Identifying and describing Web resources

European Commission DGXIII/E-4
http://www.elpub.org
Document record

Author: Laurie Causton
Oversight: Derek Kueter
Client: Bernard Smith, European Commission DGXIII/E-4

This report has been produced by TechServ

Acknowledgements

The contributions of the following in providing feedback and comment on the initial draft are gratefully acknowledged:

Dan Connolly Head of the Architecture Domain, World Wide Web Consortium
Norman Paskin International DOI Foundation
Bernard Smith European Commission DGXIII/E-4
Derek Kueter TechServ

<table>
<thead>
<tr>
<th>VERSION</th>
<th>DATE</th>
<th>STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>September 22 1998</td>
<td>Initial draft</td>
</tr>
<tr>
<td>0.2</td>
<td>November 5 1998</td>
<td>Second draft following review</td>
</tr>
<tr>
<td>1.0</td>
<td>November 17 1998</td>
<td>Version 1, after final revision</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

1. INTRODUCTION .......................................................... 4

2. SOME BASIC PRINCIPLES ............................................... 5
   2.1 Resources, objects, creations | 5
   2.2 Identifying, locating and describing | 6

3. IDENTIFYING RESOURCES ............................................. 8
   3.1 Some concepts .................................................. 8
       3.1.1 Uniqueness .................................................. 8
       3.1.2 Persistence ................................................ 9
   3.2 What the technologists have been doing ....................... 10
       3.2.1 Uniform Resource Locator (URL) ......................... 10
       3.2.2 Uniform Resource Identifier (URI) ....................... 11
       3.2.3 Uniform Resource Name (URN) ............................ 11
       3.2.4 Persistent Uniform Resource Locator (PURL) .......... 14
   3.3 What the user community has been doing ....................... 14
       3.3.1 Digital Object Identifier (DOI) ......................... 15
   3.4 Other initiatives ................................................ 17
       3.4.1 WEBDAV and the UUID ..................................... 17
       3.4.2 Persistent Document Identifiers (PDIs) ................. 17
       3.4.3 Human Friendly Names .................................... 18
   3.5 Conclusions ........................................................ 19
       3.5.1 Consensus ................................................. 19
       3.5.2 Collaboration ............................................. 22
       3.5.3 Making it work ........................................... 22

4. DESCRIBING RESOURCES ............................................... 24
   4.1 The concepts .................................................... 24
   4.2 What the technologists have been doing ....................... 25
       4.2.1 Basic metadata in HTML .................................... 25
       4.2.2 Uniform Resource Characteristics (URC) ................. 26
       4.2.3 Platform for Internet Content Selection (PICS) .......... 27
       4.2.4 Channel Definition Format (CDF) .......................... 27
       4.2.5 Meta Content Format (MCF) ............................... 28
       4.2.6 XML Web Collections and XML-Data ....................... 28
       4.2.7 Resource Definition Framework (RDF) ..................... 29
   4.3 What the user community have been doing ....................... 31
       4.3.1 Dublin Core ................................................. 31
       4.3.2 The Warwick Framework .................................... 36
       4.3.3 The DOI and metadata ..................................... 37
   4.4 Conclusions ........................................................ 38
       4.4.1 Consensus and collaboration ................................ 38
       4.4.2 Making it work ............................................. 38
       4.4.3 Dealing with diversity .................................... 39
       4.4.4 In conclusion .............................................. 41

5. IMPLEMENTING THE SCHEMES ......................................... 42
   5.1 Software .......................................................... 42
   5.2 Services .......................................................... 43

6. LOOKING AHEAD ....................................................... 45

REFERENCES AND LINKS TO OTHER RESOURCES ....................... 47

SOME OF THE CONTRIBUTING ORGANISATIONS .......................... 55
1. INTRODUCTION

The World Wide Web is now one of our primary information repositories - a vast
digital library of documents, software, images, and so on, covering a multitude of
subjects and application areas. Still, as we browse the Web, we tend not to dwell
on the technological infrastructure that makes it possible.

Even so, not only is this infrastructure fundamental to the operation of the Web,
but to ensure that the Web continues to deliver information to us, efficiently and
reliably, it will need to evolve. The core mechanisms, such as Internet Protocol
(IP), are developing to accommodate future growth and future needs. At the
application level, initiatives are also under way to impose a more coherent
structure on the information resources themselves.

These latter initiatives aim to provide better ways to identify and describe the
resources that are available on the Web. And in case it’s not immediately clear
what that means, we will take a little time to explore the concepts, before looking
in more detail at what they are doing.

