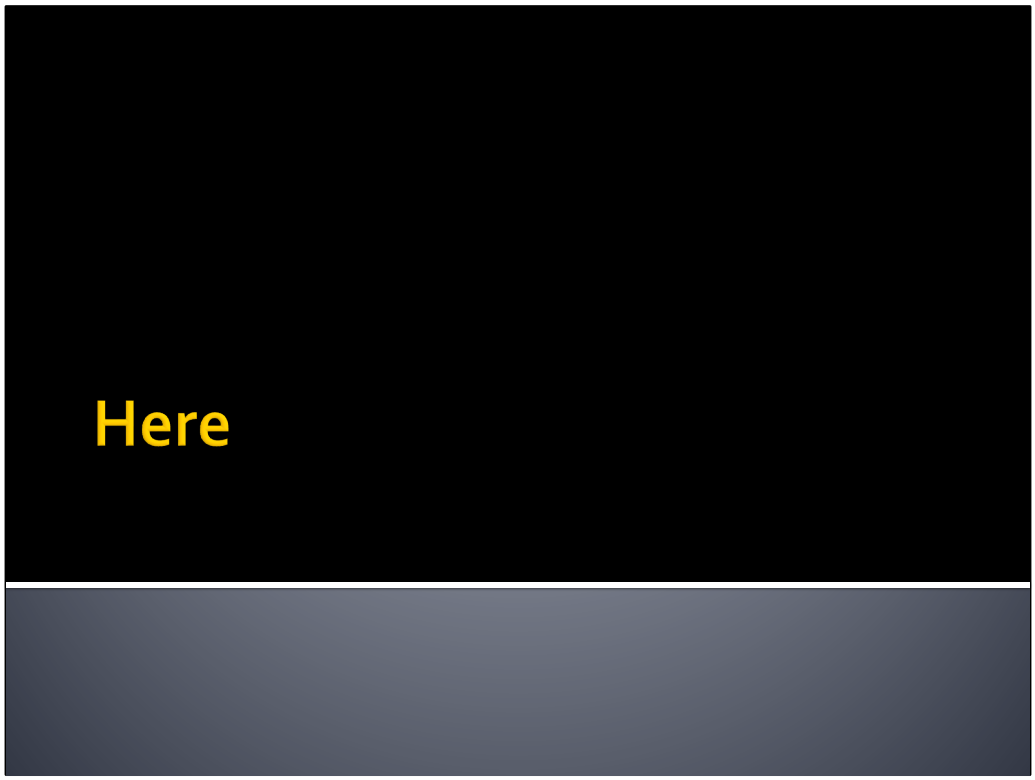
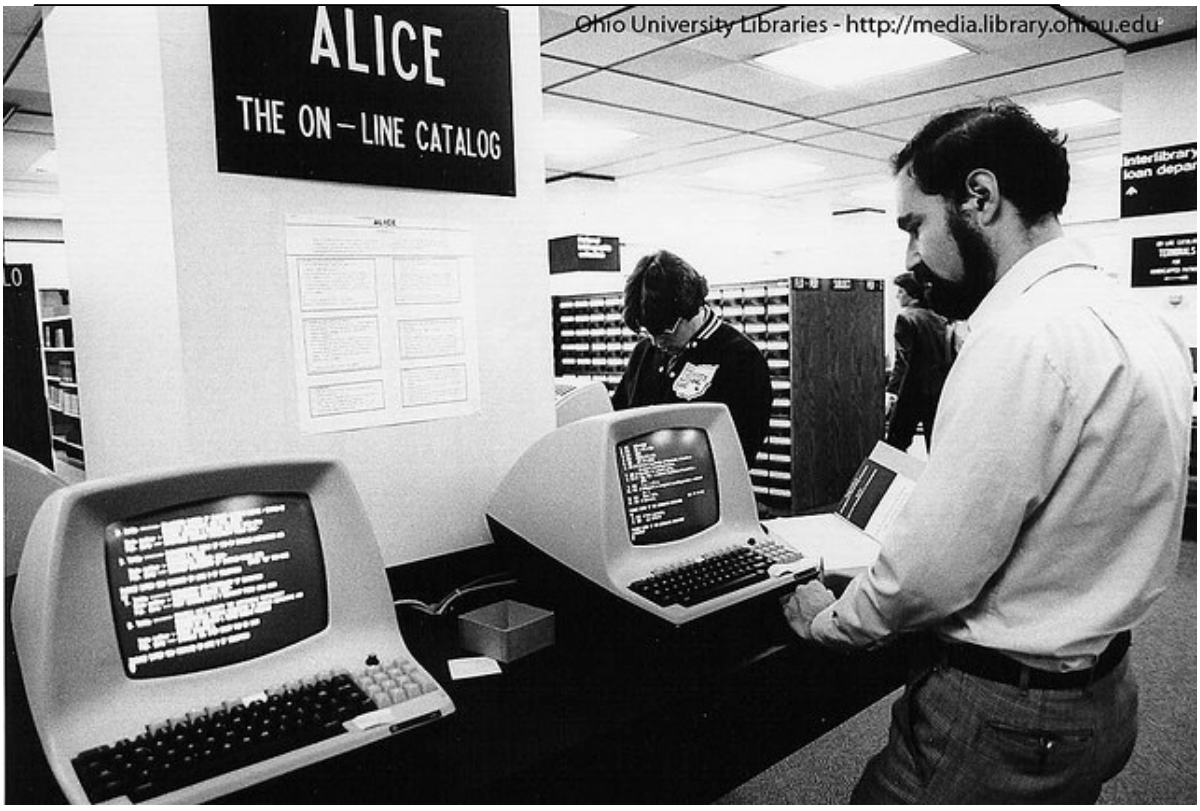
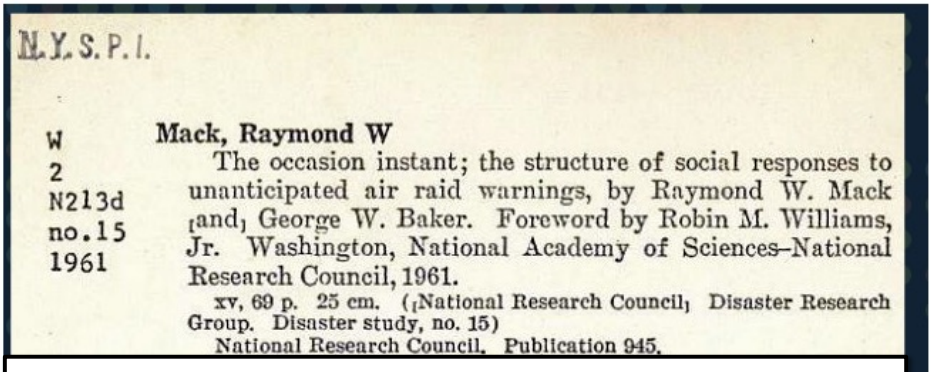
NISO/DCMI Webinar
February, 2012
Karen Coyle


From Here to There
Megan Brown
Beckut83@hotmail.it


Here


ALICE
THE ON–LINE CATALOG
Ohio University Libraries - http://media.library.ohiou.edu



N.Y.S.P.I.

