRAFT, February 2008, submitted to <http://events.linkeddata.org/ldow2008/>

The MARC standard for exchanging data has been around for more than 50 years. It is a structured binary format that has allowed libraries to exchange bibliographic data very successfully. So successfully, in fact, that the Library of Congress and British Library have around 10 million records in this form each. Most national libraries have a similar number. OCLC Worldcat, a US database of library information has many tens of millions. The data is not readily available for reuse outside of the library community. Talis has, for more than 40 years, maintained a database of such bibliographic records currently numbering in the tens of millions, a mixture of contributed data from libraries and commercial data from suppliers.

The Semantic Web, a web of data linked through the use of URIs and accessible over HTTP, offers the opportunity to create large, interconnected sets of data.

This paper aims to discuss the possibilities of representing the most prevalent form of MARC, MARC21, as RDF for the Semantic Web.

2. MARC21

MARC21 is used to describe several different types of record in library catalogues. Bibliographic records describe publications, authority records list the names of authors, names, titles or subject headings. All of the major library management systems in use in English-speaking countries are able to import and export data in this form.

There are other flavours of MARC: Unimarc, UNIMARC and UNIMARC are just some examples. The different MARCs essentially all share an underlying record system. UNIMARC has very much in common with MARC21, but vary in the semantics assigned to different parts of the record. They differ in the level of granularity at which they store data, so single name fields versus separate fore and surname being one example, and also in where they locate data within a record – that is what meaning is assigned to each position.

Given an increasing volume of online knowledge due to their massive digitisation projects, use a mixture of MARC21, UNIMARC and UNIMARC. With the volume of data available in MARC21 and the global connectivity provided by the internet, MARC21 is rapidly becoming the lingua franca for libraries globally. The techniques described in this paper are equally applicable to all flavours of MARC as well as other data formats.

<http://events.linkeddata.org/ldow2008/#program>









<http://www.youtube.com/watch?v=6eGcsGPgUTw>

- Resources v Literals
- Synthetic or Natural Keys
- Dealing with Ambiguity

00673nam a2200217 a 45040010033000000030009000330
0500170004200800410005901500190010002000170011903
5001700136040003100153082001600184100001900200245
0062002192600033002813000020003146500060003346500
031003946550030004259cbbe7fc3a7346d99c281979d45b6
79cUK-BiTAL20050705133033.0990831s1999 enk
j 000 ||eng|d aGB99Y57412bnb a0747542155 :
a()0747542155 aStDuBDScStDuBDSDUK-BiTAL04a823.
9142211 aRowling, J. K.00aHarry Potter and the Pr
isoner of Azkaban /cJ.K. Rowling. aLondon :bBloom
sbury,c1999. a317p. ;c21 cm. 0aPotter, Harry (Fi
ctitious character)vJuvenile fiction. 0aWizardsvJ
uvenile fiction. 7aChildren's stories.2lcs


```
=LDR 00673nam a2200217 a 4504
=001 9cbbe7fc3a7346d99c281979d45b679c
=003 UK-BiTAL
=005 20050705133033.0
=008 990831s1999\\\\enk j\\\\\\000\\||eng|d
=015 \\$aGB99Y5741$2bnb
=020 \\$a0747542155 :
=035 \\$a()0747542155
=040 \\$aStDuBDS$cStDuBDS$dUK-BiTAL
=082 04$a823.914$221
=100 1\\$aRowling, J. K.
=245 00$aHarry Potter and the Prisoner of Azkaban /$cJ.K. Rowling.
=260 \\$aLondon :$bBloomsbury,$c1999.
=300 \\$a317p. ;$c21 cm.
=650 \\0$aPotter, Harry (Fictitious character)$vJuvenile fiction.
=650 \\0$aWizards$vJuvenile fiction.
=655 \\7$aChildren's stories.$2lcs
```

```
=LDR 00673nam a2200217 a 4504
=001 9cbbe7fc3a7346d99c281979d45b679c
=003 UK-BiTAL
=005 20050705133033.0
=008 990831s1999\\|\\|\\|enk j\\|\\|\\|\\000\\|\\|eng|d
=015 \\$aGB99Y5741$2bnb
=020 \\$a0747542155 :
=035 \\$a()0747542155
=040 \\$aStDuBDS$cStDuBDS$dUK-BiTAL
=082 04$a823.914$221
=100 1\\$aRowling, J. K.
=245 00$aHarry Potter and the Prisoner of Azkaban /$cJ.K. Rowling.
=260 \\$aLondon :$bBloomsbury,$c1999.
=300 \\$a317p. ;$c21 cm.
=650 \\0$aPotter, Harry (Fictitious character)$vJuvenile fiction.
=650 \\0$aWizards$vJuvenile fiction.
=655 \\7$aChildren's stories.$2lcs
```

0747542155

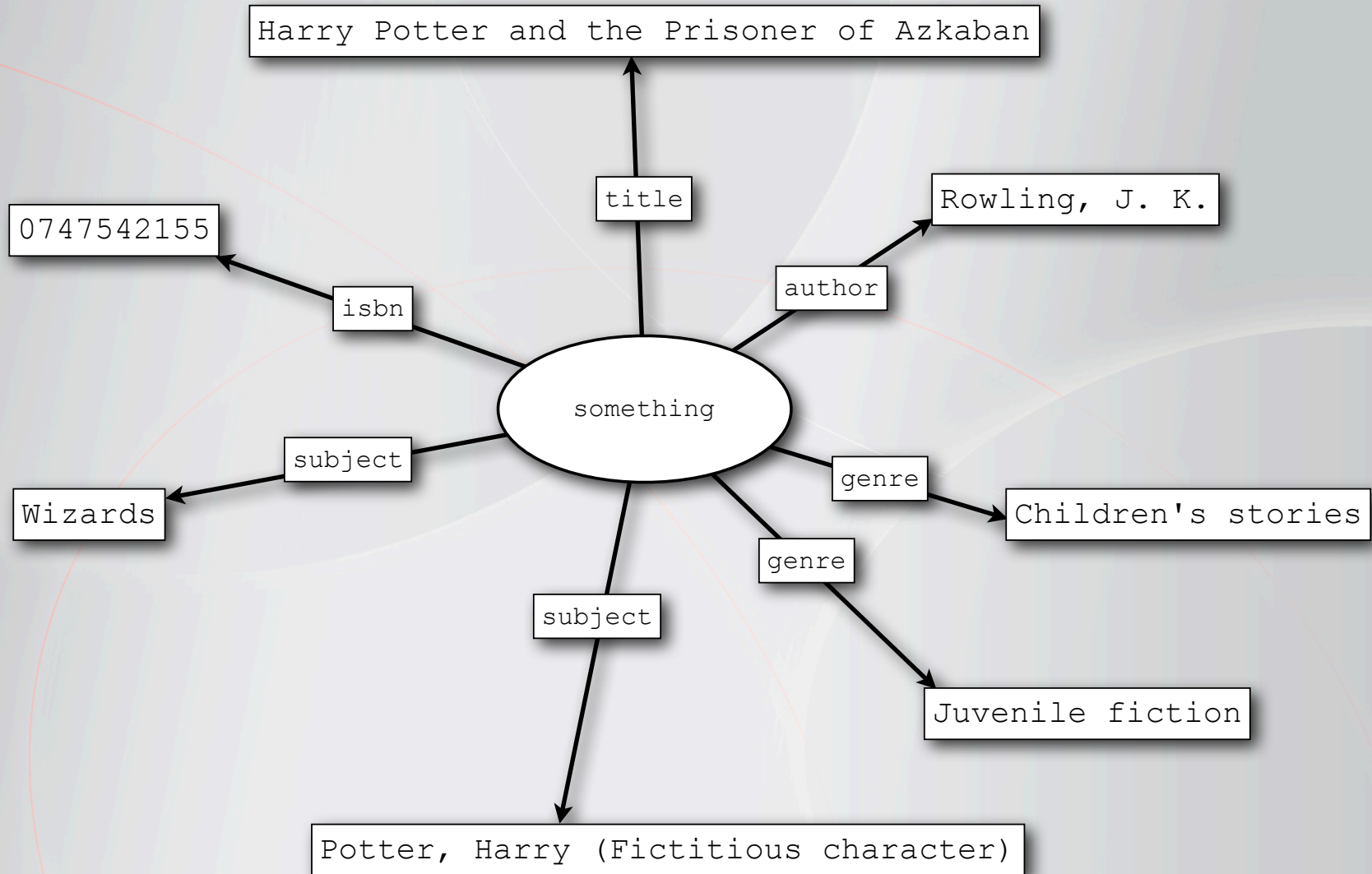
Rowling, J. K.