In fact we now have a number of initiatives, some of which are happening
concurrently, and some of which are addressing the same issues but with
different objectives and using different terminology. And this means that it can be
hard to get a clear picture of what is happening. Here you will find an overview of
the current state of affairs, focussing on the more significant developments and
highlighting any apparent problems. There will be no excursions into the depths
of arcane technical details, nor is there any aim to be exhaustive or authoritative.
Rather, the aim is simply to be informative - to describe what is being done, to
pass comment on what is being done, and to indicate what remains to be done in
the future.
2. SOME BASIC PRINCIPLES

Identification and description issues are not new - philosophers have ruminated over these ideas for many years. Imagine, for instance, that you are looking at a tree. You could probably readily identify it as, say, an oak tree, and you could probably describe it reasonably well. Now imagine that you had to identify that particular oak tree; how would you do that? And if you succeeded in doing so, how could you be sure that your specific identification of this particular tree will still be valid tomorrow, or in a hundred years?

This may seem a trivial, and possibly obscure, example but it is one with an underlying complexity which has engrossed metaphysicists over the centuries - what constitutes the identity of something, how can you describe it, what makes it unique, and how does that unique identity persist?

Metaphysics may not have found the perfect answers, but identification and description of the things around us, and indeed of ourselves, are part of our everyday life. Cars have registration numbers, products have bar codes, people have social security numbers. And now similar issues are being tackled for the World Wide Web and its resources. How can we accurately identify and describe Web resources, so that we know what they are, and provide means to locate them, so that we know where they are. Without these capabilities, the Web would be impossible to use. At present all we have in common use is the trusty Uniform Resource Locator, the URL, but as the Web grows ever larger and more complex, will this be enough?

Consequently, the cognoscenti of the World Wide Web are trying to define workable schemes to provide these capabilities.

2.1 Resources, objects, creations ....

Trees, bicycles, books, Web pages - all are examples of things that have some meaning or value to us. In our current context, books and Web pages have something else - they contain information, or to use current parlance, they have content. In the digital world, the terms that we use to define such things are object or resource. At root, we could say that these terms signify pretty much the same thing - an identifiable entity with certain attributes. Furthermore an author, or perhaps an organisation, must have created them, so in some application areas the concept of a creation is also useful.

Resource, object, entity, creation - this terminological diversity can be confusing, but we have to live with it; the name game is all too common in information technology. Finding clear definitions, nonetheless, is difficult.

The International DOI Foundation (IDF) [org1] describes its digital objects as ‘a meaningful sub-set of Internet Resources’ [i1]. Even so, it does not see the term resource itself as well defined, and refers to the definition in ‘A Beginner's Guide to URLs’ [i2]:
‘a file in a directory, [which] can exist on any machine on the network, ... and might not even be something as simple as a file: URLs can also point to queries, documents stored deep within databases, the results of a finger or archie command, or whatever.’

In truth, while the definition was probably adequate at the inception of the Web, this latter document is now quite old. More recent definitions of a resource have been less restricted, as in RFC2396 from the Internet Engineering Task Force (IETF) [org2]: ‘A resource can be anything that has identity. Familiar examples include an electronic document, an image, a service (e.g.: “today’s weather forecast for Los Angeles”) and a collection of other resources’ [rfc2396].

Even so, this definition is not particularly precise or meaningful in real terms. The point, perhaps, is that nobody has yet offered unequivocal and accepted definitions. The IDF notes: ‘the distinction of Internet Resource, Digital Object and Creation is not a wholly clear one. They are overlapping but not mutually exclusive sets.’ [id1]

Perhaps we tend to know intuitively. A document, a software program, an image – these are Web resources. For the sake of completeness the various initiatives might need to propose their own definitions for their own purposes, or might need to consider a blank document, for example, as a Web resource. Ultimately, however, we could perhaps think of a resource on the Web as we would any other – it is something we can use.

### 2.2 Identifying, locating and describing

Registration numbers for cars, social security numbers for people, ISBN numbers for books - these are all identifiers, which follow some form of consistent standard. Sometimes the standard is global in scope, such as the ISBN, sometimes local or national, such as registration numbers. Expressed another way, they conform to a particular scheme. In the same manner, if we want to reliably identify resources on the Web, we need a consistent means of doing so - we need a reliable identification scheme.

Giving an identifier to something in principle provides it with a form of unique name or identity, such as the bar code on a tin of soup. This might also provide us, directly or indirectly, with a means to locate it, such as a TCP/IP address. Alternatively the locator might be distinct from the identifier.