W
2
N213d
no.15
1961

Mack, Raymond W

The occasion instant; the structure of social responses to unanticipated air raid warnings, by Raymond W. Mack [and] George W. Baker. Foreword by Robin M. Williams, Jr. Washington, National Academy of Sciences–National Research Council, 1961.

xv, 69 p. 25 cm. ([National Research Council] Disaster Research Group. Disaster study, no. 15)

National Research Council. Publication 945.

LC control no.: 61064605

LCCN permalink: http://lccn.loc.gov/61064605

Type of material: Book (Print, Microform, Electronic, etc.)

Personal name: Mack, Raymond W.

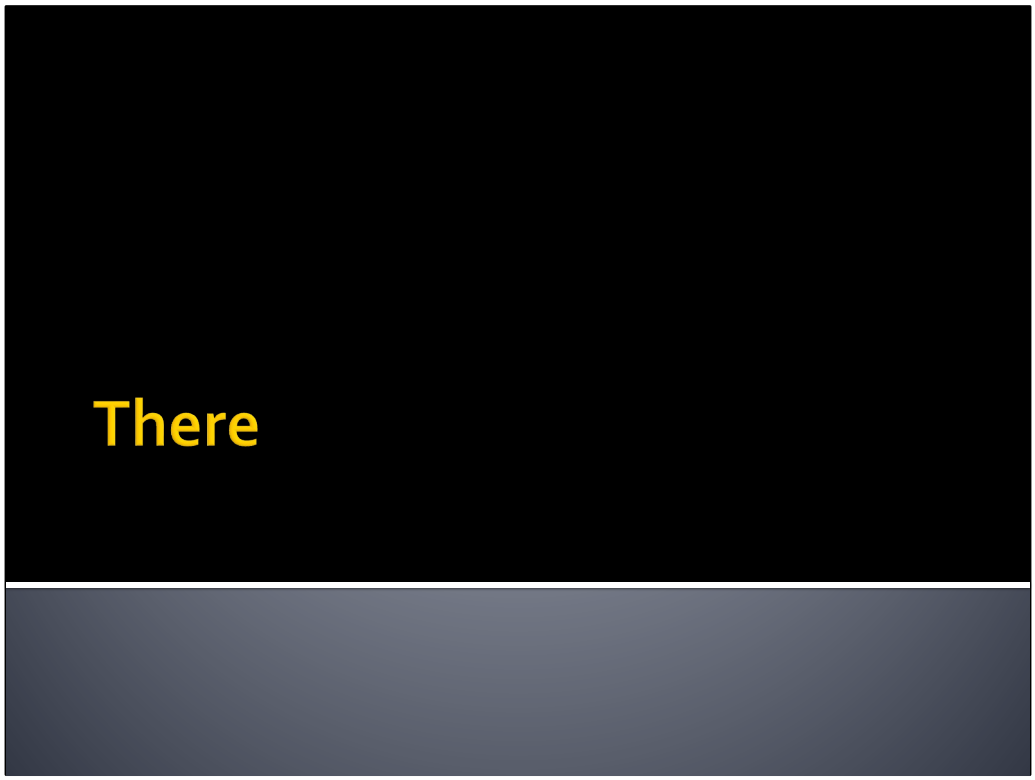
Main title: The occasion instant : the structure of social responses to unanticipated air raid warnings / by Raymond W. Mack, George W. Baker ; Foreword by Robin M. Williams, Jr.

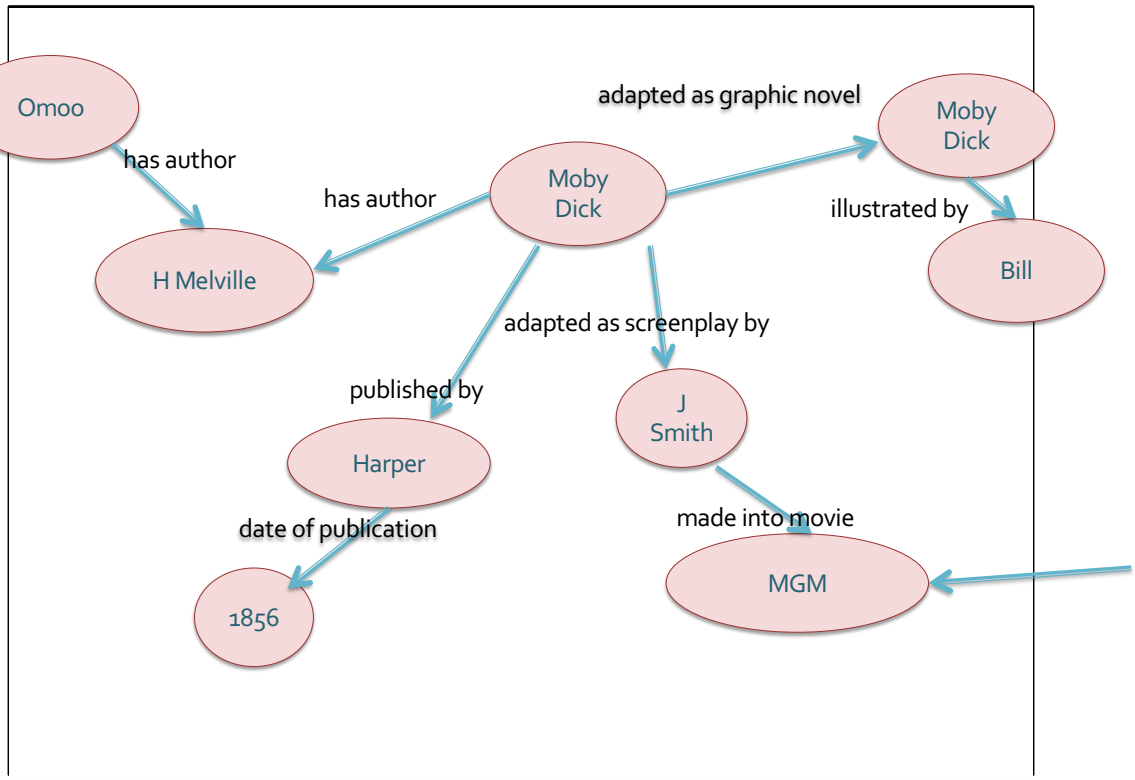
Published/Created: Washington, National Academy of Sciences-National Research Council, 1961.

