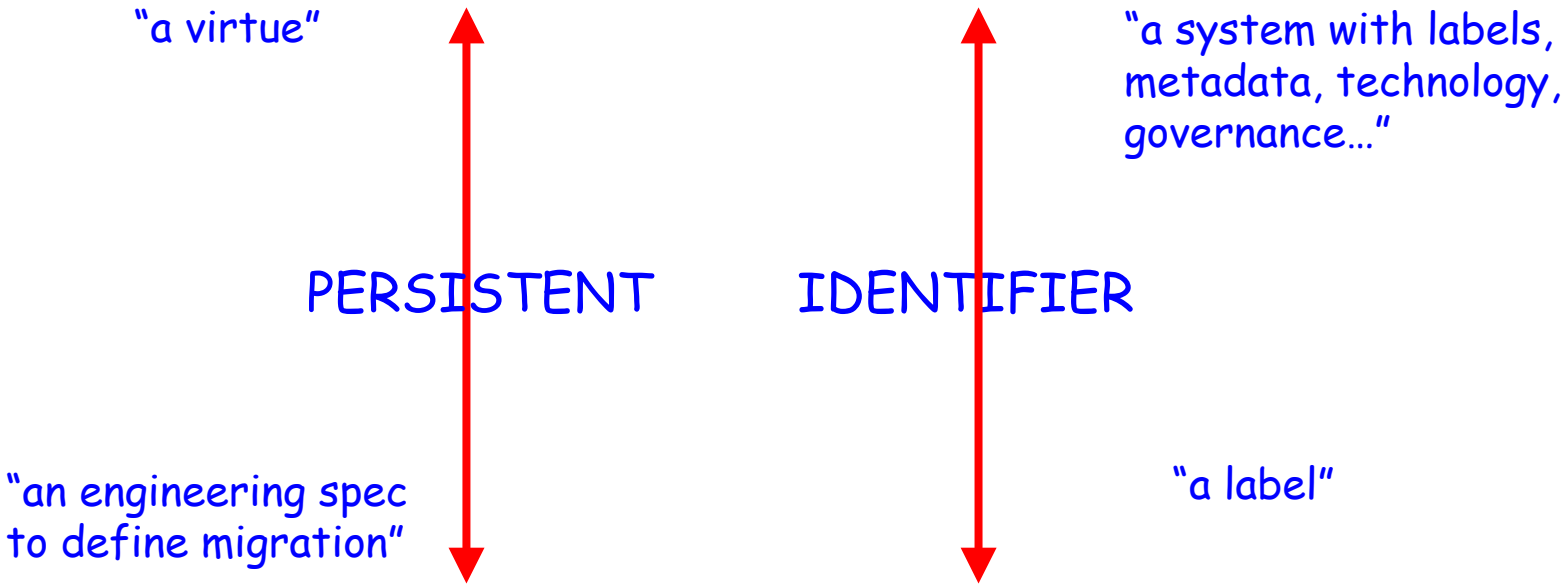
ERPANET Persistent Identifiers seminar

KEYNOTE

The development of persistent identifiers

Norman Paskin, International DOI Foundation

The word trap…



- There are several meanings for "persistent" and "identifier" , so:
  1. Even if using only one word:
     do you and I mean the same thing when we say e.g. "identifier"…?
  2. Some combinations of the two are essentially meaningless
     - category mistake ("the personality of a banana")
- Philosophers solve this problem by "defining what functions you mean by this word?" (functional decomposition); but…

Identifiers

- We all know our own back yard ("We all know what we mean")
- Q: Why do we want persistent identifiers?
- A: For interoperability
- "persistence is interoperability with the future"
- We know what we mean, but others may not.
  - Identifiers assigned in one context may be encountered, and may be re-used, in another place (or time) - without consulting the assigner. You can't assume that your assumptions will be known to someone else. Interoperability = the possibility of use in services outside the direct control of the issuing assigner
- Interoperability is accelerated through automation:
  - Two key events:
  - 1966: automation of supply chains (ISBN)
  - 1994: automation of sharing resources (WWW)
- Increasing interoperability = increasing chance of breakdown

Persistence

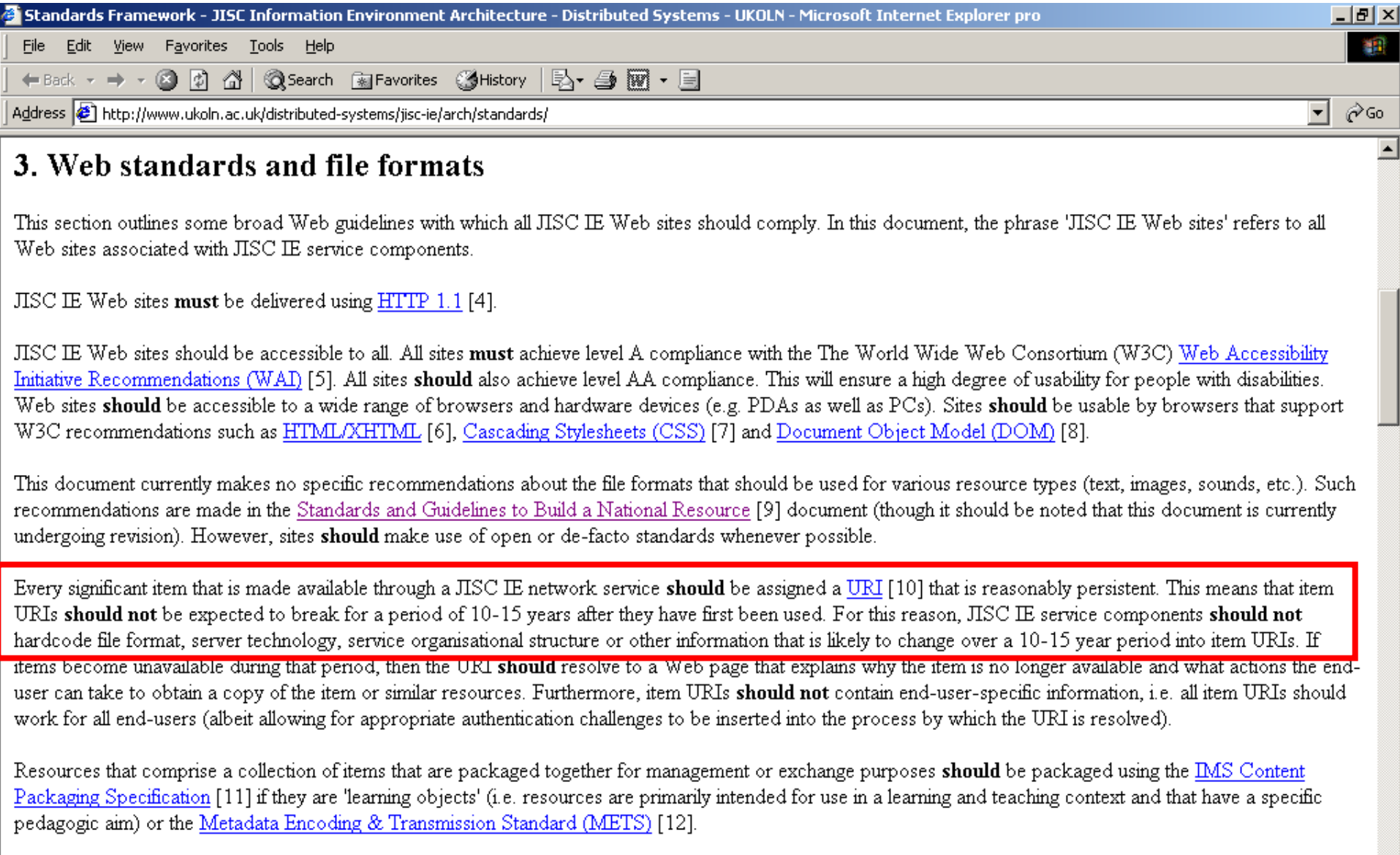
- "It is intended that the lifetime of a [persistent identifier] be permanent. That is, the [persistent identifier] will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name."

- [Persistent Identifier] = URN in IETF RFC 1737 Functional Requirements for Uniform Resource Names. (http://www.ietf.org/rfc/rfc1737.txt)

Technical and social infrastucture issues

Persistence?

JISC Information Environment Architecture Standards Framework Version 1.1 May 2004



3. Web standards and file formats

This section outlines some broad Web guidelines with which all JISC IE Web sites should comply. In this document, the phrase 'JISC IE Web sites' refers to all Web sites associated with JISC IE service components.

JISC IE Web sites **must** be delivered using HTTP 1.1 [4].

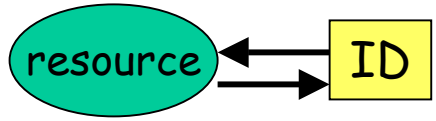
JISC IE Web sites should be accessible to all. All sites **must** achieve level A compliance with the The World Wide Web Consortium (W3C) Web Accessibility Initiative Recommendations (WAI) [5]. All sites **should** also achieve level AA compliance. This will ensure a high degree of usability for people with disabilities. Web sites **should** be accessible to a wide range of browsers and hardware devices (e.g. PDAs as well as PCs). Sites **should** be usable by browsers that support W3C recommendations such as HTML/XHTML [6], Cascading Stylesheets (CSS) [7] and Document Object Model (DOM) [8].

This document currently makes no specific recommendations about the file formats that should be used for various resource types (text, images, sounds, etc.). Such recommendations are made in the Standards and Guidelines to Build a National Resource [9] document (though it should be noted that this document is currently undergoing revision). However, sites **should** make use of open or de-facto standards whenever possible.

Every significant item that is made available through a JISC IE network service **should** be assigned a URI [10] that is reasonably persistent. This means that item URIs **should not** be expected to break for a period of 10-15 years after they have first been used. For this reason, JISC IE service components **should not** hardcode file format, server technology, service organisational structure or other information that is likely to change over a 10-15 year period into item URIs. If items become unavailable during that period, then the URI **should** resolve to a Web page that explains why the item is no longer available and what actions the end-user can take to obtain a copy of the item or similar resources. Furthermore, item URIs **should not** contain end-user-specific information, i.e. all item URIs should work for all end-users (albeit allowing for appropriate authentication challenges to be inserted into the process by which the URI is resolved).

Resources that comprise a collection of items that are packaged together for management or exchange purposes **should** be packaged using the IMS Content Packaging Specification [11] if they are 'learning objects' (i.e. resources are primarily intended for use in a learning and teaching context and that have a specific pedagogic aim) or the Metadata Encoding & Transmission Standard (METS) [12].