Neither, however, necessarily tells us what the thing consists of. In other words, they do not provide the information which describes it; they do not provide its attributes. Attributes, of course, vary depending on the thing being described - creation date, engine size, number of pages, percentage fat content, and so on. Equally, the number of attributes varies, and for more complex entities the number can be enormous. So for practical purposes, we tend to collect together a set of attributes that is sufficient for our needs. On the Web, this descriptive attribute set for a resource is what we call its metadata.
Inevitably perhaps, in the digital world the schemes tend to concentrate predominantly on documents or, let’s say, works with an intellectual or creative content, so we can include musical compositions, images, and so on. The BIBLINK project maintains a useful list of identification schemes and initiatives [i5] which reflects this tendency, but it does include one or two other flavours, such as the EAN bar coding scheme and a digital data identification standard for television.

Other industries, such as engineering, have recognised the need for robust schemes for identifying and describing the resources that they deal with, and often incorporate these into more wide-ranging standards. An example is STEP, the International Standard for the Exchange of Product Data (ISO standard 10303), designed to provide a complete and computer-readable definition of the physical and functional characteristics of a product through its life cycle. This is a comprehensive standard covering everything from protocols for drafting, design and assembly through to a detailed modelling language. Products of course are made up of component parts, and themselves typically belong to a hierarchy, so STEP necessarily includes means of identifying these objects and defining their attributes.

As another example, the geographic information domain has similar needs, and is preparing a standard (ISO 15046) which will include schemes for cataloguing, referencing and metadata.

As the World Wide Web is pervading more and more aspects of our commercial and professional activity, we will begin to encounter many of these various schemes on the Web. Here we will concentrate on those which are specifically focussed on Web resources.
3. IDENTIFYING RESOURCES

3.1 Some concepts

The number of accessible resources on the Web increases daily, and we need to be able to identify them - to give them, let us say, specific names. But first there are some basic concepts that we need to think about.

Number one - uniqueness. A name will be practically useless if it does not uniquely identify the resource in question. If two or more resources have the same name, how do we know which is the one that we want? The name, or identifier, given to a resource should belong only to that particular resource. Should the resource have only that particular name? Not necessarily – there may be practical benefits in allowing more than one.

Number two - persistence. We cannot expect that the name will be much use to us if it is only valid for a certain period of time. If the name becomes invalid, how do we find the resource? And because we may never know how long we will want to be able to find it, it should keep its name, in principle, forever. In other words, it should be persistent.

Both, of course, are fine and necessary concepts, but neither is wholly straightforward in practice.

3.1.1 Uniqueness

Here, we first need to consider exactly what the resource is. A book has its own unique International Standard Book Number, but this does not mean that there is only one, unique, copy. Moreover, the ISBN denotes a particular publication of a particular work. Another publication of the same work will have a different ISBN; we can say that it is a different manifestation of the same work.

And so it is with Web resources. Presuming that this paper, as a work, is unique does not mean that there will only be one manifestation. Apart from any printed version, on the Web there could be a copy in HTML format, another in PDF format, and so on. Should they all have the same name? To answer this, we must ask what identification means.

A resource available on the Web needs to be uniquely identified, in the sense that its intellectual or informational content is unique, but this is not to say that it can only be found in one unique place or in only one manifestation. We might find a document on the Web at, for instance:

http://www.somewebsite.com/thisdoc.html

or at

http://anotherwebsite.org/docs/samedoc.htm
We also have to ask what constitutes a unique resource. If this report is revised, should it retain its unique identifier? If it became the intellectual property of another body, should its identifier then change? This issue can be related to the form of the identifier, of which there are two basic types.

‘Simple’ or ‘dumb’ identifiers offer no intrinsic information about the resource; they consist essentially of a structured - and uniquely assigned - but functionally meaningless string of characters. By contrast, ‘compound’ or ‘intelligent’ identifiers might contain certain information about, for example, the publisher concerned, the date of publication, and so on. So if this report has an intelligent identifier containing references to the publisher and publication date, then any revision or change of publisher would warrant a change of identifier.

Also, resource content is not necessarily permanent. We can say that, for instance, an agreed standard such as an RFC from the IETF is a fixed resource, but others may be dynamic. If we assign an identifier to a resource offering ‘today's weather forecast for Los Angeles’ then we want this to be unique, so that we can always use that identifier for this particular weather forecast, even though the content of the resource might be changing daily.

Nevertheless, at root these issues are not technological; rather, they are issues of definition and of policy, concerned with the rules for unique identifier assignment, and they are still being debated.

3.1.2 Persistence

Persistence can also have its problems. Just how long should an identifier persist? Is there really a forever? There is talk of identifiers needing to exist for several centuries into the future, but of course we can have no idea what the technology will be over that considerable timespan.