Description: xv, 69 p. ; 25 cm.

Links: Table of contents

CALL NUMBER: UA927 .M23

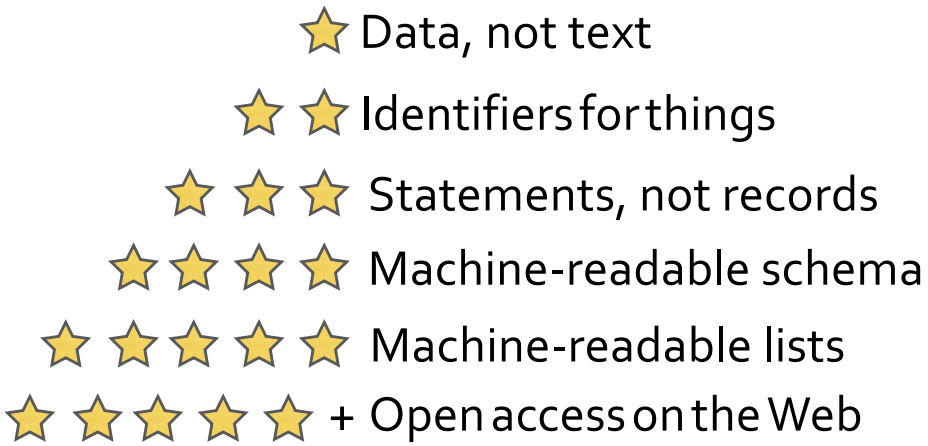

There



Mostly what we do today, but using a different technology, a different data model.


LINKED DATA
On the web, open licens
Machine-readable data
Non-proprietary format
RDF standards
Linked RDF
YOUR DATA 5 ★?

Linked Data


Data, not text
Identifiers for things
Statements, not records
Machine-readable schema
Machine-readable lists
+ Open access on the Web

Data

Data Text

It is a Law of Nature with us that
side than his father, so that each gener
in the scale of development and no
is a Pentagon; the son of a Pentagon,
But this rule applies not alway
often to the Soldiers, and to the W
said to deserve the name of human F
sides equal. With them therefore t
and the son of an Isosceles (*i.e.* a Tria
Isosceles still. Nevertheless, all hop
Isosceles, that his posterity may ul
condition. For, after a long series
and skilful labours, it is generally fou
the Artisan and Soldier classes manife
or base, and a shrinkage of the two ot
by the Priests) between the sons and
members of the lower classes genera
mating still more to the type of the

Once text becomes digitized, it is ones and zeroes. Is it data? Not in the way that I'm using the term data.

Data	Text
▪ Machine-readable ▪ Machine-actionable ▪ Computable	▪ Human readable ▪ Natural language ▪ Ambiguous

TEXT

- 020 $a 0439064864 (hardcover)
- Concertos (Violin)
  Concertos (Violins (2))
- 300 $c 23 cm.

DATA

- ISBN:0439064864
- instrument: [violin]
  number: 1
  instrument: [violin]
  number: 2
- height: 23
  unit: [centimeters]

The Physical Object

What sort of book is it? Paperback; Hardcover, etc.

How many pages?

211

Pagination? Note the highest number in each pagination pattern.

x, 211 p. ;

For example: *xii, 346p.*

How much does the book weigh?

( ) grams (•) kilos ( ) ounces ( ) pounds

Dimensions

Height:

Width:

Depth:

( ) centimeters (•) inches

Data IS the information

```
008   711222s1961 dcu b 000 0 eng
```

260 __ $a Washington, $b National Academy of Sciences-National Research Council, $c 1961.

We tend to think of the text of the record as the information, but in a machine-processing world, the information that you can make use of is in the data portion of the record. The text is only usable AFTER the item has been retrieved, sorted, faceted, etc.

humans may look at one or two pages of display. if you want to make sense of a retrieved set of 1K items, you are going to have to computer to do so

catalogoing rules without data is like designing the body of a new car and forgetting to put in the engine

Identifiers

dog = 

We humans have a great tolerance for language ambiguity. We know if we are talking about a dog, or a dog, usually by the context. The proof that there is ambiguity is that we have puns, which couldn't exist without it.

Identifiers must be precise

IDa87nn3

66.ah.773m

Machines are really, really stupid. They need a separate identity for each thing, no ambiguity allowed.

Identifiers don't change

Smith, John, 1946-

Smith, John, 1946-2006

Change the identifier, change the identity

There are a lot of rules that pertain to identifiers, but one of the key ones is that identifiers do not change. If you change the identifier, you have changed the identity. It becomes a different thing This is why you cannot use display text to identify things, because sometimes displays need to change.

In libraries, we have used display text as identifiers in a very clever way. This was a smart thing to do in the pre-computer environment, but it no longer is a valid practice.

Internationalization (i18n)

dog (lang=en)

hund (lang=de)

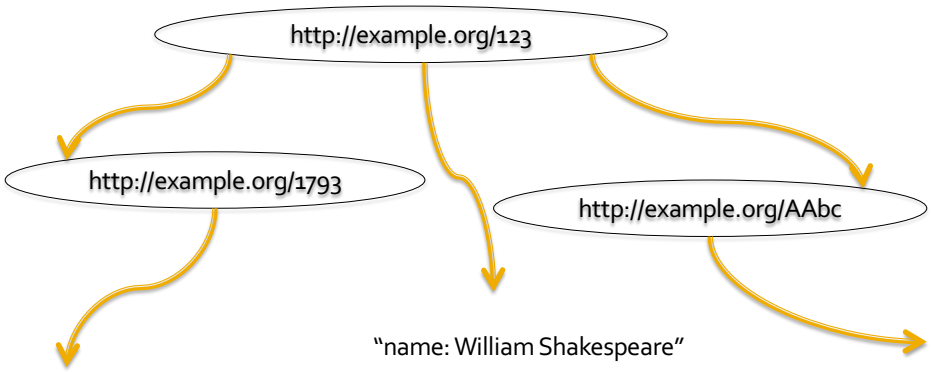
perro (lang=sp)

chien (lang=fr)

