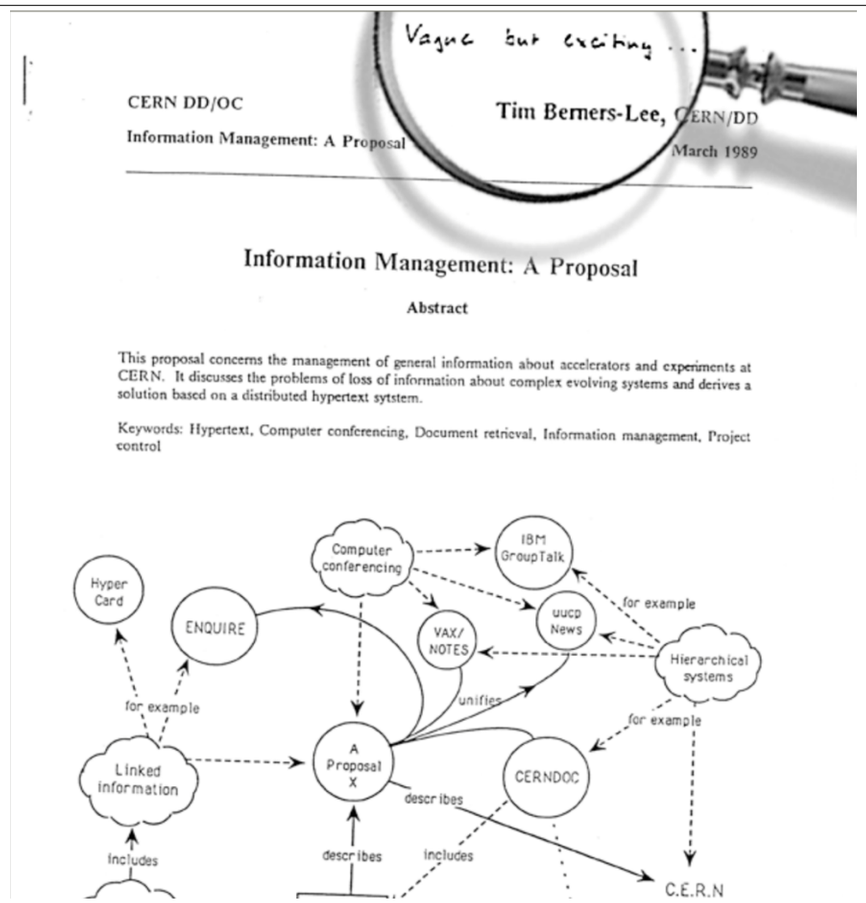
Data on the Web

Owen Stephens
British Library

Using these slides

These slides were developed by Owen Stephens (owen@ostephens.com) on behalf of the British Library.

It is suggested when crediting this work, you include the phrase "Developed by Owen Stephens on behalf of the British Library"



http://info.cern.ch/Proposal.html

Original proposal at CERN – 1989
Followed up by a more concrete proposal in 1990 which scoped down considerably – talks about ‘hypertext pages’
What is ‘the web’ – at it’s heart a set of protocols, data standards and identifiers

URIs

URIs vs URLs

HTTP

Methods: GET, HEAD, POST, PUT
Status: 200, 404, 3XX

HTML

Markup for ‘hypertext’
Document focussed

HTML elements

<html>

<head>

<body>

<h1><h2><h3>...

<p>

...

Document focussed – describe structure in terms of document parts – headings; paragraphs; images. Also a few more structural things – tables; lists.

```
<html>
<body>

  <h1>A Tale of Two Cities</h1>
  <h2>Book the First—Recalled to Life</h2>
  <h3>I. The Period</h3>

  <p>It was the best of times, it was the worst of
  times ...</p>

</body>
</html>
```

What are these three headings? Title? Book? Title?
Humans might have difficulty. Machines have no chance because next time it could all be different

Structured Documents

```
<dl xmlns:lndl="http://copac.ac.uk/schemas/library.name.display.lookup"
class="bibliographicdetails">
  <dt class="title">Title</dt>
    <dd class="title"><h2>
      <a href="/search?title=Winnie-the-Pooh%20story
      %20books">Winnie-the-Pooh story books</a> / A.A. Milne ;
      illustrated by Ernest H. Shepard.</h2></dd>
  <dt>Author</dt>
    <dd class="notlistview">
      <a href="/search?author=Milne,%20A.%20A.%20(Alan
      %20Alexander),">Milne, A. A. (Alan Alexander),</a> 1882-1956.</dd>
  <dt>Published</dt>
    <dd><ul><li>
      <span class="notlistview">London : Methuen Children's,</span>
      1990-1991.</li></ul></dd>
```