This might not necessarily be a major problem, however. James Miller of the W3C, underlining that good engineering design is the hallmark of the Web, insists that ‘To an engineer, there is no forever. Instead, there is a fixed lifetime and a mechanism for moving forward before that lifetime expires’ [i4]. This at least means that whatever schemes may come in the far-off future, all that matters is that there will be a means of migrating existing (centuries-old?) schemes into the new ones.

But whatever is meant by forever, it has been argued that unique identifiers should be persistent in that they will last longer than the availability of the resource that they identify. This might seem a little redundant, but in a sense a persistent identifier should be a permanent identifier – once it is assigned to a resource, it is always assigned to that resource, even if the resource disappears, and it will never be reassigned to another resource. And remember that the end of availability of a resource, which has been widely distributed across the Web, might not be something that can be easily recognised. In any case, to take the converse, if the resource outlasted its identifier, then the identifier would ultimately have no practical benefit.
Note also that, while in principle intelligent identifiers offer some meaningful information about a resource, the existence of this information in the identifier renders it less persistent; consider the examples given above about revision and change of publisher. Dumb identifiers have an inherent advantage in this respect. Tim Berners-Lee, Director of the World Wide Web Consortium, points out: 'if you put information in a name, it decreases its longevity' [i5]. Others suggest that in any case intelligence is on the way out: ‘In the longer run, we recognise that the argument about intelligent and unintelligent numbers is likely to come down in favour of the essentially random, unintelligent number - the only intelligence being incidental to mechanisms devised to ensure uniqueness’ [i6].

3.2 What the technologists have been doing

The initiatives of the technologists in the identification arena have largely been though the IETF [org2], typically resulting in one or more Requests for Comment (RFCs), the quasi-standards of the Internet. And it was Tim Berners-Lee's concepts of HTTP, HTML and the URL which laid the original foundations of the World Wide Web. While it is not useful to delve into all of these concepts here, a few words on the URL would not be amiss, in view of its fundamental role in Web resource access.

3.2.1 Uniform Resource Locator (URL)

The Internet hinges on the Internet Protocol (IP), which provides the basic addressing and routing mechanisms to allow us to locate a computer on the Internet, and hence on the Web. But IP, with its strings of digits, was not designed for the common man, nor was it designed to provide an easy means to access files on a computer.

The ‘visible face’ of the World Wide Web as we know it today is the Uniform Resource Locator (initially the Universal Resource Locator), which allows IP addresses to be represented by more meaningful character strings, and allows reference to directories and files at the computer located by the IP address. The URL in turn is dependent on the Domain Name System, a controlled resolution service (more on this concept later) which provides a ‘look-up’ mechanism to map a URL to the corresponding IP address.

**URL syntax**

It is unlikely that the format of the URL needs any explanation here, but for completeness we can look at the structure:

http://www.ncsa.uiuc.edu/demoweb/url-primer.html

This URL defines an HTML document named url-primer.html, which can be accessed by an application using the HyperText Transfer Protocol (http), and is located in the directory /demoweb on the computer defined in the Domain Name System as www.ncsa.uiuc.edu.
What the URL appears to provide is a distinctive and memorable name for an Internet resource, so we could think of a URL as an *identifier* for a Web-accessible resource. But is it? The URL provides a unique, possibly distinctive and possibly memorable, name to *locate* a Web resource, although identical resources might be found at other URL-referenced locations. Whether a URL can be seen as, or employed as, an identifier has become a subject of debate, and we will return to this issue later.

And the URL is not *necessarily* persistent. We often find ourselves in what has been called ‘404 limbo,’ reading the well-known Error 404 informing us that the URL was Not Found. Some estimates have put the average life of a URL as 44 days, so it seems to score low marks for persistence. Remember, however, that this is not the result of any technological constraint - it is the people and organisations creating and administering URLs who are responsible for ensuring their persistence.

### 3.2.2 Uniform Resource Identifier (URI)

Confronted with a diversity of schemes for Web resources, the IETF addressed the issue of providing an all-embracing definition for all the existing or proposed schemes - http, ftp, mailto, urn, urc, etc. This is the idea of the URI, and for a basic indication of what it signifies, the title of the original IETF Request for Comment [rfc1630] rather says it all:

> ‘Universal Resource Identifiers in WWW - A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web’

Thus the URI aims to provide syntax rules which can encompass any name or address scheme for the Web, including those which will arise in the future.