Two principles for persistent identification



1. *Obvious:* Assign ID to resource
   - Once assigned the number must identify the same resource
   - Beyond the lifetime of the resource, or the assigner

2. *Less obvious:* Assign Resource to ID
   The resource must be "identified"
   - Must ensure it is always the same thing (bound)
   - Describe the resource "content" [with precision]
   - Failure to do this will ultimately break interoperability

How far do we go in each? Depends on what we think is "good enough"
- Technologists have focussed on (1) [and "bags of bits/data structures"].
- The content/rights world (2) [and focus on "intellectual content"]
- Both viewpoints valid
- (2) is now becoming more relevant

1966: ISBN began "identification numbering"

- "In 1965 the largest British book wholesaler WH Smith announced their intention to move their wholesaling and stock distribution operation to a purpose built warehouse in Swindon [in 1967]. To aid efficiency they would install a computer, and this would necessitate the giving of numbers to all books held in stock…"

- "The idea of numbering books is not new. One British publishing house has been giving numbers to its books for nearly a hundred years. What is an entirely new concept, however, is that numbers should be given to all books; that these numbers should be unique and non-changeable; and that they should be allocated according to a standard system…"

  (David Whitaker, The Bookseller, May 27 1967)

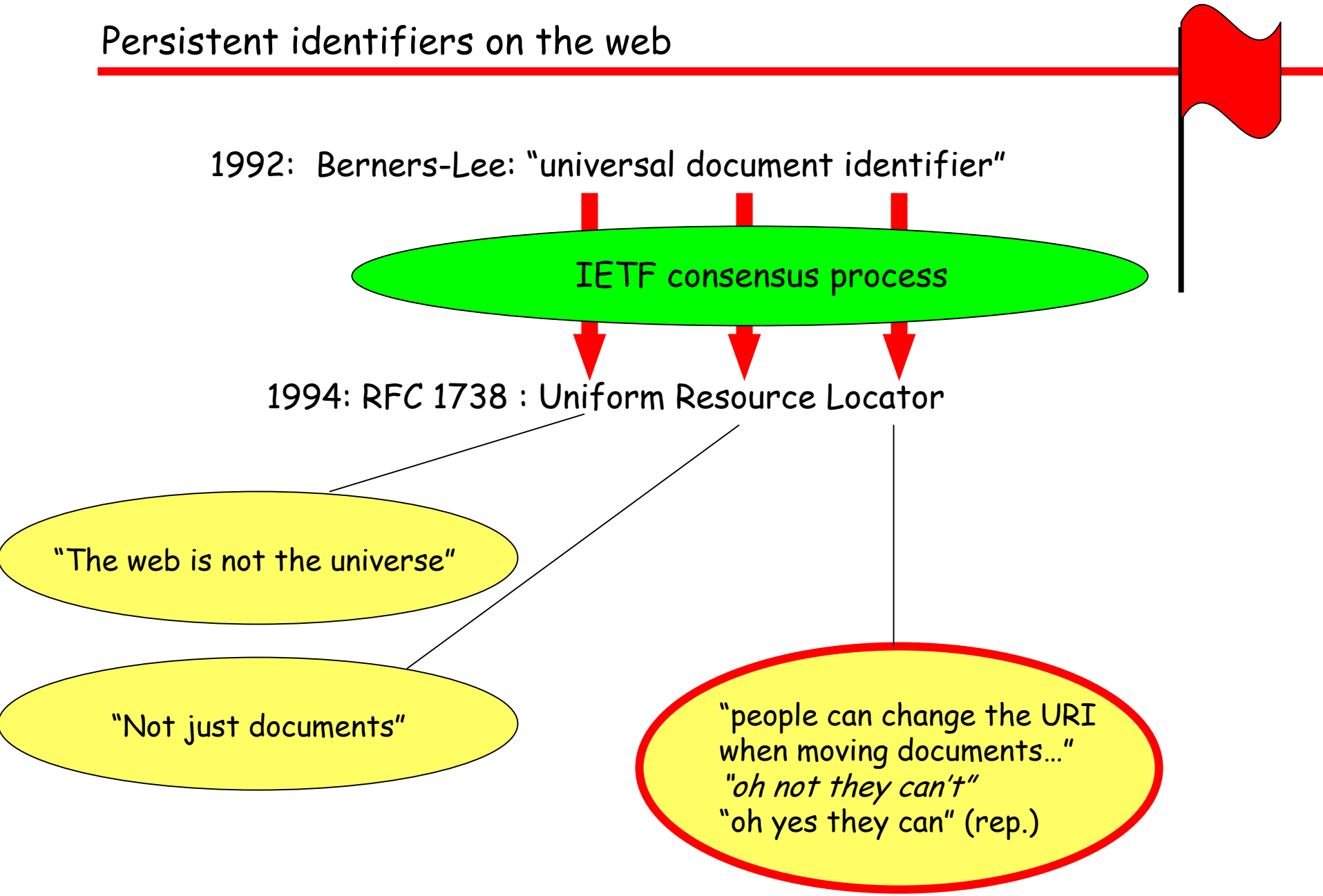
ISO continues "identification numbering"

http://www.collectionscanada.ca/iso/tc46sc9/index.htm
Information and Documentation - Identification and Description

ISO 2108	International Standard Book Numbering (ISBN)
ISO 3297	International Standard Serial Number (ISSN)
ISO 3901	International Standard Recording Code (ISRC)
ISO 10444	International Standard Technical Report Number (ISRN)
ISO 10957	International Standard Music Number (ISMN)
ISO 15706	International Standard Audiovisual Number (ISAN)*
ISO 15707	International Standard Musical Work Code (ISWC)*
ISO Project 20925	Version identifier for Audiovisual Works (V-ISAN)*
ISO Project 21047	International Standard Text Code (ISTC)*

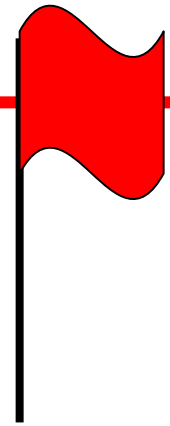
1 : trend towards identifiers of * abstract entities
2. all ISO TC46SC9 identifiers now carry mandatory structured metadata to specify the item identified (either from start, or when revised)

Persistent identifiers on the web

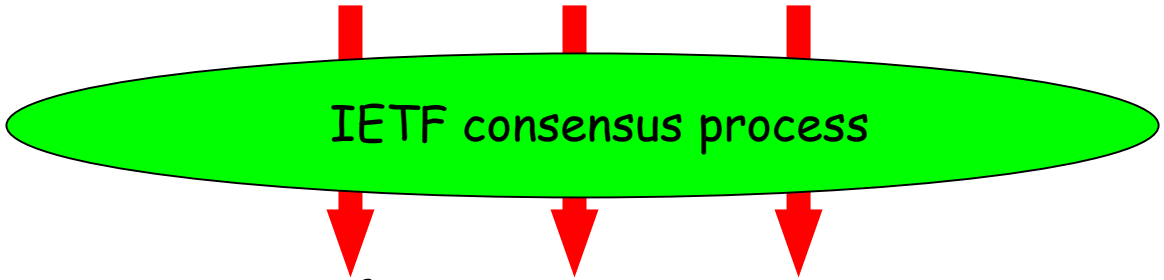


See "Weaving the Web" Ch.5

Persistent identifiers on the web



1992: Berners-Lee: "universal document identifier"



1994: RFC 1738 : Uniform Resource Locator

1995: RFC 1808 : Relative Uniform Resource Locators

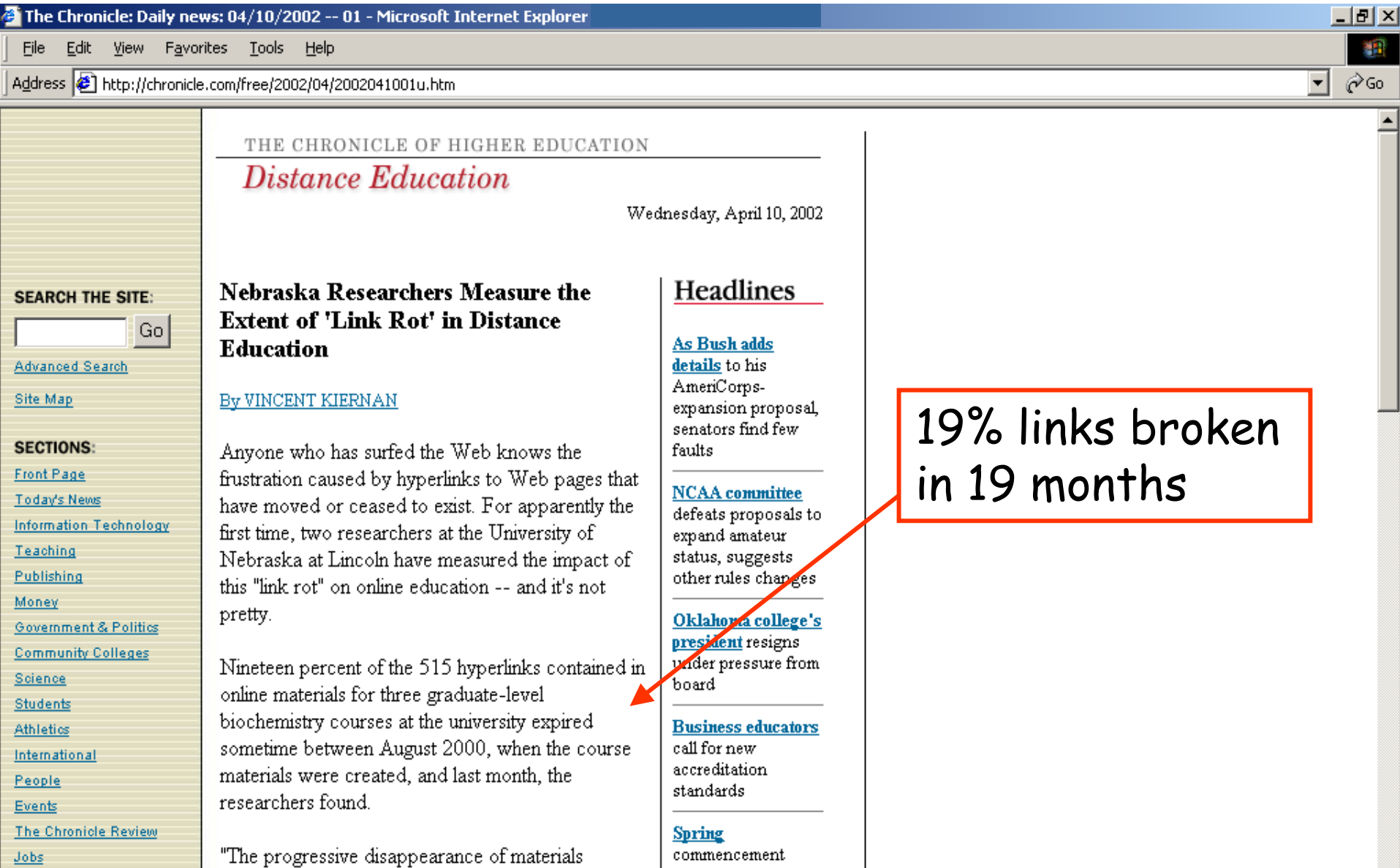
1998: RFC 2396 URI Generic Syntax ("replaces 1738 and 1808")

2004: RFC 2396 bis (revision) ?