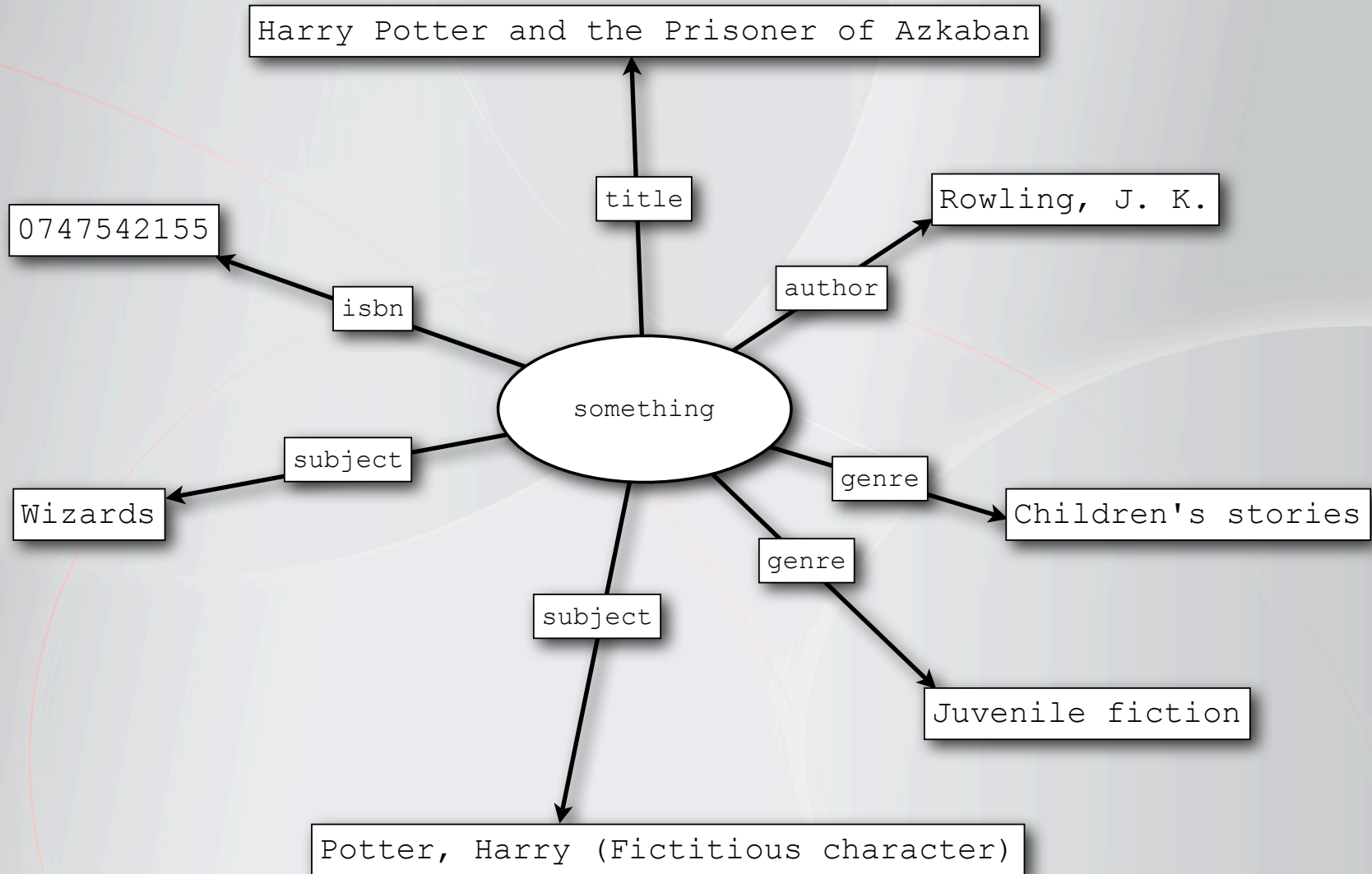
Harry Potter and the Prisoner of Azkaban

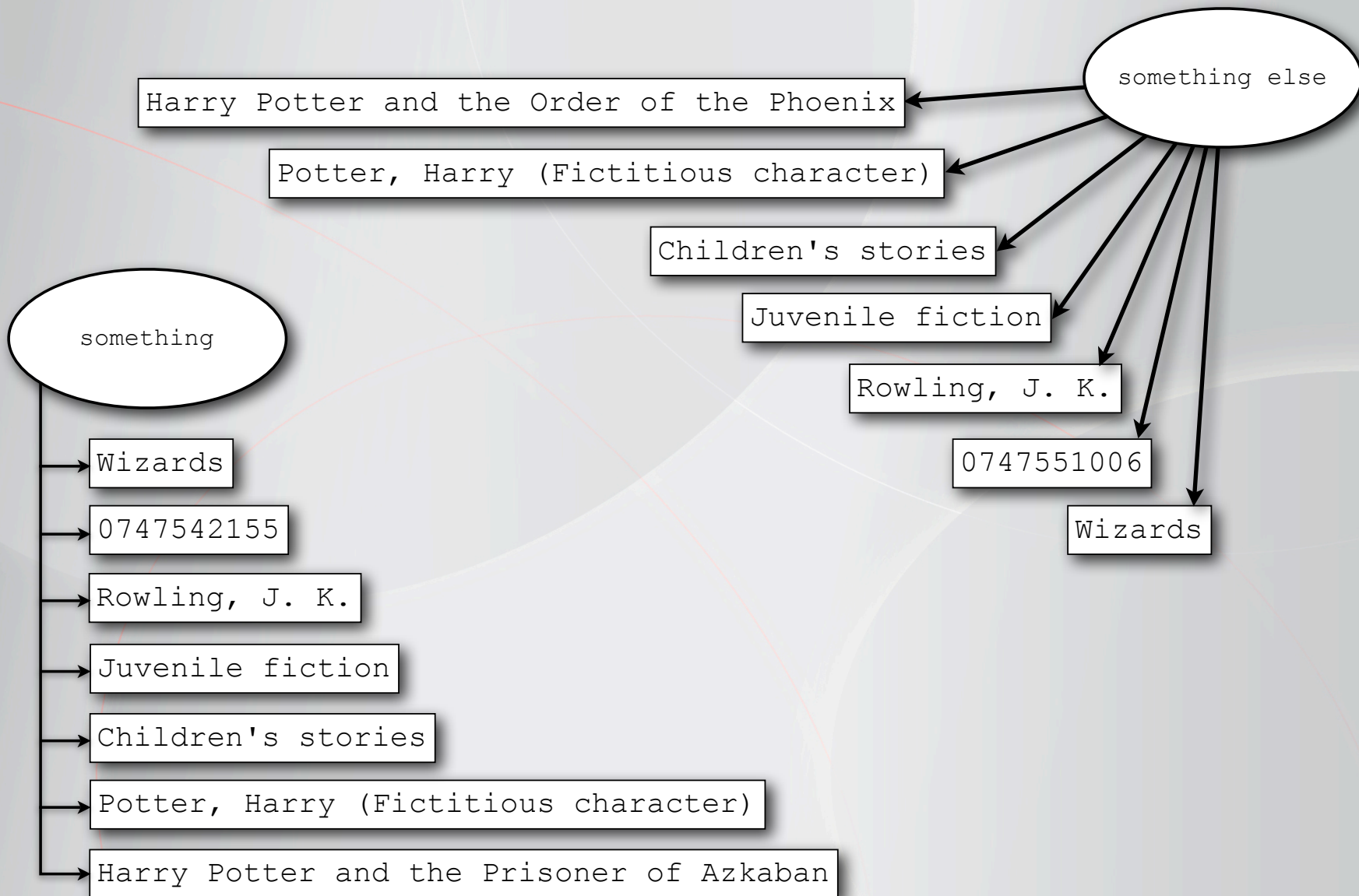
Potter, Harry (Fictitious character)