A big advantage of identifiers, not display: internationalization

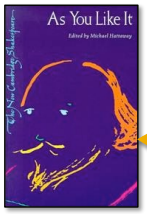
identifiers for machines – display in any language

Strings are endpoints
Things can connect to other things



Statements

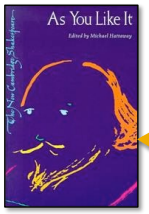
A record



	|a 822.3/3 |2 22
1001	|a Shakespeare, William, |d 1564-1616.
24510	|a As you like it / |c edited by Michael Hattaway.
250	|a Updated ed.
260	|a Cambridge, UK ; |a New York : |b Cambridge University Press, |c 2009.
300	|a xv, 240 p. : |b ill. ; |c 24 cm.
4901	|a New Cambridge Shakespeare
504	|a Includes bibliographical references (p. 240).
5050	|a Introduction. Journeys ; Plays within the play ; Theatrical genres ; Pastoral ; Counter-pastoral ; The condition of the country ; Politics ; 'Between you and the women the play must please' ; Gender ; Nuptials ; Sources ; Date and occassion ; Stage history ; Recent critical and stage interpretations -- Note on the text -- List of characters -- The play -- Textual analysis -- Appendixes : 1. An early court performance? ; 2. Extracts from Shakespeare's principal source, Lodge's Rosalind ; 3. The songs.
650 0	|a Fathers and daughters |v Drama.
650 0	|a Exiles |v Drama.
60010	|a Shakespeare, William, |d 1564-1616. |t As you like it.
655 7	|a Pastoral drama. |2 gsafd

We know these talk about the same thing because they are in the same record.

A record

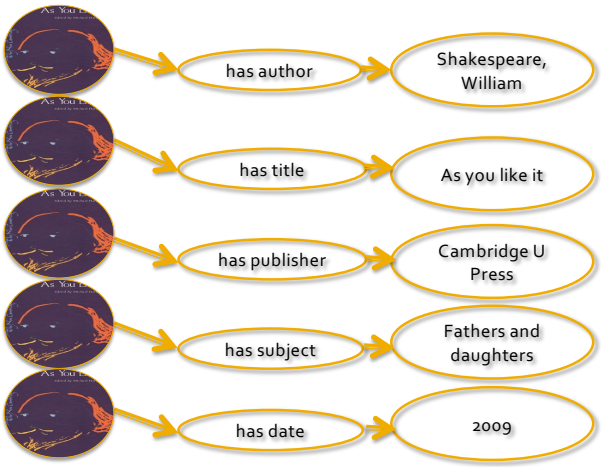


a 822.3/3 |2 22
001 |a Shakespeare, William, |d 1564-1616.
24510 |a As you like it / |c edited by Michael Hattaway.
250 |a Updated ed.
260 |a Cambridge, UK ; |a New York : |b Cambridge University Press, |c 2009.
300 |a xv, 240 p. : |b ill. ; |c 24 cm.
4901 |a New Cambridge Shakespeare
504 |a Includes bibliographical references (p. 240).
5050 |a Introduction. Journeys ; Plays within the play ; Theatrical genres ; Pastoral ; Counter-pastoral ; The condition of the country ; Politics ; 'Between you and the women the play must please' ; Gender ; Nuptials ; Sources ; Date and occassion ; Stage history ; Recent critical and stage interpretations -- Note on the text -- List of characters -- The play -- Textual analysis -- Appendixes : 1. An early court performance ? 2. Extracts from Shakespeare's principal source, Lodge's Rosalynde ; 3. The songs.
650 0 |a Exiles |v Drama.
60010 |a Shakespeare, William, |d 1564-1616. |t As you like it.
7 |a Pastoral drama. |2 gsafd

100 $a Shakespeare, William

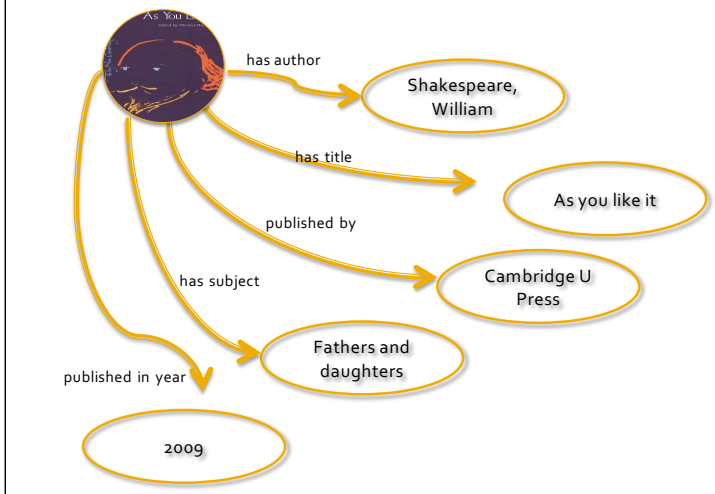
If we take a single field or statement out of the record, it is no longer meaningful because we don't know what it refers to – we don't know what it is talking about. But inside the record you can't do anything with it – it doesn't interact with any other data.

Statements



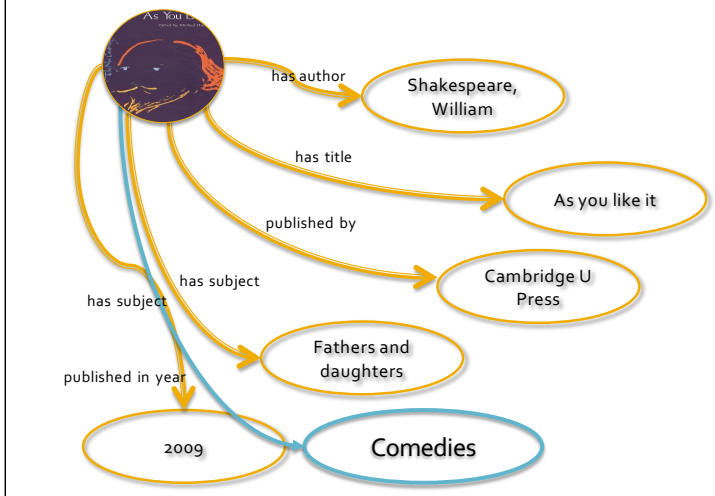
In the idea of linked data, each “bit” of data is a statement that includes what it is talking about. So instead of a record with data, we have data statements that are like a sentence. Each one has a subject, the oval on the left, and then what it is saying about the subject. Now you can use them in different combinations and you know that they are information about the book (the subject).