2001: Persistence on the web??

- "One of the web sites I maintain is the Lisweb directory of library homepages. Every week, I run a link checker that contacts each page to see if it is still there, and every week about 20 sites that were in place seven days before have vanished. Across the Internet, the rate at which once-valid links start pointing at non-existent addresses -- a process called "link rot" -- is as high as 16 percent in six months. That means that about one sixth of all links will break."
  - *NetConnect,* Thomas Dowling, Library Journal, Fall 2001, p. 36

2002: Persistence on the web??



Back to 1994…

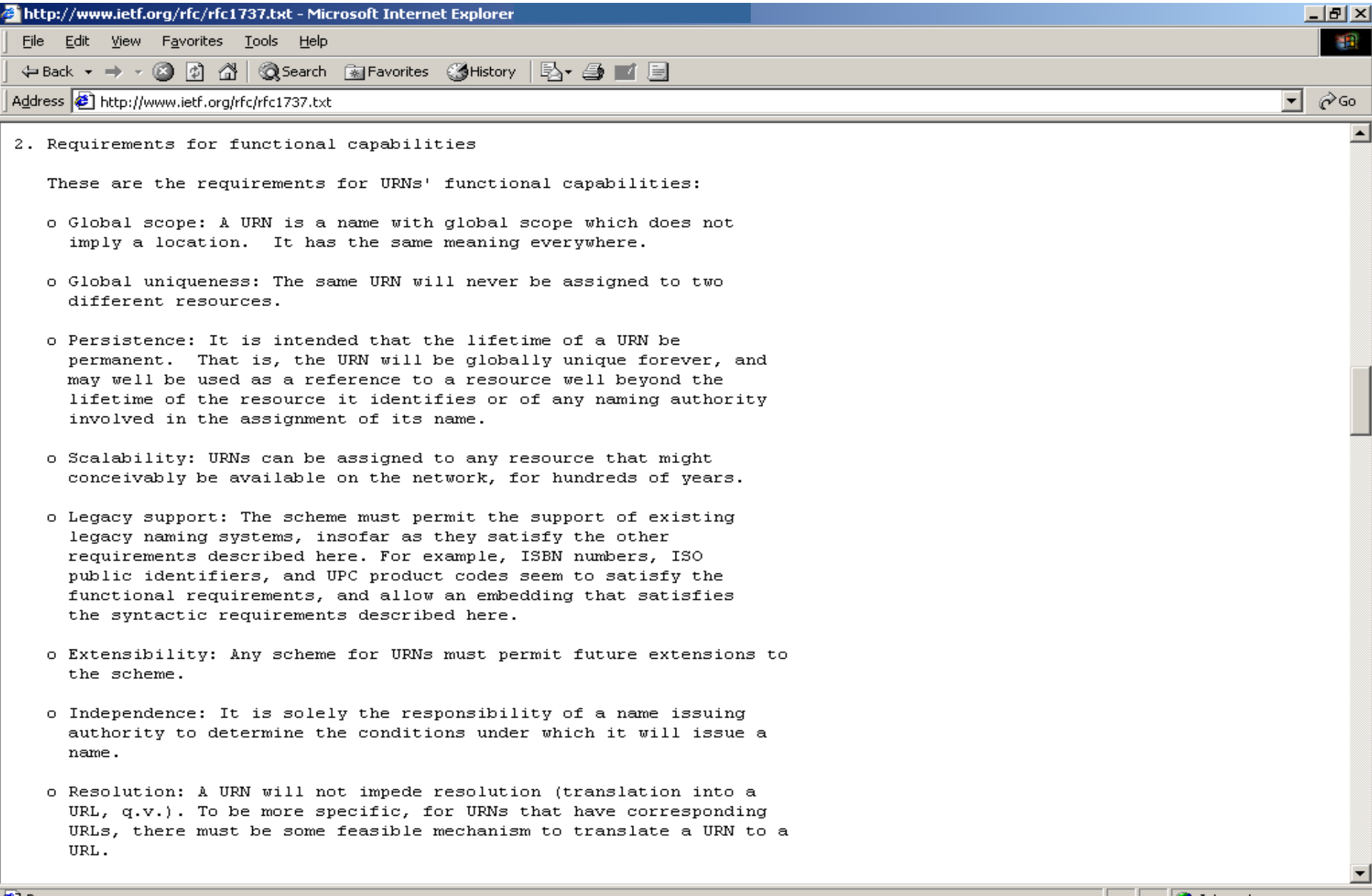
- URL (Uniform Resource Locator) is a location
- Managing by locations alone is not sustainable
- What we need is: a solution for redirecting…
- Or something like a name for the object
  - Treating the object as as "First class object";
  - A Name that could then be relied on even if moved anywhere
- Name would resolve to location (N → L)
  - A Name that is easy to automate, using simple characters

- Hey, didn't those old-fashioned text people do something like that..?

What are we identifying: "actionable identifiers"

- Resolution: The process in which an identifier is the input (a request) to a network service to receive in return a specific output
- "Point and click" is what I do (URL model), so:
- "what I point to (resolve to and get) is what is identified", right?
- It may be – but usually isn't. Consider:
  - Point and click is not referencing
  - Can identify things that are intangible (works), or fugitive (performances)
  - Or that change: "Todays NY Times"
  - People and concepts can be identified but can't be "returned"
  - Pointing and clicking can return different things in different contexts
  - Pointing and clicking can give multiple options
- Identifier identifies an entity. Pointing and clicking is a service about that entity
  - even if a very simple one like "locate an instance"
  - which often really means "locate a derivation"
- Entities can be physical, abstract, tangible, intangible, things, people, concepts, colours… (see later)

1994: Functional Requirements for Uniform Resource Names N → L



2. Requirements for functional capabilities

These are the requirements for URNs' functional capabilities:

- Global scope: A URN is a name with global scope which does not imply a location. It has the same meaning everywhere.

- Global uniqueness: The same URN will never be assigned to two different resources.

- Persistence: It is intended that the lifetime of a URN be permanent. That is, the URN will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name.

- Scalability: URNs can be assigned to any resource that might conceivably be available on the network, for hundreds of years.

- Legacy support: The scheme must permit the support of existing legacy naming systems, insofar as they satisfy the other requirements described here. For example, ISBN numbers, ISO public identifiers, and UPC product codes seem to satisfy the functional requirements, and allow an embedding that satisfies the syntactic requirements described here.

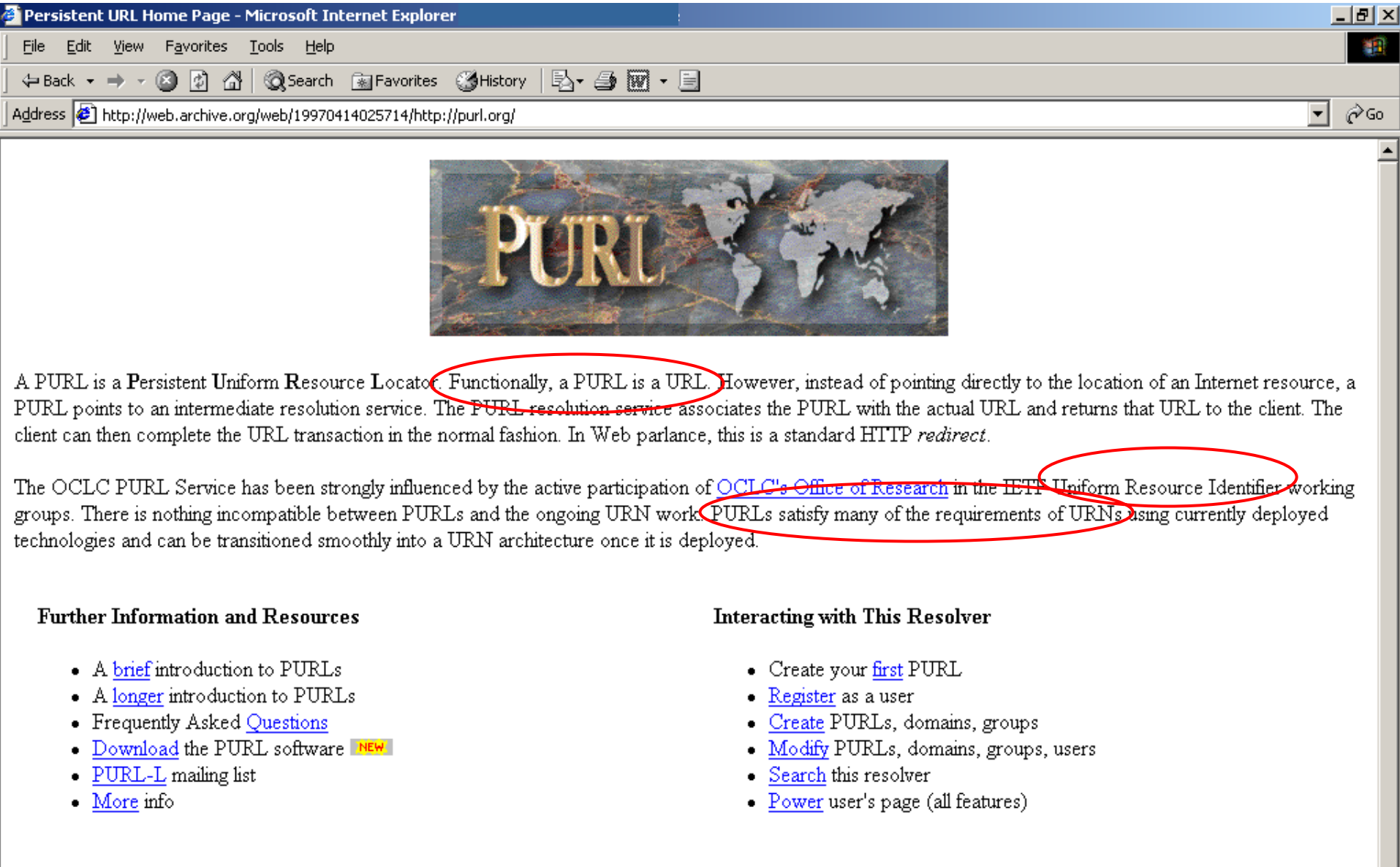
- Extensibility: Any scheme for URNs must permit future extensions to the scheme.