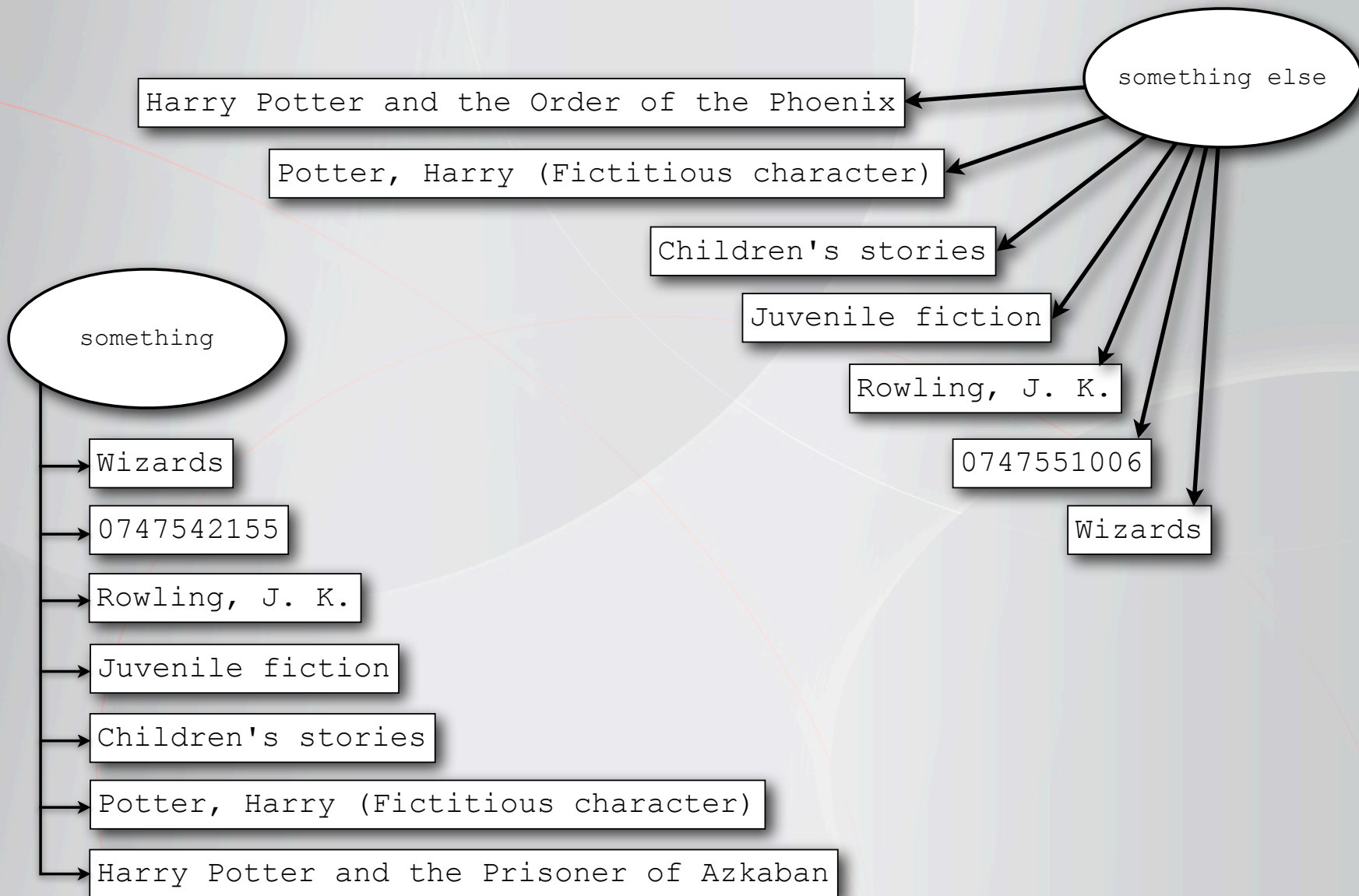
Wizards Juvenile fiction

Children's stories











Joanne K. Rowling



Charles Ammi Cutter (1837 – 1903)

[Whole Number 540]
U. S. BUREAU OF EDUCATION
SPECIAL REPORT ON PUBLIC LIBRARIES—PART II

RULES

FOR A

DICTIONARY CATALOG

BY
CHARLES A. CUTTER
LIBRARIAN OF THE FORBES LIBRARY, NORTHAMPTON, MASS.