Graphs



Relationships with centers

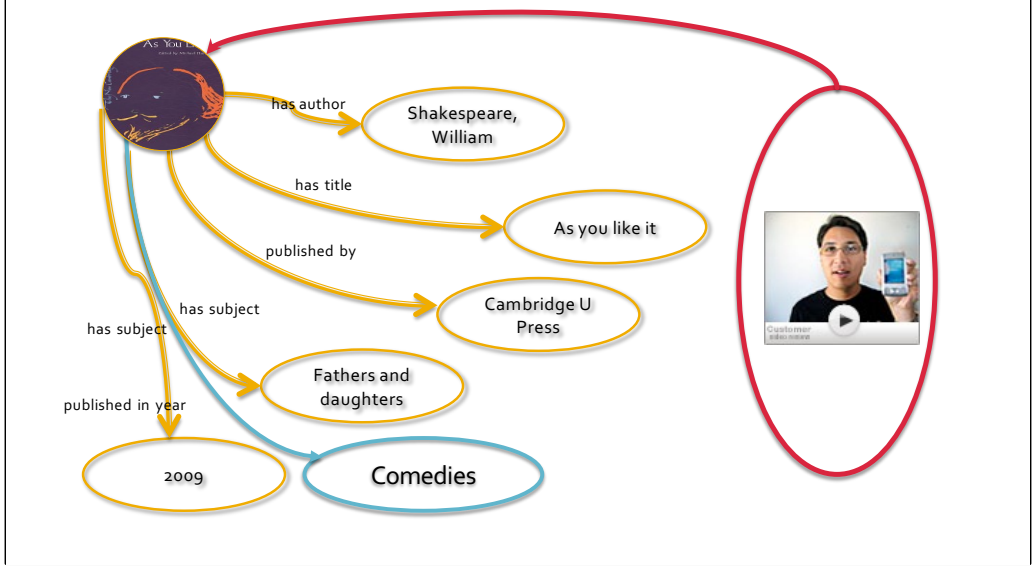
Advantages: flexibility



This format has many advantages. It is easy to add new information because you just add another triple and it does not disturb the data that is already there.

Even add new data types without re-writing your entire set of data.

Advantages: flexibility



The information that you add can be in many different formats. You are free to ignore the ones you do not want or cannot understand. There is ongoing work to make sure that you can know where the information linked to a subject has come from, so you can choose based on who has provided the information, and even how old or new it is.

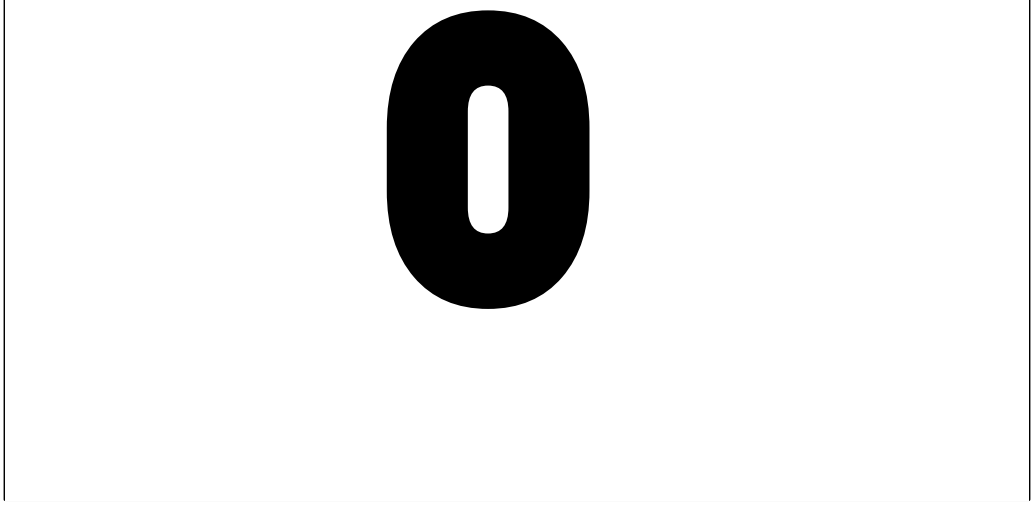
This is a big change, and it isn't just a matter of taking the records we have today and breaking them up into triples. If we want the data to be usable, there is much more to it.

Things (entities) and the properties of those things. What we call “data elements” in more traditional metadata are called “properties” in Semantic Web metadata. These are also what are called “attributes” in the FRBR model.

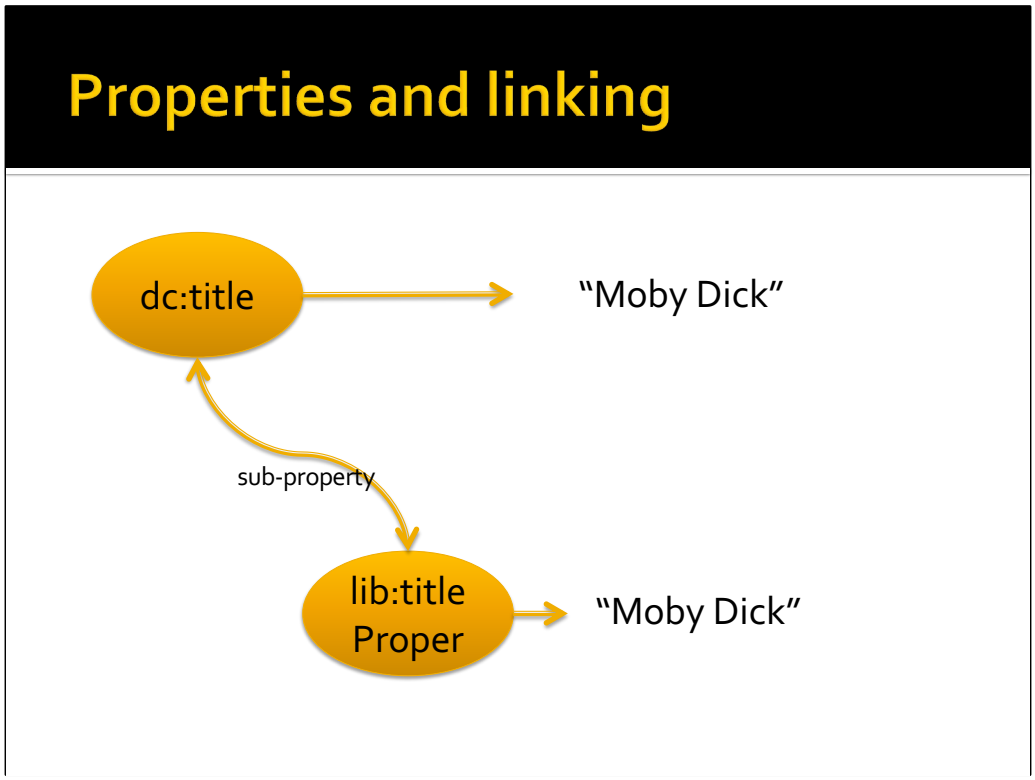
What are the things in our metadata?