An all-embracing scheme, and one which is not short of alternative definitions. The more recent ‘Uniform Resource Identifiers (URI) : Generic Syntax’ [rfc2396], issued in March 1998, defines a URI as simply ‘a compact string of characters for identifying an abstract or physical resource.’ The W3C offers ‘the generic set of all names/addresses that are short strings that refer to resources’ [i7]. Others are more prosaic: ‘URIs are the union of URLs ...., URNs ..., URCs ... and whatever else the URI working group of the Internet Engineering Task Force decides to throw into the pot.’ [i8]

For our purposes, perhaps, the URI is indeed the pot into which the schemes are mixed according to a well-defined, but complex, recipe [rfc2396]. For the proponents of specific identification schemes in the technology world, the principal ingredient is surely the URN.

### 3.2.3 Uniform Resource Name (URN)

So what is a URN? The seminal definition from the IETF: ‘The purpose or function of a URN is to provide a globally unique, persistent identifier used for recognition, for access to characteristics of the resource or for access to the resource itself.’ [rfc1737]
Unique? Persistent? Wait a minute, isn’t this what we were looking for? And there’s more - it can not only get us the resource, but it can get us its characteristics, the information about the resource.

And there is still more, because the RFC imposes other requirements on URNs. They must be future-proof - meaning that they are scaleable, and thus applicable to any network resource well into the future, and they must support any potential future extensions. They must support existing ‘legacy’ naming systems, such as ISBN, and they must allow resolution into corresponding URLs, where these exist.

The RFC explains further: ‘A URN identifies a resource or unit of information. It may identify, for example, intellectual content, a particular presentation of intellectual content, or whatever a name assignment authority determines is a distinctly nameable entity.’ In other words, it’s a name, but you can choose what it names.

So the URN, in concept, sounds promising. It is defined as an identifier that is unique, persistent, future-proof, and all of those other good things, and it can identify pretty much whatever we want it to. The aim is that the URN can accommodate a variety of independent identification and naming schemes, each of which would have its own naming conventions and syntax. But do all these objectives mean that it is horribly complicated? In technical terms, no. The syntax of the URN itself [rfc2141] is quite simple.

**URN syntax**

URN:NID:NSS

URN is the required prefix indicating that this is a URN
NID is the Namespace Identifier
NSS is the Namespace Specific String

The first variable component is the Namespace Identifier. A namespace in this context is at its simplest level a controlled set of names, conforming to a defined syntax laid down by the organisation assigning the Namespace Identifier - the name assignment authority, or Naming Authority. The NID defines the namespace - or, to put it another way, the naming scheme - that is used for the particular identifier which forms the last component, the Namespace Specific String. The NSS must conform to the rules of its namespace.

It may be clearer to look at some examples (largely taken, with due acknowledgement, from Renato Iannella’s paper ‘The User Interface of URNs and URCs’ [i9]):

- urn:isbn:0-550-10255-8: URN using the existing ISBN scheme
- urn:inet:dtsc.edu.au:tr008: URN for an Internet resource
- urn:telecom:61733654310: URN using a (hypothetical) phone number scheme

In the first example, **urn** identifies this as a Uniform Resource Name; **isbn** identifies the NID as the International Standard Book Number Scheme; 0-550-10255-8 is the NSS, in this case the ISBN code.
Note that, provided that it conforms to the syntax of a URN (avoiding ‘illegal’ characters, for example) the NSS is independent of the URN structure, depending for its own structure only on the naming scheme identified by the NID.

Also, it is designed only to identify the resource, not to locate it. This requires that there must be a means of retrieving information on the resource including its location. Thus, like the Domain Name System there is a need for a resolution service, to map from the URN to the locations of the resource or, indeed, to other information or services related to the resource.

The principle is that once a URN is assigned to a resource, users will always be able to use it to reference that resource. The resolution service would return what is essentially the metadata of the resource, which would include a URL - or multiple URLs - that could then be invoked to retrieve the resource. Equally, the metadata could be used for other purposes - linking to other resources, providing rights information, and so on.

So the claimed benefit of the URN, through its use of a resolution service, is that the URN can remain unique and persistent. The perceived volatility of URLs is avoided, with any changes to the locations of the resource or to its metadata requiring only a change to the resolution database.

Clearly, there are some basic organisational prerequisites for a practical URN implementation. There must be agreed authorities that will have the responsibility for allocating the namespace identifiers and others for controlling the namespaces themselves (the Naming Authorities); and there will have to be effective resolution services available.

Neither of these requirements is, in principle, problematic. There are well-established examples of controlling authorities such as the Internet Assigned Number Authority, with the responsibility for IP address assignment; and the Domain Name System is an operational resolution service. In fact in the Internet Draft “URN Namespace Definition Mechanisms” [id1], the IANA is been suggested as the body to control the assignment of NIDs.