- Independence: It is solely the responsibility of a name issuing authority to determine the conditions under which it will issue a name.

- Resolution: A URN will not impede resolution (translation into a URL, q.v.). To be more specific, for URNs that have corresponding URLs, there must be some feasible mechanism to translate a URN to a URL.

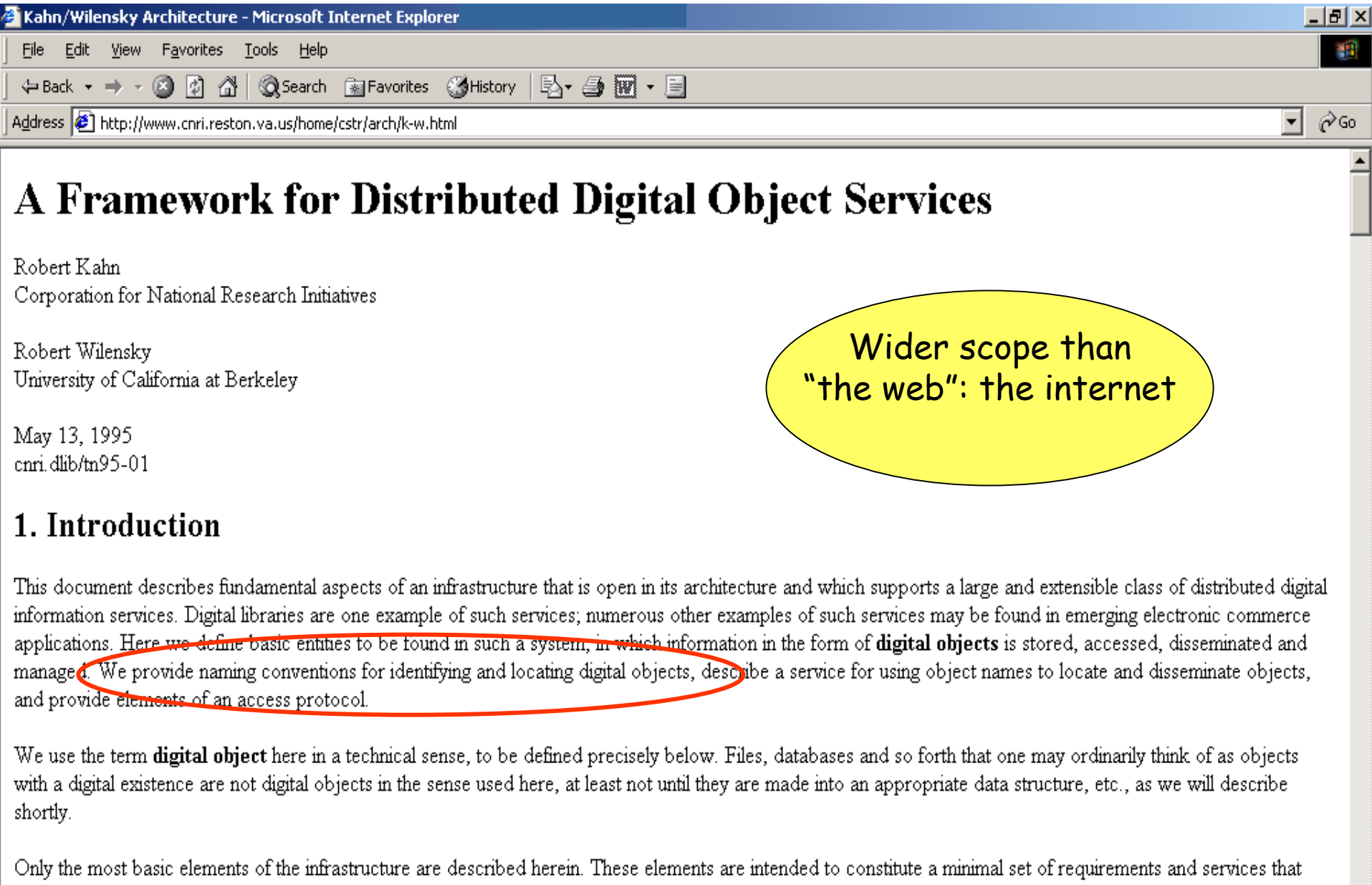
1995: Persistent URLs – redirection service

L → L



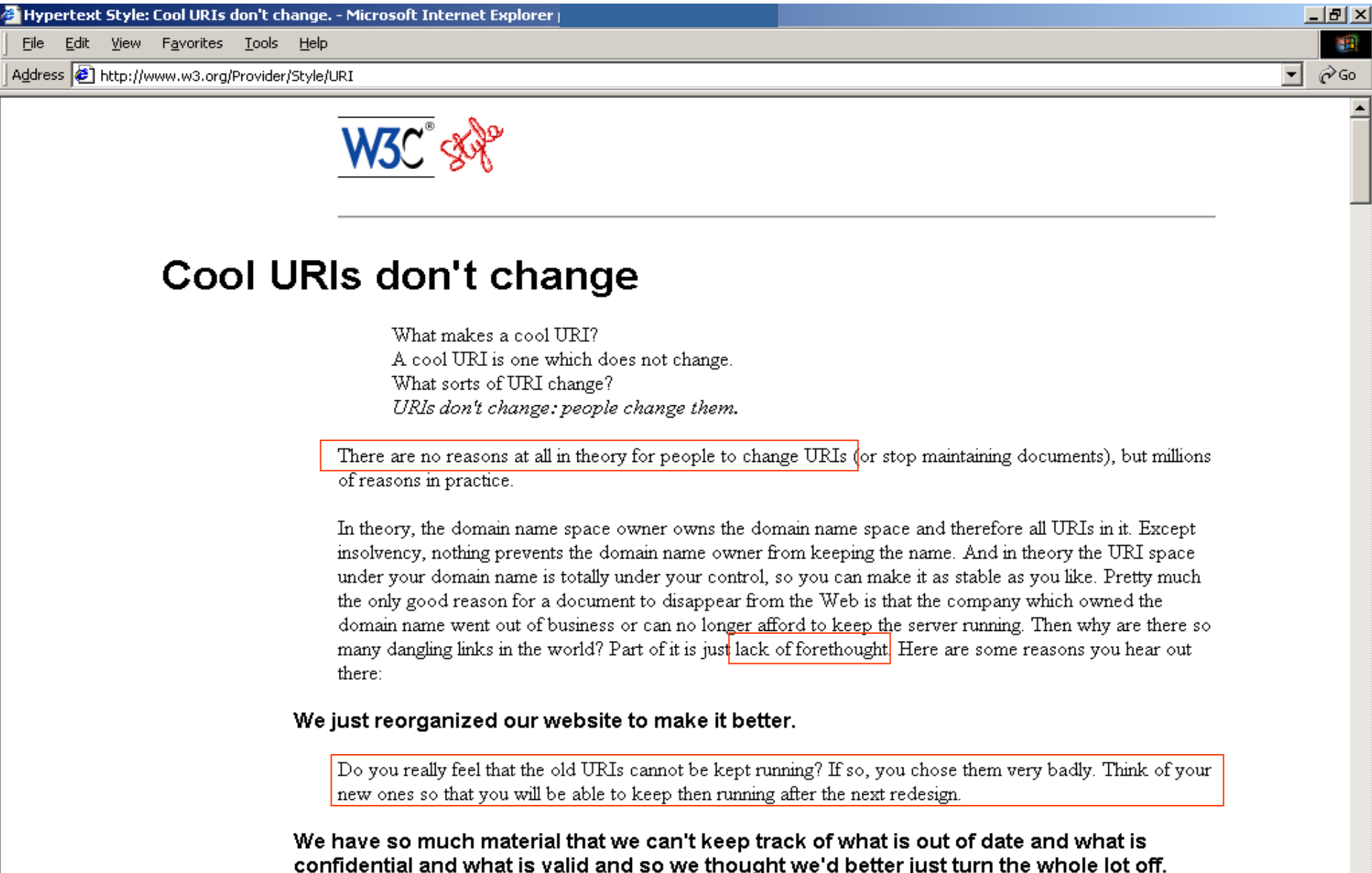
1988-95: DO architecture, Handles

N → L



Wider scope than "the web": the internet

1998: "Cool URI's don't change"



Blackwell Publishing	BMJ Publishing Group
For Immediate Release	**NEWS RELEASE**

Blackwell Publishing Acquires BMJ Books

Oxford & London, UK— April 8, 2004—Blackwell Publishing Ltd and BMJ Publishing Group Ltd announced today that they have finalized the sale of BMJ Books, the book-publishing arm of BMJ, to Blackwell. Each year, Blackwell publishes over 100 books and 350 journals in medicine, biomedicine and allied health. With the acquisition of BMJ Books, titles in specialties such as pediatrics, accident and emergency medicine, anesthesia and intensive care, cardiology, and evidence-based medicine will be added to the company's list.

"Blackwell has a history of providing the medical community with a comprehensive collection of high-quality medical books. The addition of BMJ titles only strengthens the scope of Blackwell's service to clinicians world-wide, and demonstrates our commitment to growth in medical publishing," said René Olivieri, CEO of Blackwell Publishing. "As this transfer moves forward, Blackwell will further the high service standards set by BMJ, reinforcing the strength of publishing excellence for which Blackwell is known."