- people, corporate bodies, families
- places (as subjects, as locations, and other)
- events (conferences, legislative actions)
- topics (classification, subject headings)
- resources (books, sound recordings; works, expressions)
- physical formats, extents
- …

How many are unique to library data?

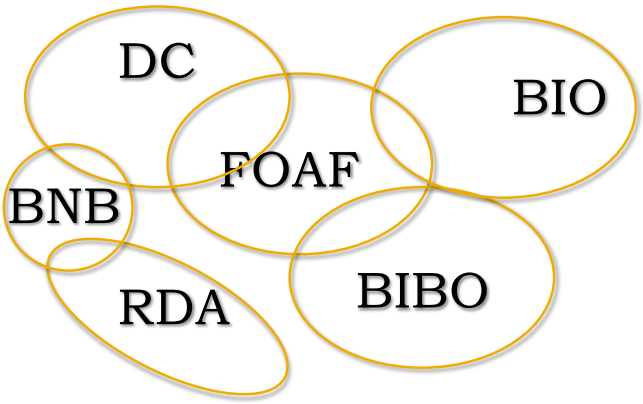


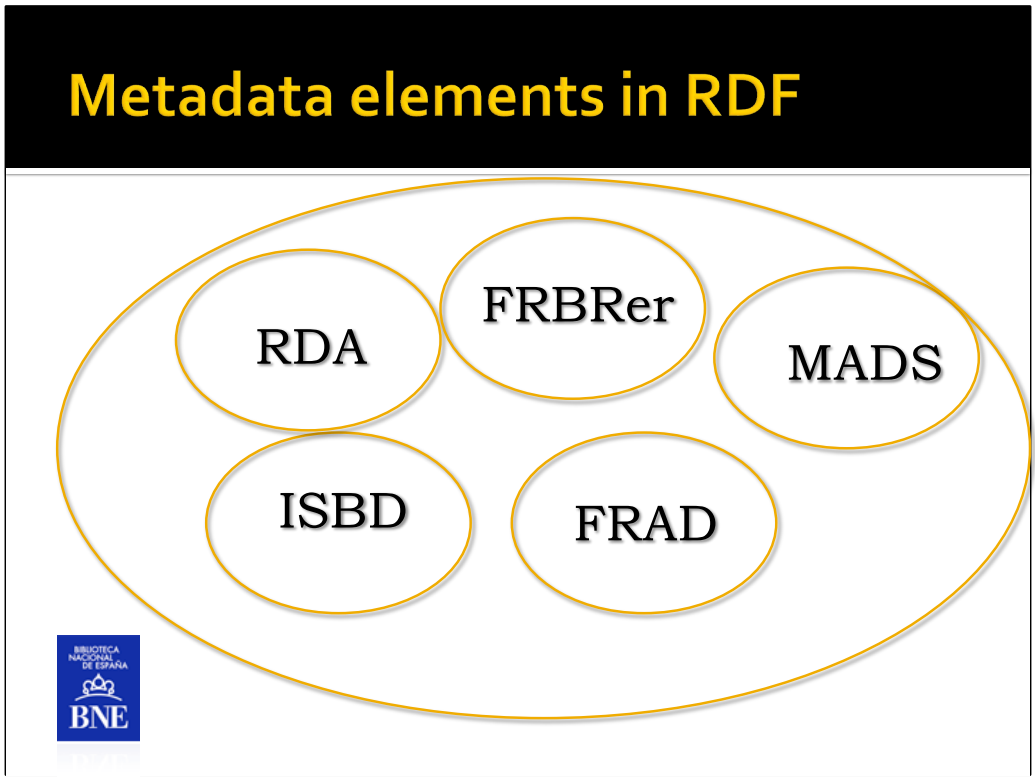
This is a big plus! Great opportunity for sharing. Our data should connect to the data created by others because we have the same things.



How do you make sure that your metadata will link? You can use existing and common properties that will be used by others, such as the dublin core "title" property. However, you should only use that property if it perfectly suits your need. DC title is defined as "the name of the resource". That may not