Moreover, it has often been pointed out that the technology infrastructure is more than capable of dealing with all the technical requirements. It is the organisational and social issues that have to be ironed out before any real deployment can happen, and this is certainly true for NID assignment, the creation of resolution services and for maintenance of the metadata. Only when these are in place will we see a practical implementation of unique and persistent URNs.

How close are we? Not very close at present; there are many hurdles to overcome, particularly concerning the establishment of the organisational infrastructure. Nevertheless there are, or have been, some trial URN implementations - TURNIP [s2], mu [s3], and BURNS [s5].
3.2.4 Persistent Uniform Resource Locator (PURL)

Although deployment of a true URN service remains some distance into the future, in the meantime the Online Computer Library Center (OCLC) [org5] have developed the Persistent URL, or PURL.

A PURL remains a URL in form, in fact it looks just like a URL, but it points to an intermediate resolution service [org6] rather than directly at the resource in question. This service then retrieves the actual URL.

**PURL syntax example**

```
http://purl.oclc.org/OCLC/PURL/FAQ
```

- `http` identifies the protocol, as for any URLs
- `purl.oclc.org` identifies the resolver, or resolution service
- `OCLC/PURL/FAQ` identifies the address name, used to retrieve the true URL

Like URIs, PURLs can be assigned to changing resources, such as ‘today’s edition of the Wall Street Journal.’

Persistence, of course, is a fundamental aspect of PURLs. Changes of location of the target require only a change to the PURL database - the PURL itself remains valid. Moreover, for access to Web resources, the PURL uses existing and established HTTP mechanisms, so it works with the installed browser base.

This is not meant to be a long-term solution. OCLC itself sees PURLs as an intermediate step towards the time when URNs are an integral part of the Internet information architecture [i11]. Indeed OCLC are an active participant in the URN development process.

But while it is readily admitted to be an interim solution, it has been used operationally; by August 1998, nearly 190,000 PURLs had been created, and some 3,500,000 resolutions had been performed. Small numbers, perhaps, on the URL scale, but significant nonetheless.

### 3.3 What the user community has been doing

The Internet world - essentially the IETF and the W3C - have been instrumental in progressing the developments described above, but while they recognise the non-technological issues that have to be faced, their interest perhaps remains the production of a technologically coherent solution. In other quarters, there are more commercially and practically driven motivations, and consequently other identification schemes have arisen.

The prime movers here have been principally the publishing and libraries communities, who have a need to deal with intellectual property and copyright, e-commerce, bibliographic information, and so on. Just as the International Standard Book Number and International Standard Serial Number arose from
these communities as identification schemes for physical publications, and have made publication trading and library management so much easier, so they have now turned their attention to digital resources.

The needs here are arguably far more complicated. Remember the earlier discussion on works and manifestations. For example, a novel - such as Charles Dickens’ *Hard Times* - can be considered a work (or maybe a creation) and should be uniquely identified as such. But then there could be many different manifestations of that work - different publications of that novel by different publishers. There could then be different versions of each manifestation - revised reprints, perhaps. Plus we have the question of granularity - should we stop at simply identifying the work (or the manifestation), or should we provide means to refer to a specific chapter of that work? The issues can become still more complex, but this should suffice to highlight the problems.

So not surprisingly, in this user community different schemes abound, all trying to address some or all of the problems. Amongst these we find the PII (Publisher Item Identifier), BICI (Book Item and Component Identifier), SICI (Serial Item and Component Identifier), and ISWC (International Standard Work Code). But when we talk of Web resources, one stands out from the rest - the DOI.

### 3.3.1 Digital Object Identifier (DOI)

The DOI is designed as an identifier for all digital content. Like the URN and the PURL, it aims to be both unique and persistent. More than the URN or PURL, however, it is also seen as ‘an enabler for processing some routine transactions such as document retrieval, clearinghouse payments, and licensing’ [i12]. So it is much more application-specific, and much more commercially oriented.

The DOI was developed originally by the American Association of Publishers and the US Corporation for National Research Initiative (CNRI) [org8], but it is now gaining support and input from other corners of the globe. The co-ordinating body is the International DOI Federation (IDF) [org1] which, along with the DOI itself, was formally launched at the Frankfurt Book Fair in October 1997.

When a DOI is accessed, perhaps by clicking on a DOI icon or button on a Web page, the DOI is resolved by a central directory service which returns a Web location. The resolution technology is currently based upon the Handle system [i13], with the resolving service operated by CNRI, although the DOI definition allows other resolving systems to be used.