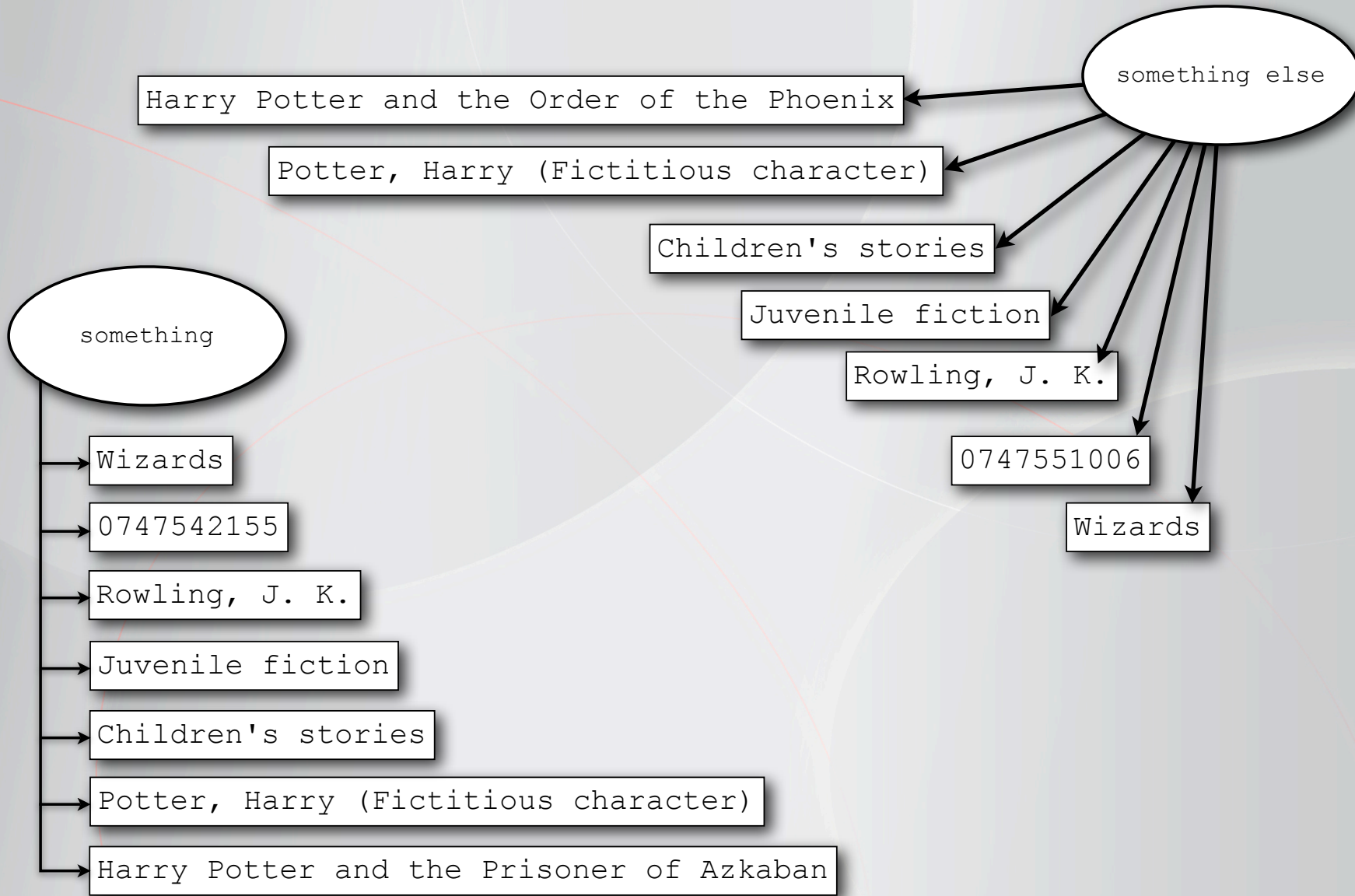
FOURTH EDITION, REWRITTEN

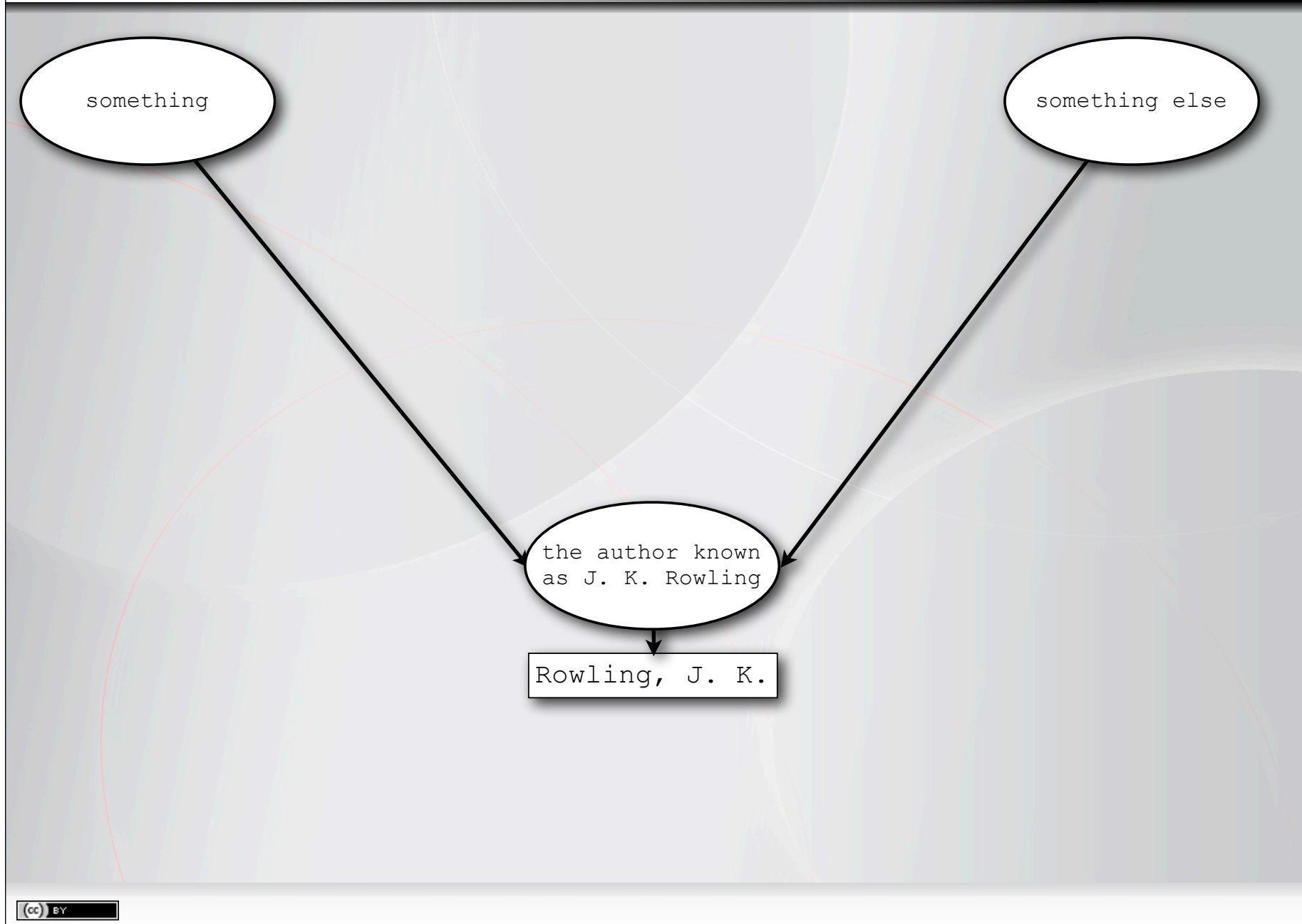
WASHINGTON
GOVERNMENT PRINTING OFFICE
1904