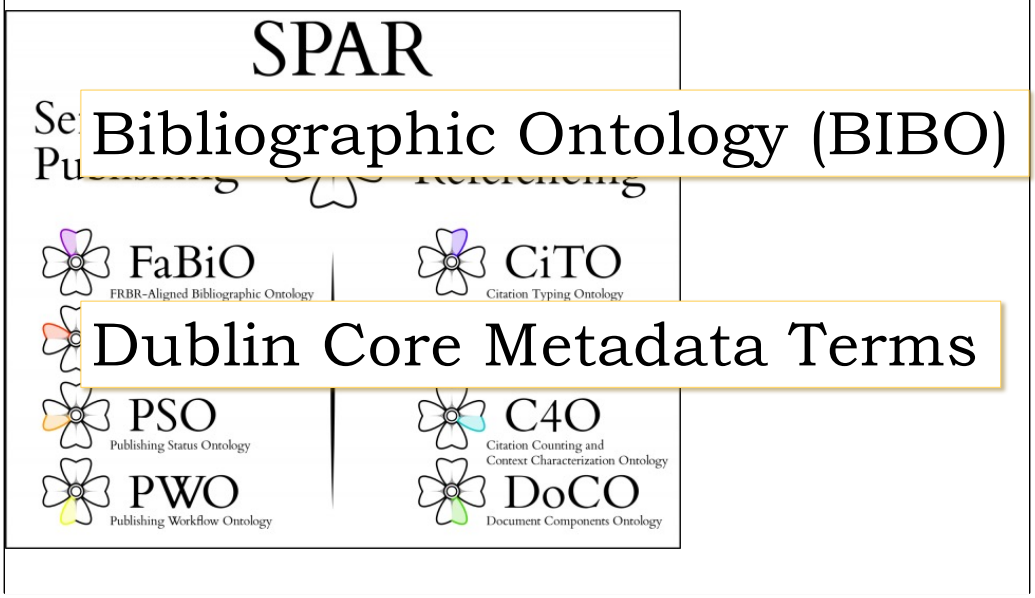
British National Bibliography in RDF





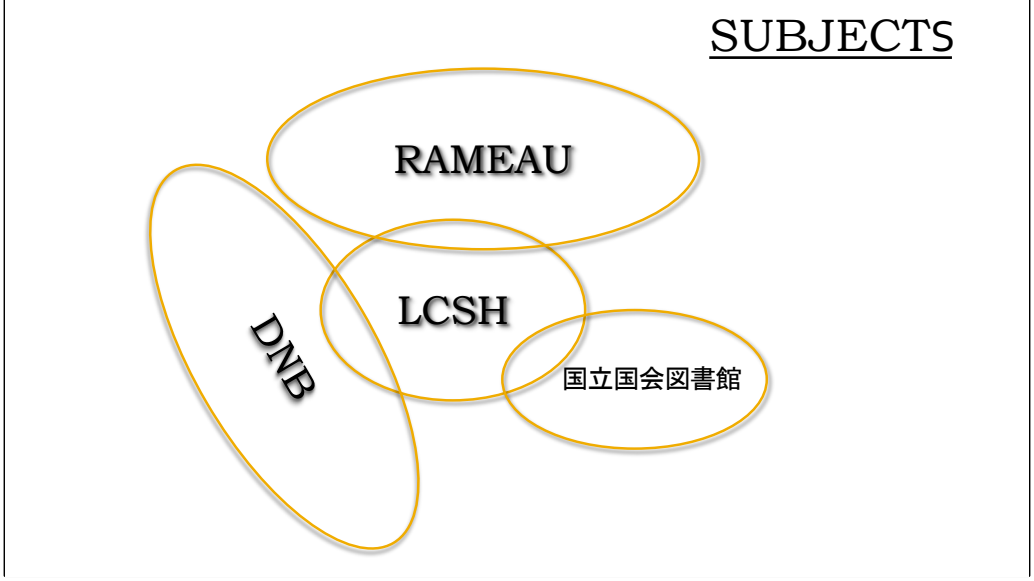
RDF silos. It isn't enough to take the records we have today and "translate" them into RDF. It requires a re-thinking of how we imagine our data.

Linked bibliographic metadata

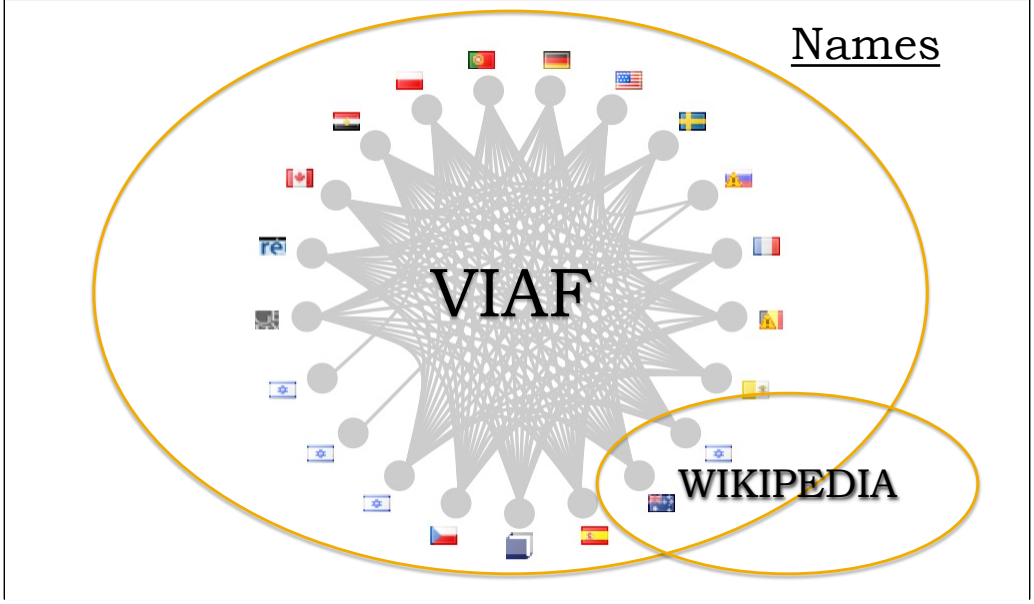


Controlled vocabularies

Controlled vocabularies & authorities



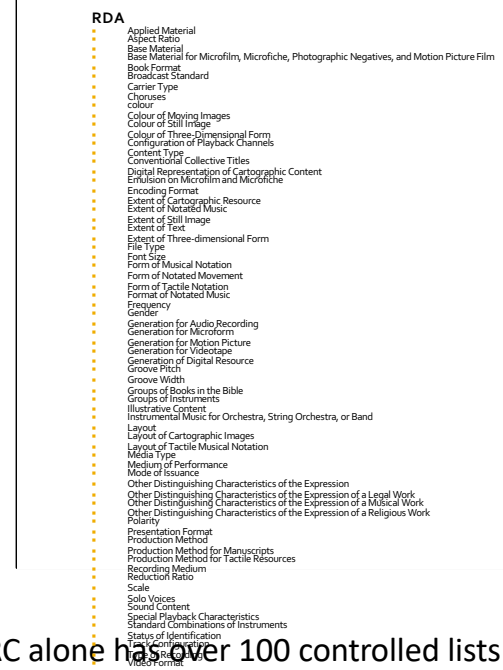
Controlled vocabularies & authorities



VIAF gathers name authority files from about 20 national libraries. Has about 20 million “triples”

”links to Wikipedia (which links to DBPedia)

Lists, lists, lists



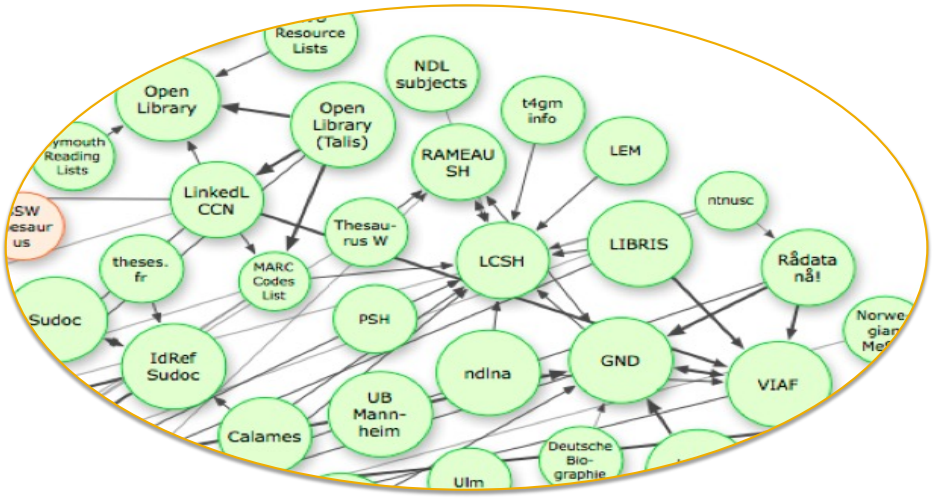
MARC
- Genre/form
- Relators
- Countries
- Geographic Areas
- Languages