The rights owner decides on the content at the Web location. It could be the ‘digital object’ itself - for instance a magazine article - or information about the object, or an order form for acquiring the object. This flexibility allows the rights owner to offer more to the user, who could be directed to related material, to order the object, to refer to other works by the same author, and so on. Equally, if the ownership of an object changes, the new owner registers the change with the central agency. If the object moves location, the DOI entry for the object is updated. In this way, the DOI preserves its uniqueness.
DOI syntax

The DOI syntax consists simply of a prefix and a suffix. Publishers - or more generally, rights owners - are assigned a unique prefix by the Directory Manager (see below). The suffix is the ‘designation assigned by the publisher to the specific content being identified’ [112]. The rights owner is free to choose whatever structure he chooses for the suffix, and equally can assign it to anything he chooses - a book, an item of software, and so on. Often, publishers adopt existing standards for the suffix. Two examples of DOIs:

10.1002/123456

Prefix: Within the prefix, 10 identifies a particular Directory, and 1002 the registered publisher or rights owner. The full Directory structure has not yet been defined; it could be that different Directories are set up for each country, or for each industry sector.

Suffix: In the first example, a dumb identifier 123456 has been used as the suffix, whereas the suffix in the second example uses [ISBN] to represent a code, in this case an International Standard Book Number 0-471-58064-3 as the item identifier. Thus, other schemes such as ISBN or PII can also be accommodated with the DOI structure.

In addition, the DOI identifier itself can be expressed as a URN, such as (potentially):

urn:doi:10.1002/123456

although there are possible conflicts with URN syntax, but these could be overcome relatively easily. For example, the “/” is a reserved character in the URN syntax, but could be used within the DOI suffix in an encoded (‘escaped’) form:

urn:doi:10.1002%2F123456

DOI implementation is in its very early stages, and at present, all DOIs map onto URLs. Moreover, a DOI can only point to one URL, although more than one DOI can point to the same URL. This might not always be the case; the IDF recognises that in order to extend the scope of the DOI, extended mapping would be required to provide the flexibility discussed above. The current mapping principles concern what have been termed ‘level 1’ DOIs.

The next step would be to allow the more flexible resolution of a DOI noted above, into a number of data types - the content itself, or services such as ‘describe’ to retrieve metadata, ‘purchase’ if one wishes to buy, and so on. This is the idea of a ‘level 2’ DOI.

There is also a proposal to add an ‘argument’ or ‘option’ to the DOI syntax, to allow the DOI to indicate a specific action, such as ‘!buy’ to generate a purchase request. Nevertheless, these extensions to the DOI scheme will require a major development program, and we are unlikely to see them in the immediate future.

And like the URN, the DOI needs a supporting organisational structure to be defined. Currently the IDF is responsible for licensing the Directory Managers, who will assign prefixes, and for the overall management of the DOI, but it has been suggested that in the longer term an established body such as the International ISBN Agency should take over management responsibility.
The principal difference between the DOI and the identification schemes from the technology camp is its more commercial objective, to assist in electronic trading and other services. Nevertheless the current prototype is simply a resolution service to retrieve a URL from a DOI. The IDF accepts this: ‘In essence ... the DOI prototype offers little more than other “get” redirection services for persistent URLs; to achieve the full potential we need to offer mechanisms which can offer a variety of services’ [i1]. So it still has a long way to go.

The European Commission’s Libraries sector [org9] recognises the importance of the DOI and hosted a DOI workshop in May 1998 [i14]. This workshop concluded that further work, in particular more practical work, was needed, along with better liaison with other sectors of the information community and the various standards bodies.

### 3.4 Other initiatives

There are a few other initiatives that, while not perhaps in the mainstream, do concern themselves, directly or indirectly, with identification of Web resources, and as such are worthy of a brief mention.

#### 3.4.1 WEBDAV and the UUID

The IETF’s World Wide Web Distributed Authoring and Versioning initiative (WEBDAV) [i17] was set up last year, and aims to define the HTTP extensions needed to allow distributed Web authoring tools to be interoperable and meet user needs. These extensions would provide capabilities for namespace management, metadata, document locking and version management.

Document locking requires that not more than one person can update the same document at any given time, and this will be achieved by the generation of unique lock tokens. The mechanism used is a Universal Unique Identifier (UUID) for the token. This identifier is “either guaranteed to be different from all other UUIDs generated until 3400 AD or extremely likely to be different (depending on the mechanism chosen).” So within its own domain, the UUID aims to be unique and, if it remains so for about one and a half millenia, then presumably we can consider it as persistent for all practical purposes.