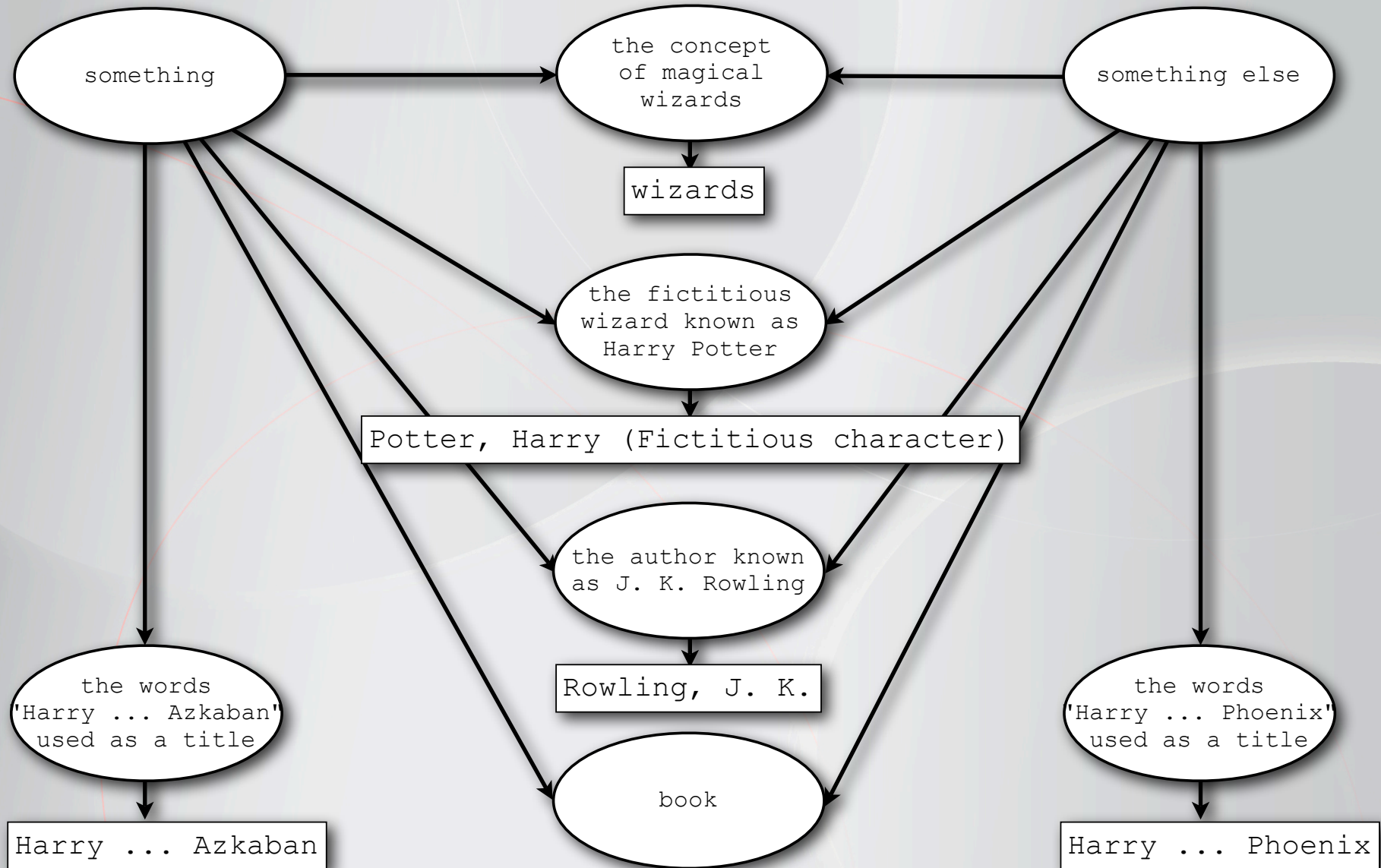
For sale by the Superintendent of Documents, Washington, D. C. Price 20 cents

U. S. GOVERNMENT PRINTING OFFICE
1904
WASHINGTON









/resource/Dog

/3020251/

/factbook/resource/China

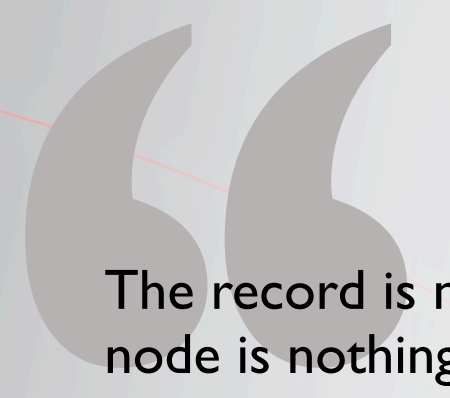
/music/artist/60d41417-feda-4734-bbbf-7dcc30e08a83

/dblp/resource/record/journals/ac/DavisR61

/rdf/usgov/geo/us/or

/bookmashup/books/006251587X

/bookmashup/doc/persons/Iain+M+Banks



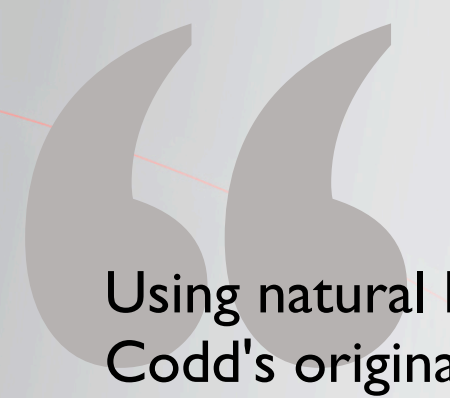
The record is nothing but the content of its fields, just as an RDF node is nothing but the connections: the property values. The mapping is very direct

- * a record is an RDF node;
- * the field (column) name is RDF propertyType; and
- * the record field (table cell) is a value.

Indeed, one of the main driving forces for the Semantic web, has always been the expression, on the Web, of the vast amount of relational database information in a way that can be processed by machines.


Relational Databases on the Semantic Web, Sir Tim Berners-Lee
<http://www.w3.org/DesignIssues/RDB-RDF.html>

URI \supseteq Primary/Foreign Keys



Using natural keys is the traditional approach, in line with Codd's original relational model. When you use them, you have only natural data that means something to users. This is good if users will ask ad hoc queries directly to the database in raw SQL. You can also often reduce the numbers of joins when using natural keys because you don't have to go to a lookup table to convert an ID to a description.

The Cost of GUIDs as Primary Keys Jimmy Nilsson
<http://www.informit.com/articles/article.aspx?p=25862>



The fundamental issue is that keys are a significant source of coupling within a relational schema, and as a result they are difficult to change. The implication is that you generally want to avoid keys with business meaning because business meaning changes.

Choosing a Primary Key: Natural or Surrogate? Scott W Ambler
<http://www.agiledata.org/essays/keys.html>

```
=LDR 00673nam a2200217 a 4504
=001 9cbbe7fc3a7346d99c281979d45b679c
=003 UK-BiTAL
=005 20050705133033.0
=008 990831s1999\\\\enk j\\\\\\000\\||eng|d
=015 \\$aGB99Y5741$2bnb
=020 \\$a0747542155 :
=035 \\$a()0747542155
=040 \\$aStDuBDS$cStDuBDS$dUK-BiTAL
=082 04$a823.914$221
=100 1\\$aRowling, J. K.
=245 00$aHarry Potter and the Prisoner of Azkaban /$cJ.K. Rowling.
=260 \\$aLondon :$bBloomsbury,$c1999.
=300 \\$a317p. ;$c21 cm.
=650 \\0$aPotter, Harry (Fictitious character)$vJuvenile fiction.
=650 \\0$aWizards$vJuvenile fiction.
=655 \\7$aChildren's stories.$2lcs
```



```
=LDR 00673nam a2200217 a 4504
=001 9cbbe7fc3a7346d99c281979d45b679c
=003 UK-BiTAL
=005 20050705133033.0
=008 990831s1999\\|\\|\\|enk j\\|\\|\\|\\000\\|\\|eng|d
=015 \\$aGB99Y5741$2bnb
=020 \\$a0747542155 :
=035 \\$a()0747542155
=040 \\$aStDuBDS$cStDuBDS$dUK-BiTAL
=082 04$a823.914$221
=100 1\\$aRowling, J. K.
=245 00$aHarry Potter and the Prisoner of Azkaban /$cJ.K. Rowling.
=260 \\$aLondon :$bBloomsbury,$c1999.
=300 \\$a317p. ;$c21 cm.
=650 \\0$aPotter, Harry (Fictitious character)$vJuvenile fiction.
=650 \\0$aWizards$vJuvenile fiction.
=655 \\7$aChildren's stories.$2lcs
```

0747542155

Rowling, J. K.