PREMIS
- Pres. events
- Pres. level role

MARC alone has over 100 controlled lists of terms. RDA has about 70. These cover everything from tape track formats to music terms to languages and geographic codes. MARC lists are being developed in SW format by LoC, at the site id.loc.gov. There are about a half a dozen now, plus some lists from the PREMIS standard. RDA lists are at rdvocab.info – these links are all available on a page that I'll give you at the end.

Open and on the Web

Open access on the Web



Not *all* of your data – you can choose what to expose. Obviously, we won't put patron data out on the Web as open access. But anything you want to link has to be open.

So far a small number. Many more authority files. harder to 1) keep up to date 2) larger in most instances 3) less clear what to do with them.

We're at a tipping point – we've beaten the dead horse of MARC for a decade, insisting that it get up and live again. Finally, we appear to have collectively faced the reality that the MARC horse is indeed dead, and it is time to move on.

Moving on, however, isn't all that easy. We have a huge installed base of systems that only know about MARC records; many thousands of practitioners who have only ever cataloged into MARC records; and we're in the middle of a world financial crisis. These aren't positives.

We also just spent 10 years and all of our available capital (both in dollars and human resources) creating the Framily and RDA. Unfortunately, I don't think that we're going to get very far with them – they are already obsolete technology. This isn't good news.

But I'm not ready to give up. Otherwise I wouldn't be here. There are things that we can do, even though they will be somewhat painful, especially for those who feel strongly about the traditional cataloging culture.

We need to finally get over the card. Contrived headings and alphabetical order just don't have the role they once had.

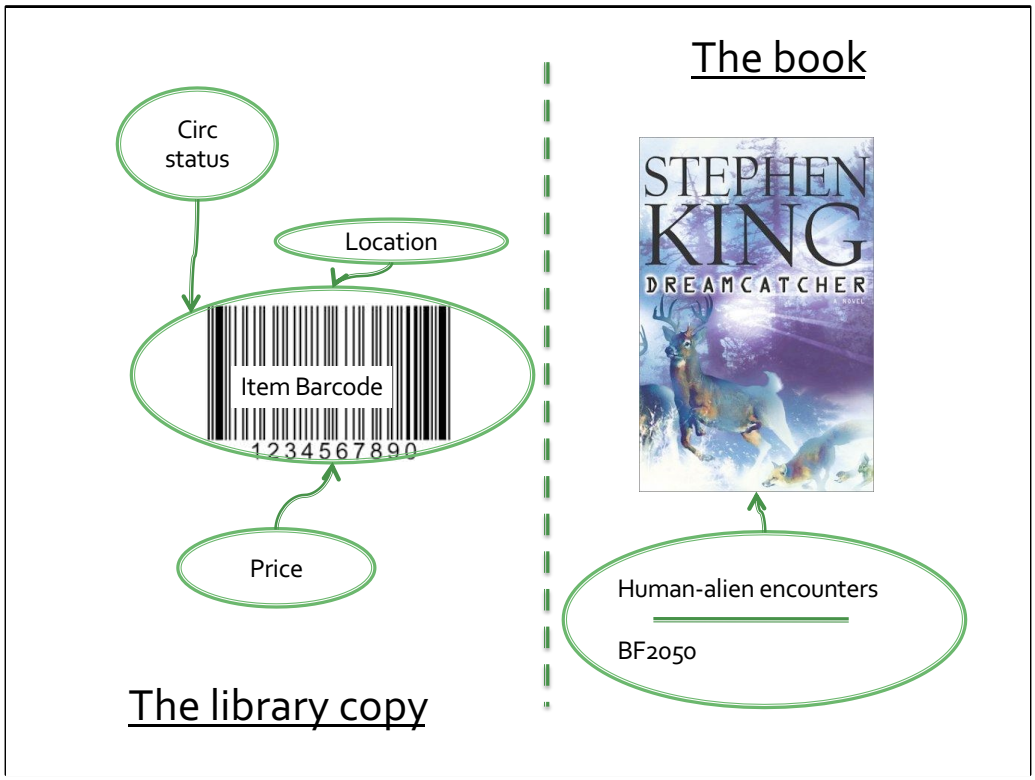
We probably have to give up the artisanal approach to cataloging: the great effort spent over commas and whether or not to use abbreviations.

We need to recognize that users need more than a bibliographic description in order to make their selection from a wide variety of resources.

In all of things, linked data can help us.

First, we need to step back and take a new look at our data. In a sense, FRBR did this, and did it well. But FRBR did it over 10 years ago based on technology that was already decades old, the relational database.

We also need to look at what it is that libraries have that no one else can provide, or provide as well as we can. It seems to me that there are two areas where libraries have extra value.



Next, we have to look at what libraries have to offer that no one else on the cloud can provide. Lots of folks have bibliographic data – even in large quantities: Amazon, Google, Barnes and Noble, article databases, repositories, publisher sites,

We do have things that we can offer uniquely that are of great value. Among these are:

1) library holdings – the link between a bibliographic description online and an actual location where the person can access the item. This isn't needed for digital materials, but there is a lot that isn't digital, or things that aren't easily used in digital form. There really isn't much use getting library data online if it doesn't lead back to the library, and it's the holdings data that leads back.
2) Classification and subject headings. Others have some subject access, although mostly keyword. Amazon uses BISAC headings from publishers. for better or worse, only libraries provide LCSH. No one else has actually classifications, anything like DDC or LCC. These are highly under-utilized even in libraries, where they mainly serve to assign a shelf location. But the structure of classification, combined with linking technology, could allow some very interesting navigation.

Do not look on LD as another way to code our data for machines, but as a new way of thinking about data. We need to get our heads around the idea of library data being not only *in* the Web of *of* the Web.