Richard Smith, editor of the BMJ and Chief Executive of BMJ Publishing Group Ltd said, "This decision has not been easy to make, but it is no longer feasible for us to continue an extensive book publishing business. We have published great books of which we are

Edit View Favorites Tools Help

Back Search Favorites History

Free Hotmail Windows Customize Links

http://www.bmjbooks.com/

arch All go how do I? home

arch search in

BMJ Books

news / events products books

products the group advertising

BMJ Books

Journals Conferences

CME/CPD BMJ Careers

inical Evidence Evidence-Based Products

rofessional and Scientific Publications Library/ Resource Centre

print friendly

BMJ Books

For Over 20 years, BMJ Books has been publishing high quality medical books and other materials for practising clinicians in the UK and around the world. Read more about BMJ Books

eBooks are now available for you to view on your desktop, laptop or PDA - more...

Search books Browse by subject 2004 catalogue (pdf)

Featured Title

Evidence-based Rheumatology

Evidence-based Rheumatology discusses the application of evidence based principles in rheumatological practice. The clinical section provides an overview of the evidence for optimum management in the key areas. Drawing on the expertise of rheumatologists working in evidence based medicine worldwide, the book is an important contribution to the literature, and includes a unique consumer summaries and decision aids section.

one Internet

From: ISO/TC 46/SC 9 Committee [mailto:ISTC-L@INFOSERV.NLC-BNC.CA]
Sent: 15 April 2004 15:36
To: ISTC-L@INFOSERV.NLC-BNC.CA
Subject: Change of e-mail & Web addresses for ISO/TC46/SC9 work

Dear ISTC Colleagues,

Due to the creation of the new Library and Archives Canada, the Internet domain that hosts ISO/TC46/SC9 and its various Working Groups has been changed from "nlc-bnc.ca" to "lac-bac.gc.ca".

Please update your address books for the following:

- Jane Thacker: jane.thacker@lac-bac.gc.ca
- ISO/TC46/SC9 Secretariat: iso.tc46.sc9@lac-bac.gc.ca

And change your bookmarks for the ISO/TC46/SC9 Working Group 3 (ISTC) Web site to:

  http://www.lac-bac.gc.ca/iso/tc46sc9/wg3.htm

The address for the server that runs the ISTC-L discussion list has NOT been changed. Continue to send your messages to:

   istc-l@infoserv.nlc-bnc.ca

Thank you.

1999: Berners-Lee summary

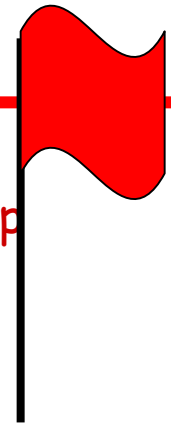
- "URI (Universal Resource Identifier) The string (often starting with http:) that is used to identify anything on the Web.
- URL (Uniform Resource Locator) *A term used sometimes for certain URIs to indicate that they might change."*
- [URNs – Uniform Resource Names: *not mentioned in text, glossary or index*]
- [URCs – Uniform Resource Characteristics: *abandoned concept not mentioned – now we call it "metadata"*]

"Weaving the Web" ,Tim Berners-Lee 1999

But others still find older concepts useful –

- "PURLs satisfy many of the requirements of URNs using currently deployed technologies" (www.purl.org)
- "The DOI can also be considered a URN…"(www.doi.org)
- "A Uniform Resource Name is a URI …that is intended to 'name' a resource in a persistent way" (UKOLN guidelines for encoding identifiers).

Persistent identifiers on the web: observations



- The 'web' has moved beyond html: anything that moves data using http and "identifies" information entities using URLs.
- URLs, as currently understood, are demonstrably not persistent
  - calling them URIs doesn't fix that
- The IETF RFC consensus process, and the separate existence of W3C, leads to ongoing debate and standards with a vague existence
  - compare with ISO standards
  - W3C web site on naming and addressing is "incomplete"
  - Current discussion of "non-IETF" namespaces as separate *(info, DOI)*
- The Web is not the universe
  - It is not all of digital information (<1%):
    - see "How much Information" [e.g. email; Blackberry services medical, legal]
  - It is not all of the internet (see "What is the Internet..")
- URI is a useful catch-all syntactic device for referencing in XML – can describe e.g. ISBNs as URIs
  - It is useful to have such a single framework which can accommodate any other identifier for referencing
  - But it is not, as such, persistent (nor predictable)

Beyond the web

- *The web now:*
  - *170 terabytes (170 x 1012 bytes) as web pages*
  - *= 17 times the size of the Library of Congress print collections*
  - *92,000 terabytes deep web (much available offline) = content in databases which can be searched on a web site, but cannot be found by web crawlers such as Google.*
  - *Three times what was there three years ago*
  - *= 7300 x Library of Congress print collection in total*
- *For comparison: beyond the web we are adding:*
  - *5 exabytes (5 x 1018 bytes) per year*
  - *= 37,000 new libraries the size of the Library of Congress print collections.*
- *Web pages are a small % of total digital information*
- *More information services are appearing on non web (eg Blackberry)*
- *Interoperability is already beyond the control of the originator*

Source: "How much information 2003":
http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

"Internet" refers to the global information system that --

(i) is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons;

(ii) is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and

(iii) provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein."

Source: U.S. Federal Networking Council 1995: ("What Is The Internet (And What Makes It Work)" - Robert E. Kahn and Vinton G. Cerf : http://www.cnri.reston.va.us/what_is_internet.html

1995: Armati Report

- *Information Identification - a report to STM publishers* (Mar 95)
- *Uniform File Identifiers - a report to AAP publishers* (Oct 95)
- "..need to unify in one scheme music, audiovisual, document management, internet engineering, digital libraries, copyright registration and object based software" *[i.e. web was not the focus]*
- "..maximise utility of digital objects; enable core interoperability; enable integration of disparate sourced data; ability to trace ownership to manage rights"
- requirements:
  - protect legacy investments
  - enable interoperability
  - provide link between digital and physical
  - maintain privacy of users
  - have persistence
  - standard syntax
  - global scalability
  - global uniqueness
  - global meaning
- Led to launch of DOI initiative *(AAP committee, Uniform File Identifier)*

1998: DOI - Digital Object Identifier system

doi> The Digital Object Identifier System®

Developed by The International DOI Foundation (IDF)

Search Guidelines | Recent Changes | Contact Us | Members Only

Site Search [Tips]

Search

Learn About DOIs
- Overviews
- Frequently Asked Questions
- Factsheets
- DOI Handbook

International DOI Foundation
- Director's Message
- Membership
- Information Kit
- IDF Staff

Activities
- News/Events
- Mailing Lists/Working Groups
- Reviews

Resources
- Registration Agencies
- White Papers
- DOI Demonstrations
- DOI Tools

IDF Members Only
- Member's Site
- Director's Report

Subscribe to DOI News!

Home

Welcome to the Digital Object Identifier System

The Digital Object Identifier (DOI) is a system for identifying and exchanging intellectual property in the digital environment.

The DOI System provides a framework for managing intellectual content, for linking customers with content suppliers, for facilitating electronic commerce, and enabling automated copyright management for all types of media. The system is managed and directed by the International DOI Foundation. Several million DOIs have been assigned by DOI Registration Agencies in the US, Australasia, and Europe.

DOIs are names (characters and/or digits) assigned to objects of intellectual property (physical, digital or abstract) such as electronic journal articles, images, learning objects, ebooks, images, any kind of content. They are used to provide current information, including where they (or information about them) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI will not change.

Using DOIs as identifiers makes managing intellectual property in a networked environment much easier and more convenient, and allows the construction of automated services and transactions for e-commerce.

To learn more about DOIs, see the Overviews, and begin with the Introductory Overview and Introductory Slide Presentation. The Frequently Asked Questions (FAQs) cover a wide range of topics, both general and technical. For the most complete description of all aspects of the DOI System technology and policy, consult the DOI Handbook.

In the News

R. R. Bowker appointed as DOI Registration Agency

Nielsen BookData appointed as DOI Registration Agency

IDF recommended as MPEG Rights Data dictionary authority"

Analysis of Economic Benefits of DOI Published

EPS Focus Report Published: "DOI in 2004"

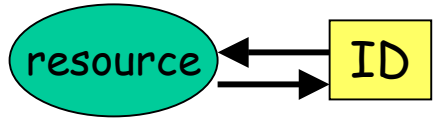
Resolve a DOI!

Type or paste a DOI (i.e., 10.1000/182) into the text box below.

And other things….

- National Bibliography Numbers…
- Government codes…
- Supply chain: Bar Codes, RFIDs…
- Hyper-G…

Two principles for persistent identification



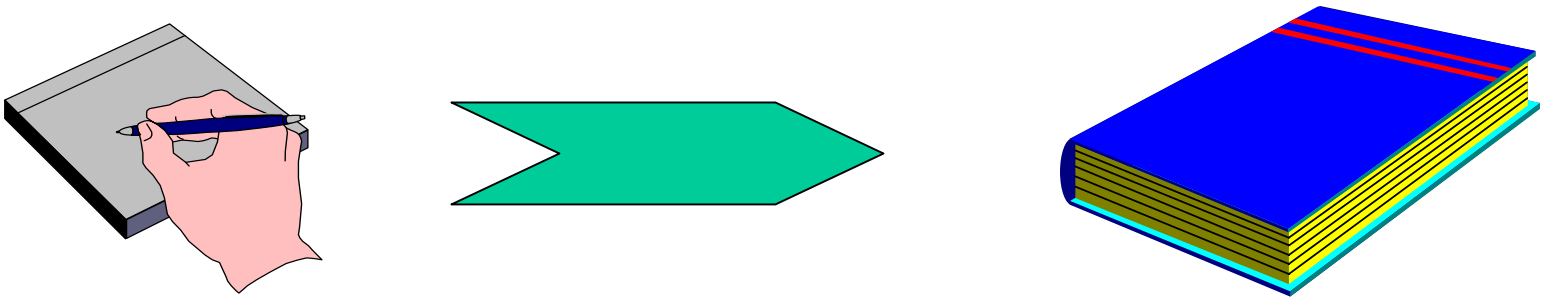
1. *Obvious:* Assign ID to resource
   - Once assigned the number must identify the same resource
   - Beyond the lifetime of the resource, or the assigner

2. *Less obvious:* Assign Resource to ID
   The resource must be "identified"
   - Must ensure it is always the same thing (bound)
   - Describe the resource "content" [with precision]
   - Failure to do this will ultimately break interoperability

How far do we go in each? Depends on what we think is "good enough"
- Technologists have focussed on (1) [and "bags of bits/data structures"].
- The content/rights world (2) [and focus on "intellectual content"]
- (2) is now becoming more relevant

Specifying what is identified

1. Electronic Resource Preservation
   - Usually of "tangible resources" (copies of documents, images, etc)
   - Familiar idea of "preservation metadata"
2. Other persistent applications may use identifiers of intangibles:



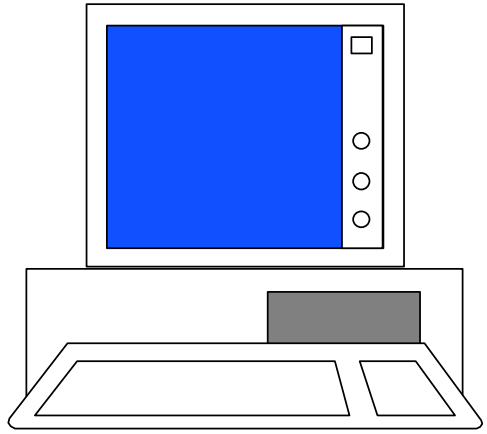
Manuscript
mss #ABC123

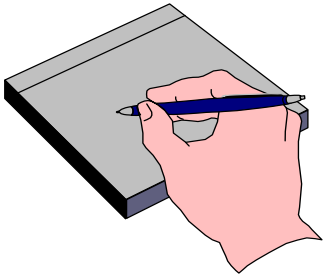
paper
journal/volume/page

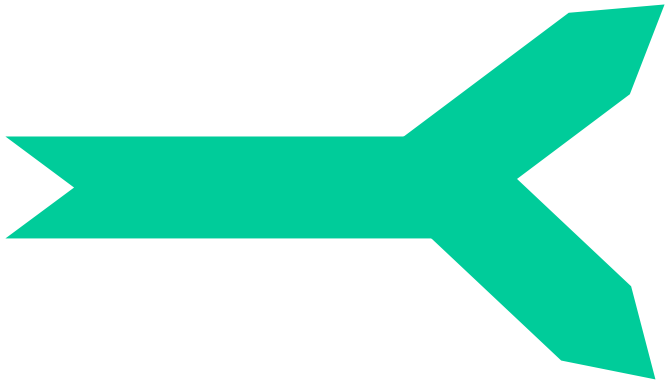
Two things in one:
Physical manifestation of intangible work
(which is identified?)

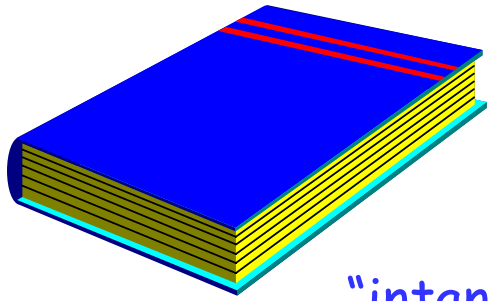
Web page URL

"intangible Work"









Vol/page; ISBN; SICI, etc

"intangible Work"

"work" used in analytical sense, not copyright sense

All Editions

Search: Books for GO!

So You'd Like to...

Offer your advice

to be a geek: by s_etoiles, linux junkie, hacker, coder, geek

You may also like

ow the Web was Born
Robert Cailliau, James Gillies (Rate it)

Page You Made

he Page You Made:
here Wizards Stay Up Late
Katie Hafner (Author)
(Rate it)

All 3 editions :

Sort by: Featured Items GO!

1. 

Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web
by Tim Berners-Lee (Author) (**Paperback**)
Avg. Customer Rating: ★★★★☆
(Rate this item)

Usually ships in 24 hours
List Price: ~~$15.00~~
Buy new: **$10.50**
Used & new from $2.97

2. 

Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by its Inventor
by Tim Berners-Lee, et al
Avg. Customer Rating: ★★★★☆
(Rate this item)

Out of Print--Limited Availability
Used & new from $1.87

3. 

Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor [ABRIDGED]
by Tim Berners-Lee (Author), Berners-Lee Tim (Reader) (**Audio Cassette**)
Avg. Customer Rating: ★★★★☆
(Rate this item)

Usually ships in 24 hours
List Price: $18.00
Buy new: **$18.00**
Used & new from $4.00

Listmania!

Add your list

History of the Net: Frontier, Revolution, Power, Symbiosis: A list by Adam Brate, Author of Technomanifestos (20 item list)

Software Industry Books: A list by Doug Phillips, Software Lawyer (8 item list)

The MBA@home syllabus: A list by drclueful, www.drclueful.com (14 item list)

content in the newspaper, but it is very hard to determine that number. As indicated above, the duplication issue is particularly serious for digital storage, since little of what is stored on individual hard drives is unique. We've tried to adjust for this the best we can, and documented our assumptions in the detailed treatment of each medium.

Compression

The advantage of using a single measurement standard such as terabytes to compare the volume of information in different formats is obvious. However, unlike paper or film, there is no unambiguous way to measure the size of digital information. A 600 dot per inch scanned digital image of text can be compressed to about one hundredth of its original size. A DVD version of a movie can be 1000 times smaller than the original digital image. We've made what we thought were sensible choices with respect to compression, steering a middle course between the high estimate (based on ``reasonable'' compression) and the low estimate (based on highly compressed content). It is worth noting that the fact that digital storage can be compressed to different degrees depending on needs is a significant advantage for digital over analog storage.

About this Report

We view this report as a "living document" and intend to revise it based on comments, corrections, and suggestions. Please send comments to how-much-info@sims.berkeley.edu.

Many thanks to our sponsors. Financial support for this study was provided by:

- Microsoft Research at http://www.research.microsoft.com
- Intel at http://www.intel.com/go/storage
- Hewlett Packard http://www.hp.com
- and EMC at http://www.emc.com.

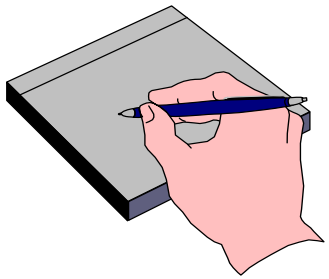
Versions – separately identified?

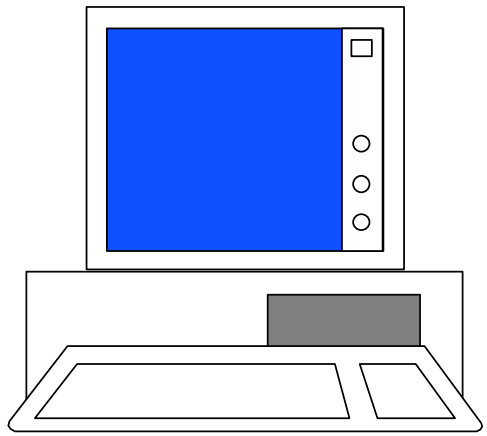
About the School of Information Management and Systems

UC Berkeley's School of Information Management and Systems is the first school in the nation to explicitly address the growing need to manage information more effectively.

With respect to education, we are training a new type of professional: "information managers". Our graduates are familiar with the latest and most powerful techniques for locating, organizing, retrieving, manipulating, protecting, and presenting information. They study not only technology, but also the institutional, legal, economic and organizational factors necessary for creating information systems that meet peoples' needs.

What are we identifying by this identifier?





Document on screen

Abstract work?
Manifestation of abstract work?
Version?
This HTML file?
All/some of these?

Does it matter in everyday life?

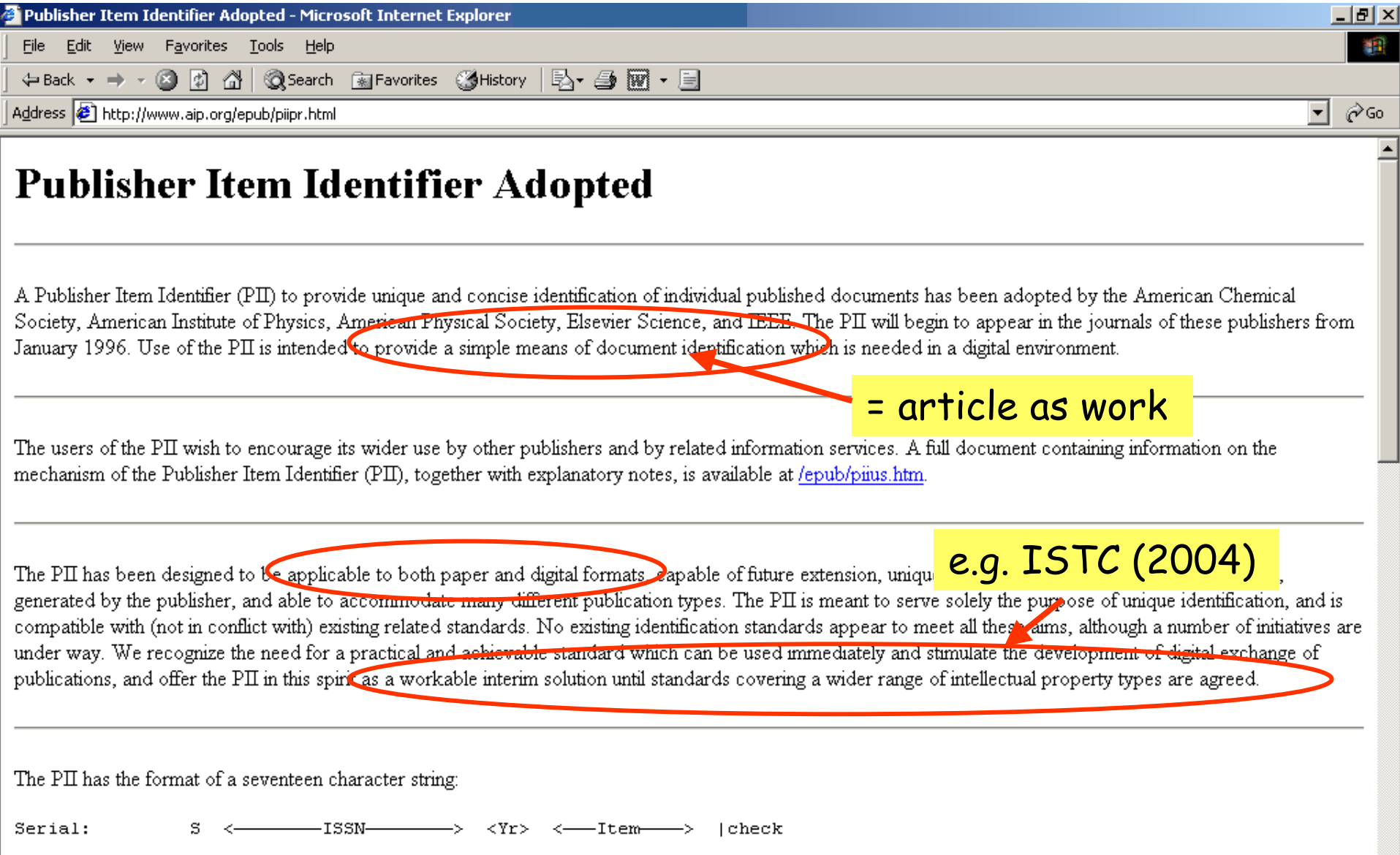
doi>

Yes, it can do. e.g.:

1. Practical use of data. Example – journal article
   - For the purpose of citation:
     - Count pdf, print, html as same
     - Citation refers to the abstract work (hence ISI, CrossRef)
   - For the purpose of purchase:
     - Count pdf, print, html as different
     - Purchase refers to the manifestation
   - Suppose I encounter a purchase system and try to use it for counting citations….
   - Can I rely on a system now if I don't know what is being identified? Can others rely on the system long after I'm gone?
2. Legal implications: copyright

   "My A is the same as your B and is my copyright…"

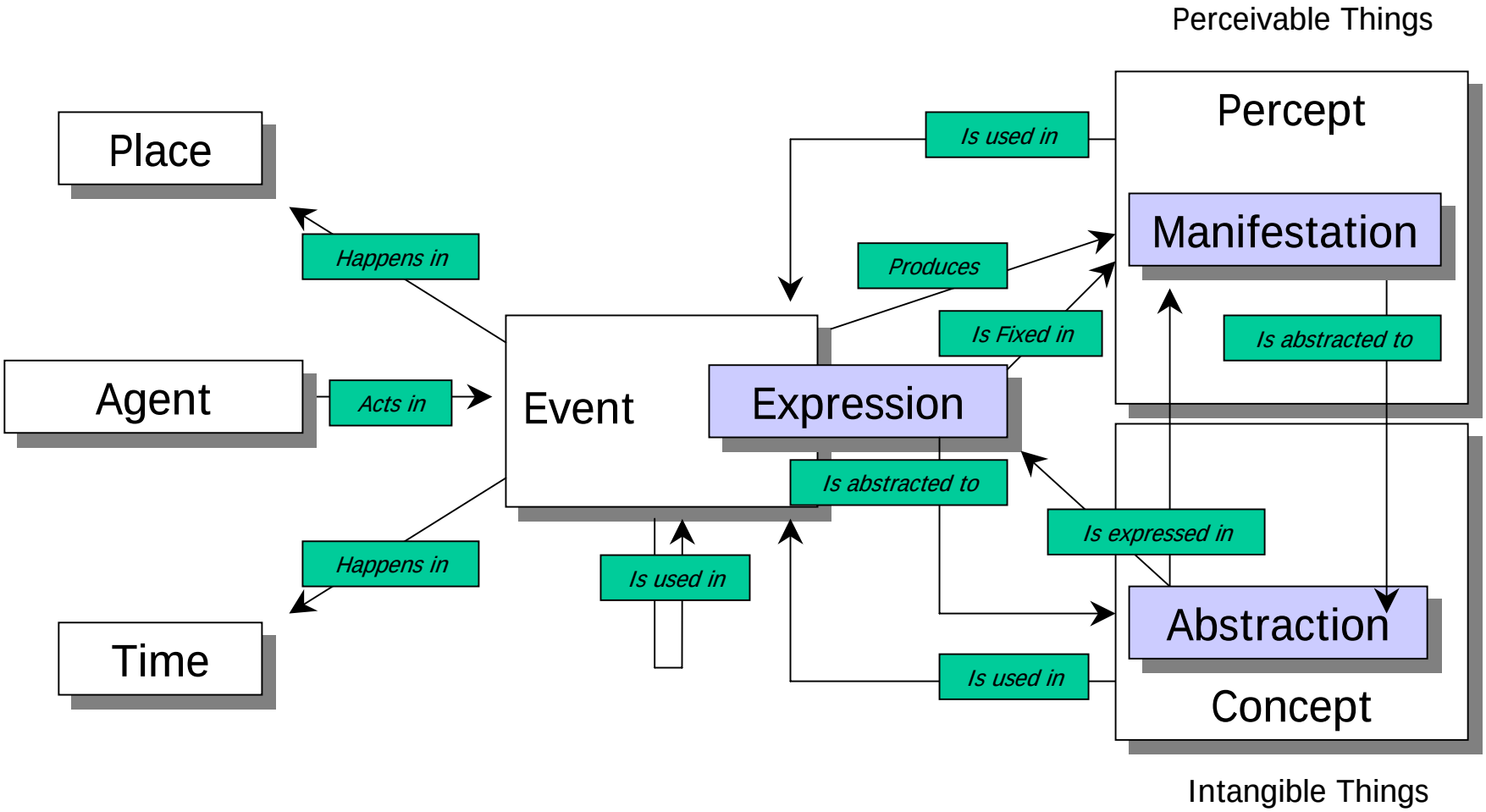
1995: PII as identifier for underlying works



1995-2004: Defining what is identified through metadata

- Many individual metadata schemes for specific sectors, applications, etc.; vary from simple to complex data models
- 1995+: Dublin Core: need for standardisation on WWW
  - 15 (+) elements for "output" for simple resource description
  - Now ISO 15836
- Ontology-based activities:
  - 1995+ : Common Information System "CIS" (CISAC) – *rights, music*
  - 1998: Functional Requirements of Bibliographic Records, "FRBR" (IFLA) – *library cataloguing*
  - 1998-2000: Interoperability of Data in E-Commerce Systems, "indecs" (multiple partners) – *generic intellectual property*
    - For "e-commerce" read "automation"
    - Influenced by CIS and FRBR
  - 2000: ABC/Harmony – *generic events-aware model*
  - *Should enable re-use of existing metadata*

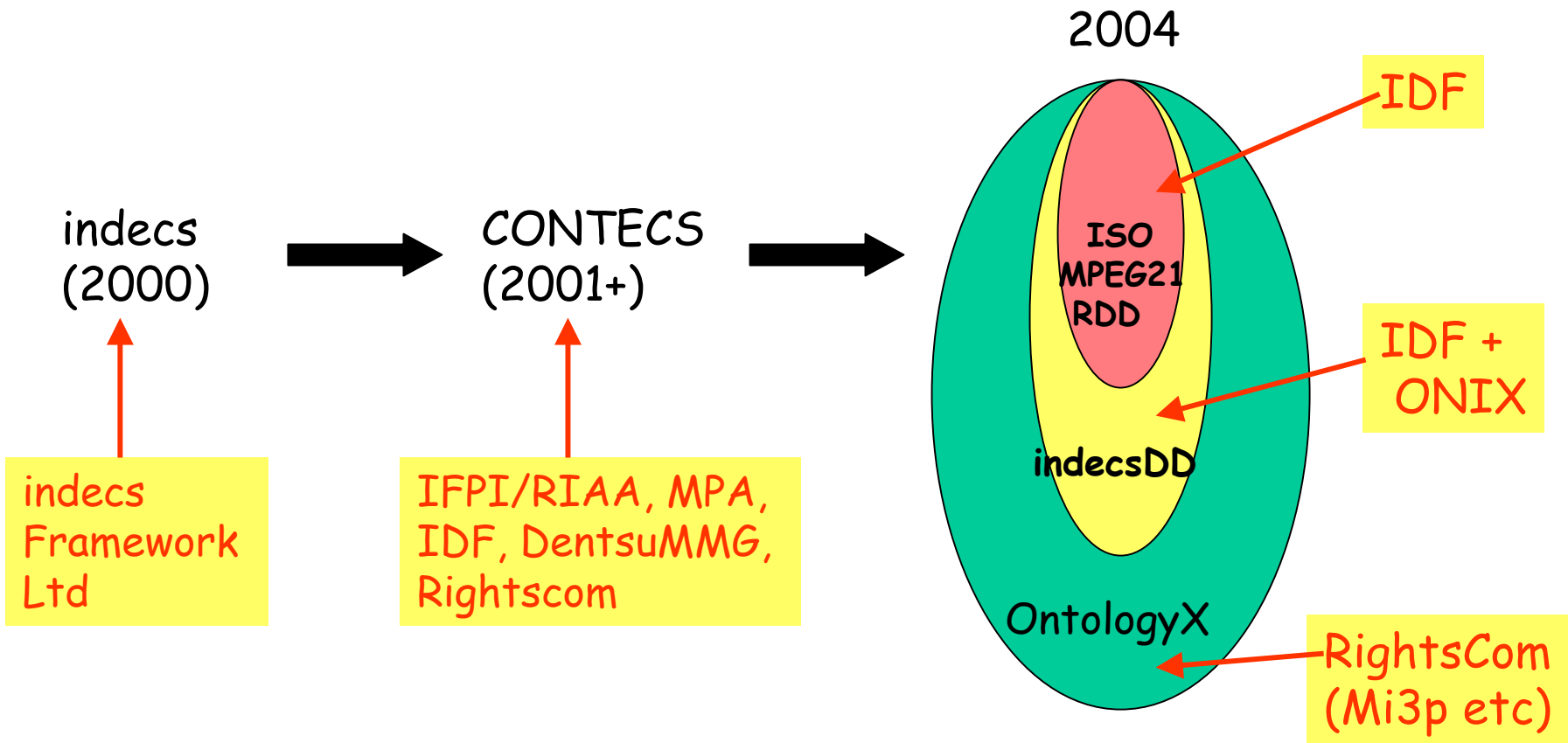
2000: Relating entities through "a model of making"



Source: in*d*ecs Sydney Conference March 2000
[FRBR, ABC have similar schemes]

1995-2004: Defining what is identified through metadata

Development of indecs 2000-2004
Black = what
Red = who



2001: Ontologies and Semantic Web

Scientific American: The Semantic Web - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Links Free Hotmail Windows Customize Links

Address http://www.scientificamerican.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 Go

SCIENTIFIC AMERICAN.COM

Find out with Ask the Experts

May 17, 2001

The Semantic Web

A new form of Web content that is meaningful to computers will unle

By Tim Berners-Lee, James Hendler and Ora Lassila

The entertainment system was belting out the Be
down by sending a message to all the other
to see a specialist and then has to have a
appointments." Pete immediately agre

BY MIGUEL SALMERON

Almost instantly the new plan was presente
reschedule a couple of his *less important* appo
company's list failing to include this provider under
reassured him. "(Details?)"