Harry Potter and the Prisoner of Azkaban

Potter, Harry (Fictitious character)

Wizards Juvenile fiction

Children's stories

0747542155

urn:isbn:0747542155

Rowling, J. K.

/people/36082b69-ba77-486b-b27d-bf3ac3f1bfe7

Harry Potter and the Prisoner of Azkaban

/titles/08944d4d-5b46-4bf5-9acf-3102b181de95

Potter, Harry (Fictitious character)

/character/e8b7ae0c-f465-4251-9bc9-bc4b6a61eb21

Wizards

/topics/08f0fa23-0cb8-4a66-a310-dfd8ed95e0ae

Juvenile fiction

/genres/ea65a567-bc36-4a23-a9de-bad053d18568

Children's stories

/genres/f96eda4a-42ab-4d57-8fc9-96e6f6f81e98

Conclusion...

Synthetic Keys are a Closed-World Mechanism.

Conclusion...

Natural Keys are Open,
difficult and require some
additional thinking.

Example...

Rowling, J. K.

Example...

/people/**rowling, j. k.**

Example...

Harry Potter and The Prisoner of Azkaban

Prisoner of Azkaban, Harry Potter and The

Example...

harry potter and the prisoner of azkaban

prisoner of azkaban harry potter and the

Example...

and azkaban harry of potter prisoner the

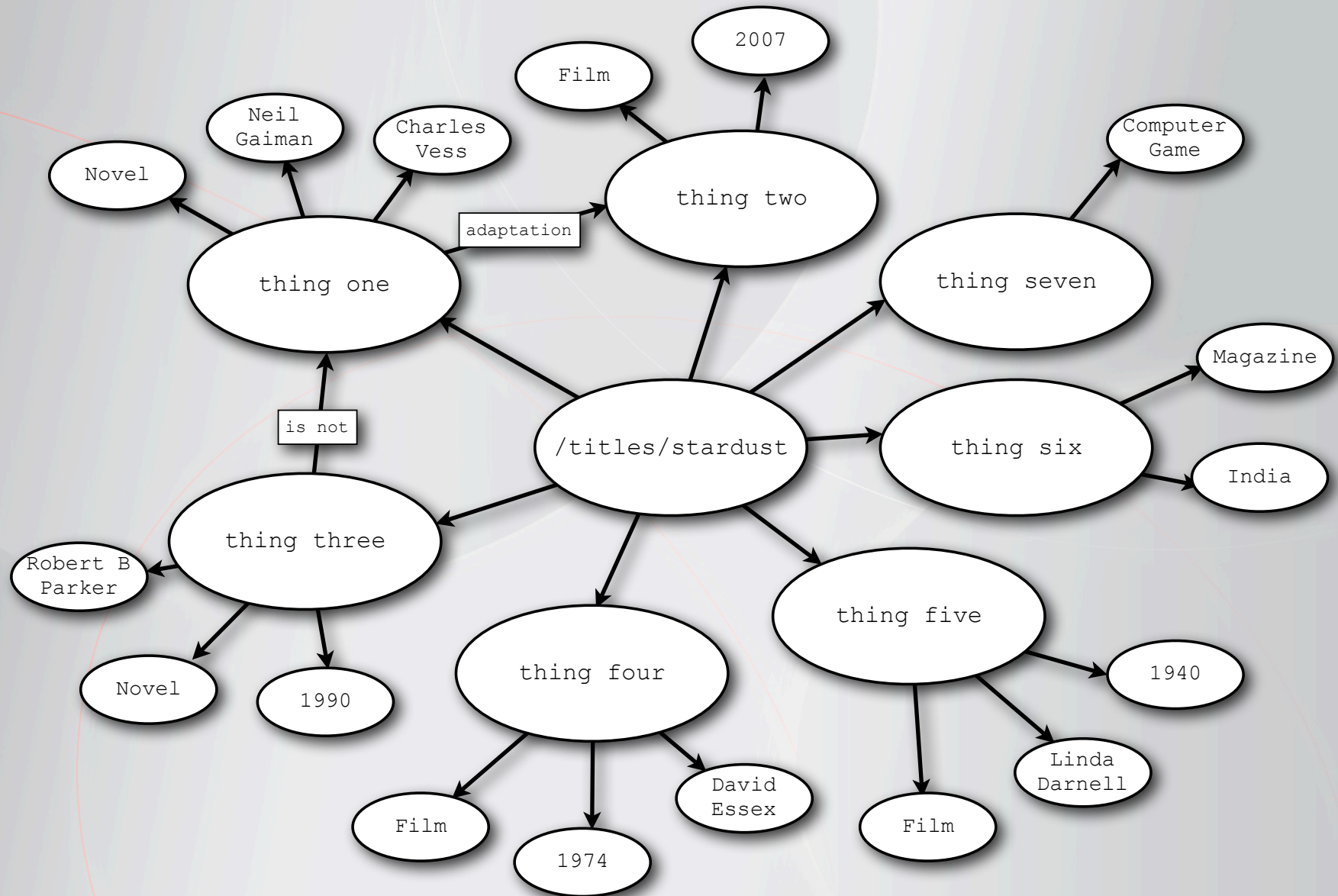
and azkaban harry of potter prisoner the

Example...

Harry Potter and The Prisoner of Azkaban
andazkabanharryofpotterprisonerthe
Prisoner of Azkaban, Harry Potter and The

Example...

/titles/
andazkabanharryofpotterprisonerthe



<http://purl.org/vocab/frbr/core#>



*Functional Analysis of the MARC 21 Bibliographic and Holdings Formats***FRBR Display Tool
Version 2.0**

Network Development and MARC Standards Office
Library of Congress

Contents

- [Download Tool](#) (ZIP file)
- [Introduction](#)
- [Suggested Usage](#)
- [Tool Description](#)
- [How to Use](#)
- [Matching, Sorting and Display Specifications](#)
- [Display Example](#)
- [Full Examples](#)
- [Future Enhancements Under Consideration](#)
- [MARC21FRBR Electronic Discussion List](#)
- [Comments and Suggestions](#)

Introduction

In 2001, the Network Development and MARC Standards Office released the publication, "[Displays for Multiple Versions from MARC 21 and FRBR](#)," which outlined how the [FRBR](#) (Functional Requirements for Bibliographic Records) model can be used to cluster bibliographic records retrieved via a search in more meaningful displays to assist users in selecting items from bibliographic collections. It contained several hierarchical [display examples](#) of bibliographic data using the FRBR model.