This slide shows HTML from http://copac.ac.uk/search?rn=1&cid=0710785944
DL = ‘Definition List’

```
<ul>
    <li>
        <strong>Title:</strong>
        <span class="searchword">Winnie</span>-<span
class="searchword">the</span>-<span class="searchword">Pooh</
span> / A. A. Milne ; with decorations by Ernest H. Shepard.
    </li>
   <li>
        <strong>Author:</strong>
        <a href="search.do?...">A. A.  Milne (Alan Alexander),
1882-1956.</a>
    </li>
    <li>
        <strong>Contributor:</strong>
            <a href="search.do?... " >Ernest H  Shepard (Ernest Howard),
1879-1976.</a>
    </li>
...
</ul>
```

This slide shows modified HTML markup taken from http://explore.bl.uk:80/
See how the date is merged with the author?

Structured Data
Documents

```
<titleInfo>
<title>Winnie-the-Pooh story books</title>
</titleInfo>
<note type="statement of responsibility">A.A.
Milne ; illustrated by Ernest H. Shepard.</note>
<name type="personal">
<namePart>Milne, A. A. (Alan Alexander),</
namePart>
<namePart type="date">1882-1956.</namePart>
<role>
<roleTerm type="text">creator</roleTerm>
</role>
</name>
<name type="personal">
<namePart>Shepard, Ernest H. (Ernest Howard),</
namePart>
<namePart type="date">1879-1976.</namePart>
</name>
```

This slide shows XML from http://copac.ac.uk/crn/0710785944 / http://copac.ac.uk/search?rn=1&format=XML+-+MODS&cid=0710785944 (no longer works)

```
[

 {"crn":
 "0710785944",

 "date":
 "19901991",

 "statement_of_responsibility":
 "A.A. Milne ; illustrated by Ernest H. Shepard.",

 "title":
 "Winnie-the-Pooh story books"


 }
]
```


This slide shows JSON from http://copac.ac.uk/search?rn=1&format=json&cid=0710785944 (no longer works)

Just the Data

RDF

All about URIs – hence ‘linked data’
To be revisited later in the day

5 Stars of Linked Open data

* Available on the web (whatever format) but with an open licence, to be Open Data

** Available as machine-readable structured data (e.g. excel instead of image scan of a table)

*** as (2) plus non-proprietary format (e.g. CSV instead of excel)

**** All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff

***** All the above, plus: Link your data to other people's data to provide context

5 Stars of Linked Open data

http://www.w3.org/DesignIssues/LinkedData.html

Structured Data in Documents

```
<h1 id="bookTitle" class="bookTitle" itemprop="name">Winnie-the-Pooh
    <a href="/series/87815-the-winnie-the-pooh-series"
class="greyText">(The Winnie-the-Pooh Series #1)</a>
</h1>

<div id="bookAuthors" class="stacked">
    <span class="by smallText">by</span>
    <span itemprop="author" itemscope="" itemtype="http://schema.org/
Person">
        <a href="http://www.goodreads.com/author/show/
81466.A_A_Milne" class="authorName" itemprop="url"><span
itemprop="name">A.A. Milne</span></a>,
        <a href="http://www.goodreads.com/author/show/
57085.Ernest_H_Shepard" class="authorName" itemprop="url"><span
itemprop="name">Ernest H. Shepard</span></a>
        <span class="authorName greyText smallText role">(Illustrations)</
span>
    </span>
</div>
```

This slide shows HTML with schema.org markup from http://www.goodreads.com/book/show/99107.Winnie_the_Pooh

OCLC includes schema.org markup in Worldcat

During 2013 a group collaborated to make improvements to how bibliographic data is represented in schema.org markup

Summary

- Most content on the web is in HTML; structured as document
- Popular structured data formats also used are XML and JSON
- RDF and Linked Data have gained momentum in the last few years
- Schema.org combines aspects of Linked Data with HTML markup

See also http://ptsefton.com/2013/02/05/putting-data-on-the-web.htm