Lucy registered her assent at about the same moment Pete was muttering,
couldn't resist the
details and later that night had his agent explain how it had found that provider even though it

Internet

"Ontologies

Of course, this is not the end of the story, because two databases may use different identifiers for what is in fact the same concept, such as *zip code*. A program that wants to compare or combine information across the two databases has to know that these two terms are being used to mean the same thing. Ideally, the program must have a way to discover such common meanings for whatever databases it encounters.

A solution to this problem is provided by the third basic component of the Semantic Web, collections of information called ontologies."

Semantic = "meaning". Does A "mean the same as" B ?

= in practice, does A need a different identifier from B?

- versions; works and manifestations; editions
  - [e.g. two different e-book formats of the same work]

- For a machine, "A means same as B" = "A has same attributes as B"
- Which attributes? The answer is entirely contextual :

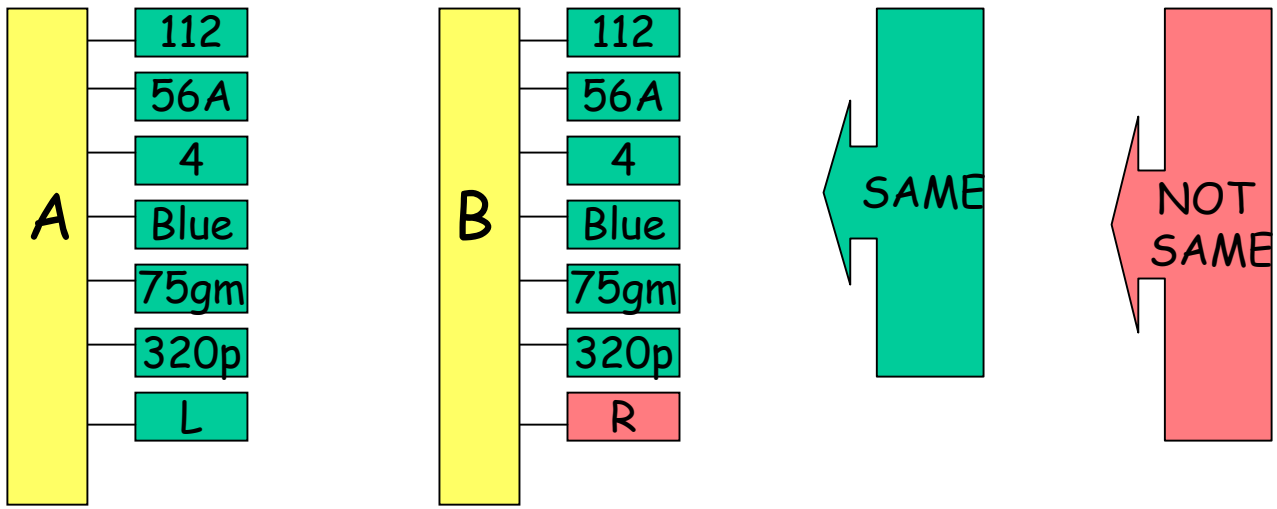
  "Is A the same as B *for the purposes of …?*"

  = Do A and B belong to the same class for the purposes of …
- For a machine, "for the purpose of" = "class having this set of attributes"
- We group similar things together; what is identified is usually a class
  - e.g. *the class of all copies of the hardback printed second edition of this book from this publisher* = the same ISBN
  - The class is defined by a set of attributes (metadata) (RDF, etc)
- No one thing is the same as another thing (or they wouldn't be two things)
  - "Roughly speaking, to say of two things that they are identical is nonsense, and to say of one thing that it is identical with itself is to say nothing at all." (L.W.)
  - Liebniz's Law (no two objects have exactly the same properties)
- Philosophy? philosophy = logic = automation

Distinguishing different entities

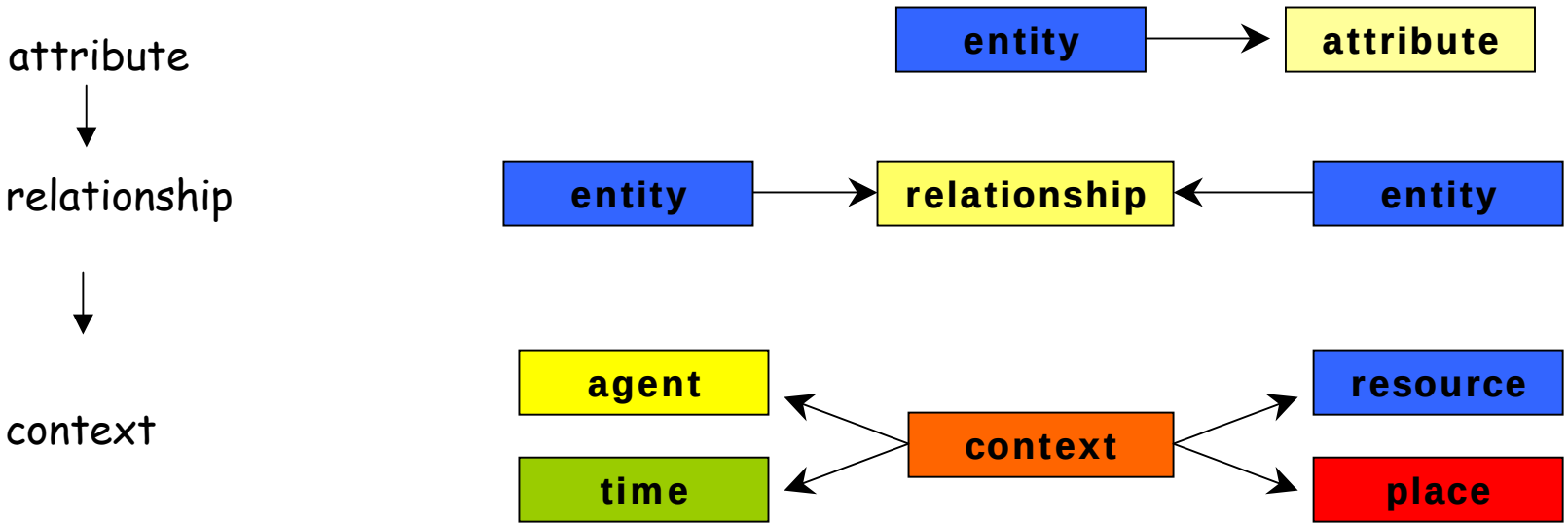
- We can always *add another attribute* to make two "like" things "unlike":
  - the class of all copies of the hardback printed second edition of this book from this publisher = the same ISBN;
  - the class of all copies of the hardback printed second edition of this book from this publisher with the luxury leather binding = different ISBN
- Consequence: *No set of metadata elements is definitive <u>for all purposes</u>*
- Practical consideration of *purpose* = some defined set of attributes

e.g. Bridgeman v Corel (2004): Bridgeman images not copyrightable UK law as they "were substantially exact reproductions of public domain works albeit in a different medium", nor US law which requires "a distinguishable variation" between two distinct copyright items.

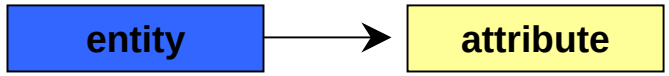


Ontology approach: deeper view of metadata

- The key to defining what is identified logically
  - enabling people to use their existing metadata
  - Ontologies can deliver data dictionaries suitable for mapping
- Fundamental, generic, extensible methods can be used to construct interoperable ontologies – by putting metadata into context:



Ontology approach: attribute



3 levels of attribution

attribute
relationship
context

1. attribute view – simplest, most direct

book
isbn “0297816470”
title “Words & Rules”
author “Stephen Pinker”
publisher “Wiedenfield”
dateOfPublication “1999”
placeOfPublication “UK”

(values may be strings, IDs etc)

Ontology approach: relationship

3 levels of attribution

attribute
relationship
context

2. association view – richer, more indirect

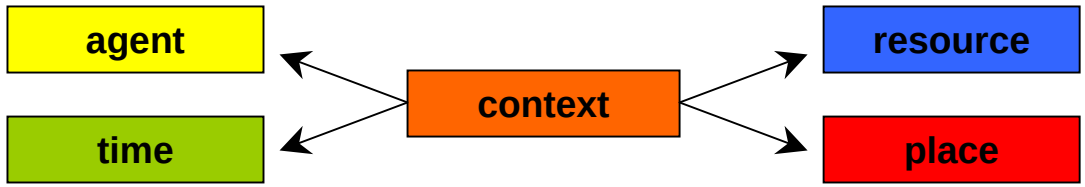
book “0297816470” hasTitle “Words & Rules”
book “0297816470” hasAuthor “Stephen Pinker”
book “0297816470” hasPublisher “Wiedenfeld”
book “0297816470” hasDateOfPublication “1999”
book “0297816470” hasPlaceOfPublication “UK”

allows multiple occurrences
allows ranges of target values
treats attributes as entities

Ontology approach: context

3 levels of attribution

attribute
relationship
context



3. context view – richest, most indirect

publishingEvent hasAgentType publisher “Weidenfeld”
publishingEvent hasResourceType book “0297816470”
publishingEvent hasTimeType dateOfPublication “2002”
publishingEvent hasPlaceType placeOfPublication “UK”

most efficient handling of complex metadata

Issues and themes for persistent identifier applications

ISSUES

- What are we identifying with this identifier? [content not just bits]
- What are we resolving to from this identifier?
- What, if any, explicit metadata are we making available?
- How will the cost of providing the infrastructure be met ?

THEMES

- Identification of entities of all forms
  - *To be used in variety of contexts*
- Appropriate use of metadata at appropriate level
  - *Development of ontology tools to describe entity relationships*
- Persistent → Interoperable → Precise → Automation → Logic
- Are we in the "word trap"?

Further reading

- "On Making and Identifying a 'Copy'". Norman Paskin D-Lib Magazine, Volume 9, Number 1, January 2003. [www.dlib.org] [http://dx.doi.org/10.1045/january2003-paskin]
- "Identification and Metadata: Components of DRM Systems" Norman Paskin; in E. Becker et al (eds) "Digital Rights Management" in the series *Lecture Notes in Computer Science* (Springer-Verlag, 2003) pp. 26-61 [http://www.doi.org/topics/drm_paskin_20030113_b1.pdf]
- DOI factsheets etc. http://www.doi.org/factsheets.html
- Other sources: http://www.doi.org/handbook_2000/bibliography.html

ERPANET Persistent Identifiers seminar

KEYNOTE

The development of persistent identifiers

Norman Paskin, International DOI Foundation
n.paskin@doi.org