The **FRBR Display Tool**, based on the above analysis, is an XSLT program that transforms the bibliographic data found in MARC record retrieval files into meaningful displays by grouping the bibliographic data into the "Work," "Expression" and "Manifestation" FRBR entities. The matching and sorting [specifications](#) for the tool are outlined below.

The **FRBR Display Tool** sorts and arranges bibliographic record sets using the FRBR model. It then generates useful hierarchical displays of these record sets containing works that consist of multiple expressions and manifestations.

The tool is very flexible. Because the tool is written in [XSLT](#), it is easy to augment based on an institution's individual needs. Likewise, the output may be augmented by simply changing the XSL stylesheet that controls display. No change in the XSLT program is needed.

The tool does not search bibliographic databases to create the record set on which it operates. A retrieved file (e.g., an OPAC search result) of MARC unit records must be created before using the tool.

In its current version, the **FRBR Display Tool** works best with record sets resulting from searches of name and title fields. Broader searches (for example, that include data matched in the 5XX note fields) promote less useful display results because the **FRBR Display Tool** does not display the field that caused the retrieval of a record unless that field was one already in the display elements.

One important factor that greatly impacts the usefulness of the **FRBR Display Tool's** results is the consistency of the bibliographic data. Data, for example, with typos or inconsistent headings, lessen the utility of the display because it prevents accurate and consistent collocation of data.

Suggested Usage

The following list indicates a few possible uses of the **FRBR Display Tool**. Please contact the Network Development and MARC Standards Office (ndmso@loc.gov) if you have used it for other purposes and would like to contribute to this list.

1. Test FRBR concepts through experimentation with collocating and sorting files by segmenting MARC 21 records into the FRBR "Works," "Expressions," and "Manifestations" entities.
2. Evaluate consistency and potential of local data for FRBR display.
3. Experiment with an alternative front end display for library catalogs, based on FRBR concepts as a user option.

Matching, Sorting and Display Specifications

The current version of the *FRBR Display Tool* is version 2.0.

The following display table outlines the matching, sorting and displaying processes used in generating the resulting FRBR display. They are given to assist analysis of results when using the tool and to help users determine where they may want to adjust the tool for their individual needs.

See the [display example](#) for further guidance on the display specifications used with the **FRBR Display Tool**.

Work Level

Define work level under: author *and* title

Author:

- **Match:** The following fields in this order: 100\$a\$b\$c\$d (or) 110\$a\$b\$c\$d (or) 111\$a\$c\$d\$n\$g
 - *Ignore:* Extra white space, case, nonfiling characters, brackets, parentheses and all punctuation
- *Sort:* Alphabetically by first sorting character in string
- *Display:* The following fields in this order: **100\$a\$b\$c\$d\$g (or) 110\$a\$b\$c\$d (or) 111\$a\$c\$d\$n\$g**
 - Maintain all punctuation
 - *Display label:* **Author:**

and Title:

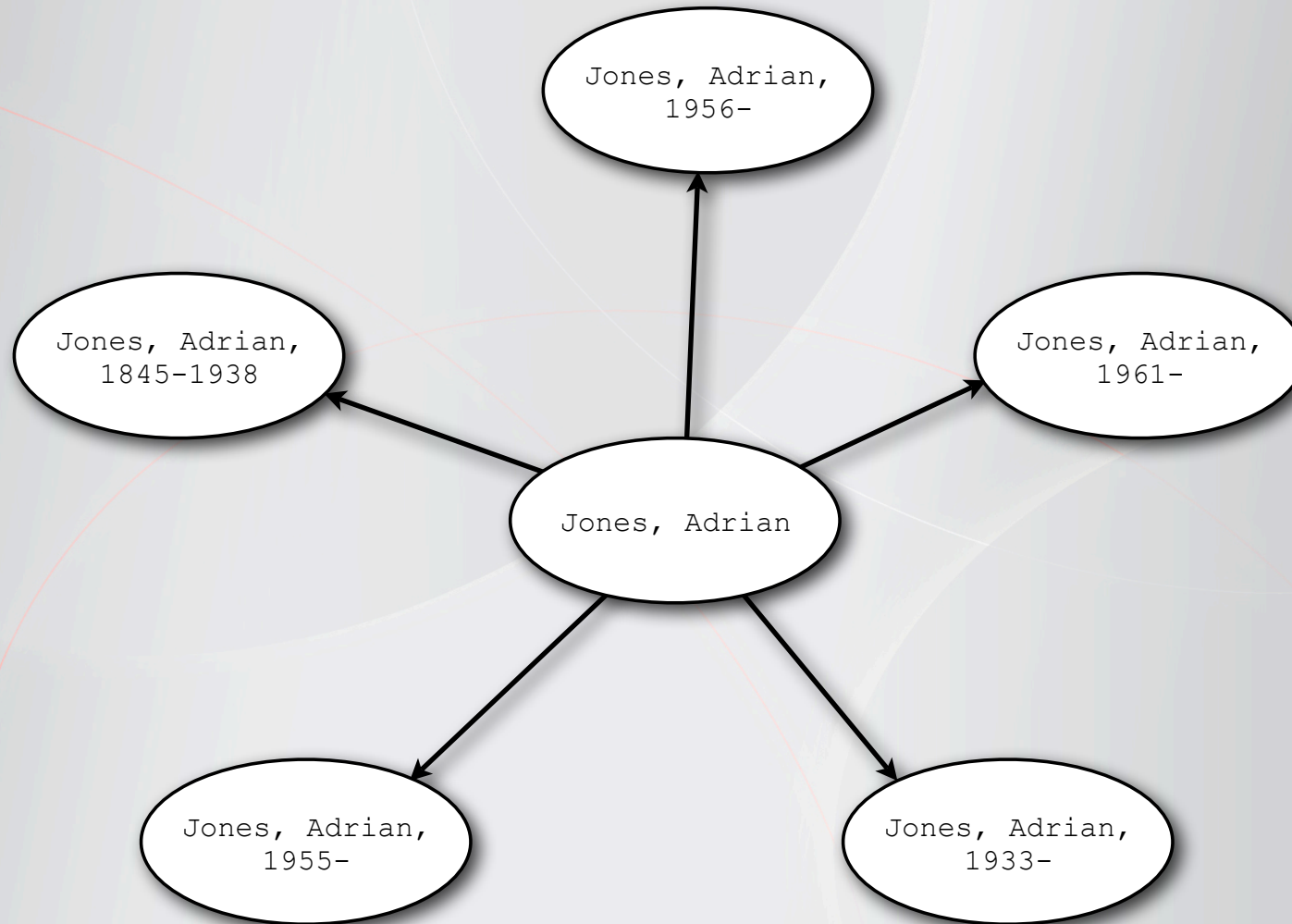
- **Match:** The following fields in this order: 240\$a\$d\$k\$m\$n\$p\$r (or) 243\$a\$d\$m\$n\$p\$r (or) 245\$a\$g\$k\$n\$p
 - *Delete:* Data contained in brackets, along with the brackets
 - *Ignore:* Extra white space, case, nonfiling characters, brackets, parentheses and all punctuation
- *Sort:* Alphabetically by first sorting character in string (beneath the content of the 1XX field)
- *Display:* The following fields in this order: **240\$a\$d\$k\$m\$n\$p\$r (or) 243\$a\$d\$m\$n\$p\$r (or) 245\$a\$g\$k\$n\$p**
 - Maintain all punctuation
 - *Display label:* **Work:**

rowlingjk
/works/
andazkabanharryofpotterprisonerthe

/works/rowlingjkandazkabanharryofpotterprisonerthe

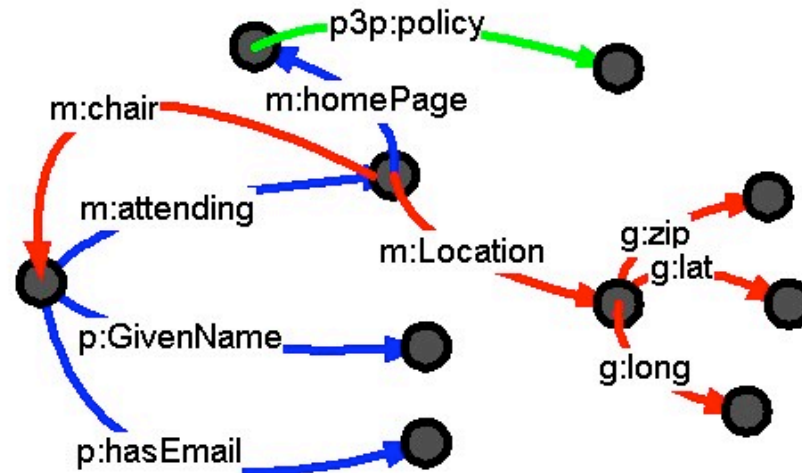
MD5

/works/4e2fc306b548098b8277c07719176998



What's in a name?

...merges just like that.



Subject and object node using same URIs

Conclusions

Pre-Publication DRAFT February 2008, submitted to <http://events.linkedata.org/ldow2008/>

SEMANTIC MARC, MARC21 AND THE SEMANTIC WEB

Rob Styles
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
rob.styles@talis.com

Danny Ayers
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
danny.ayers@talis.com

Nadeem Shabir
Talis
Knight's Court, Solihull Parkway
Birmingham, B37 7YB
+44 (0) 870 400 5000
nadeem.shabir@talis.com

ABSTRACT

The MARC standard for exchanging bibliographic data has been in use for several decades and is used by major libraries worldwide. This paper discusses the possibilities of representing the most prevalent form of MARC, MARC21, as RDF for the Semantic Web, and aims to understand the tradeoffs, if any, resulting from transforming the data. Critically our approach goes beyond a simple transformation of the MARC21 record content to develop rich semantic descriptions of the varied things which may be described using bibliographic records. We present an algorithmic approach for consistently generating URIs from critical data, discuss the algorithmic matching of author names and suggest how RDF generated from MARC records may be linked to other data sources on the Web.

Keywords

MARC, MARC21, RDF, Semantic Web, Data Conversion, Inferred Semantics.

1. INTRODUCTION

A great deal of data exists as strings of text in structured form within binary file formats. In many cases all the data is in one file (e.g. MP3) or all the data is in one file (e.g. MP3). A more complex variation is the bibliographic data created by the last work of generation of libraries, putting at that data in the purpose of this paper. The principles described here, though, are equally applicable to any form of data where humans are left to infer meaning from literal strings.

Copyright notice to go here.

The MARC standard for exchanging data has been around for more than 50 years. It is a structured binary format that has allowed libraries to exchange bibliographic data very successfully. So successfully, in fact, that the Library of Congress and British Library have around 10 million records in this form each. Most national libraries have a similar number. OCLC Worldcat, a US database of library information has many tens of millions. The data is not readily available for reuse outside of the library community. Talis has, for more than 40 years, maintained a database of such bibliographic records currently numbering in the tens of millions, a mixture of contributed data from libraries and commercial data from suppliers.

The Semantic Web, a web of data linked through the use of URIs and accessible over HTTP, offers the opportunity to create large, interconnected sets of data.

This paper aims to discuss the possibilities of representing the most prevalent form of MARC, MARC21, as RDF for the Semantic Web.

2. MARC21

MARC21 is used to describe several different types of record in library catalogues. Bibliographic records describe publications, authority records for the names of authors, names, titles or subject headings. All of the major library management systems in use in English-speaking countries are able to import and export data in this form.

There are other flavours of MARC: Unimarc, UNIMARC and UNIMARC are just some examples. The different MARC standards all share an underlying record system. UNIMARC has very much in common with the other MARC standards. They differ in the level of granularity at which they store data, a single name field versus separate first and surname being one example, and also in where they locate data within a record – that is what meaning is assigned to each position.

Given an increasing volume of online knowledge due to their massive digitisation projects, use a mixture of MARC21, UNIMARC and UNIMARC. With the volume of data available in MARC21 and the global connectivity provided by the internet, MARC21 is rapidly becoming the lingua franca for libraries globally. The techniques described in this paper are equally applicable to all flavours of MARC as well as other data formats.

Pre-Publication DRAFT February 2008, submitted to <http://events.linkedata.org/ldow2008/>

<http://events.linkedata.org/ldow2008/#program>



Rob Styles

rob.styles@talis.com
aka mmmmmRob
irc.freenode.net #talis



Nadeem Shabir

nadeem.shabir@talis.com
aka KiYanWang
irc.freenode.net #talis



Danny Ayers

danny.ayers@talis.com
aka danja
irc.freenode.net #talis

Nodalities
THE MAGAZINE OF THE SEMANTIC WEB
www.talis.com/nodalities
April 2008

INSIDE

- 1 WWW2008**
Tim Heath talks at the importance of Linked Data, and outlines key issues for discussion in Beijing.
- 4 THE VALUE OF WEB 3.0**
Mike Orlitzky of Protoprise asks how Web 3.0 differs from the previous phases of the Web.
- 6 PODCAST ROUNDUP**
News of the latest Semantic Web podcasts from Talis.
- 7 TALIS ENGAGE**
Nadine Smith reports on development work that delivered a commercially viable enterprise application with the help of the Talis Platform.
- 9 SIR TIM BERNERS-LEE TALKS WITH TALIS ABOUT THE SEMANTIC WEB**
The full transcript of this recent interview with the Father of the Web.
- 19 THE SEMANTIC WEB GANG**
Experienced Semantic Web practitioners debate the readiness of this technology for mainstream corporate adoption.

Looking ahead to Linked Data on the Web

Abstract
Image courtesy: Richard Cyganski

Ahead of the Linked Data on the Web workshop at this year's World Wide Web Conference in Beijing, Tim Heath introduces the concept of Linked Data and outlines some of the key issues that workshop participants will be discussing.

Linked Data is a style of publishing data on the Web that emphasises data reuse and connections between related data sources. To understand the value of this proposition, imagine a traditional Web site with no incoming or outgoing links. How much value would such a Web site generate for its owner, compared to another that was identical in content but richly interconnected with the Web at large?

continued on page 2

<http://blogs.talis.com/nodalities>



This work is Copyright © 2008 Talis Group Limited.
It is licensed under the Creative Commons Attribution 3.0 Unported License
Full details at: <http://creativecommons.org/licenses/by/3.0/>

You are free:



to Share — to copy, distribute and transmit the work



to Remix — to adapt the work

Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- Nothing in this license impairs or restricts the author's moral rights.
- Some Content in the work may be licensed under different terms, this is noted separately.



shared innovation™