#### 3.4.2 Persistent Document Identifiers (PDIs)

Persistent Document Identifiers are another IETF initiative, “intended primarily as permanent identifiers for archival reference to long-lived documents” [id2]. Thus PDIs do not aim to be universally applicable identifiers; rather, the scheme will concentrate on providing a specific means to refer to digital objects, and fragments of these objects, independently of their storage location and the protocol used to access them. They are nevertheless compliant with the URN syntax.

There have been a few large-scale implementations of PDIs, notably the White House Electronic Publications System.
3.4.3 Human Friendly Names

The mainstream work on unique and persistent identifiers for Web resources may turn out to be fundamental to improving the Web as a reliable and effective information repository in the long term, but what will the end result be? Could it be that, instead of typing in a URL, such as:


we will type in a URN, such as:

urn:inet:dtsc.edu.au:tr008

These are fine for a computer, and either might be OK hidden behind a bookmark in our browser, but neither is particularly snappy, meaningful or memorable. What we really need is a name which is easy for a human being to remember and which readily signifies what it is supposed to identify, in other words a Human Friendly Name (HFN).

Discussion on HFNs has only recently started, and so we are still far from any definition of what they might look like or how they would work in practice. These are very early days for the HFN, and all we can say is that they have begun to be talked about as a Very Good Idea. The principle, however, is that they will need to map to URNs, and the mechanisms for this will need to be defined.

The interesting point is that, while URNs are designed to be formalised, globally unique and persistent, an HFN will be a relatively informal device, perhaps used only within a certain geographical area or for a certain period of time. So maybe, if you are in Los Angeles, and on a certain day you type in the HFN “Weather” then you will receive the Los Angeles weather forecast for that day; if you are in London, then the same HFN would provide you with the forecast for London.

RFC2276, while declining to enter into a detailed discussion of HFNs, talks of ‘a variety of human-friendly naming (HFN) schemes supporting different suites of applications and user communities.’ [rfc2276] Significantly, it suggests that they will definitely have an important role in the future, in that users will use only HFNs - not URLs, not URNs - to access Web resources. This importance is echoed in other papers: ‘the construction of an HFN infrastructure should be a priority in the Internet’ [i17]. Impressive endorsements for the HFN concept, perhaps, but there are only one or two real signs of activity on this front at present.

Netscape’s ‘Smart Browsing’ feature in recent releases of Navigator allows, in principle, the entry of a ‘friendly’ name, but it does not directly retrieve the corresponding Web site or page. Typing, say, ‘European Commission’ in place of a URL invokes a search engine which will return its best matches to the words entered. Thus you may find yourself ultimately at the EC’s Web site, but there is no direct resolution of the name entered.
By contrast, Centraal is offering its Real Name system [i18] as a genuine means of using human-friendly names to access Web resources, and has submitted a proposal to the IETF. The system essentially offers a resolution service on top of the Domain Name System, based upon a central database of registered ‘Real Names.’ There has been some progress in implementation; Centraal has signed agreements with the LookSmart and, notably, AltaVista search engines to allow Real Name searches alongside conventional ones. Even so, for large-scale deployment, Real Name functionality would probably need to be incorporated into the leading browsers, but with Netscape, for one, having its own ideas in this context, this might not happen unless the system begins to attract market acceptance by other means.

### 3.5 Conclusions

The world seems to want a unique and persistent identifier, and to date quite a variety have been proposed. This observation is not intended as a criticism - with the diverse application needs, a diversity of identification schemes is almost inevitable. Nevertheless, perhaps there is now some convergence in sight. It may be that the original problem of proliferating standards ‘appears to be receding as different groups meet to seek consensus and discover that the issues we have in common are much more significant than the issues which divide us.’ [i6]

#### 3.5.1 Consensus

The view just quoted comes from the publishing and library camp, and it may be that for them a workable consensus is in sight. Elsewhere, however, there is less harmony.

Within the W3C, there is a good deal of scepticism about the real value of the various new identification schemes. On the Web, we already have a number of schemes within the URI framework, notably of course the URL. The URL identifies a resource – it gives it a name; it also locates it – it gives it an address. So we have the question: “Should there be separate mechanisms for names and addresses?” The W3C, for one, thinks not, and for them the issue is a long-standing one.

Consider the two views of the URN offered by the W3C [i7]:

1. An URI that has an institutional commitment to persistence, availability, etc. Note that this sort of URI may also be a URL ....
2. A particular scheme which is currently under development in the IETF ..., which should provide for the resolution of internet protocols of names which have a greater persistence than that currently associated with internet host names or organizations.’

The first presents a fairly generic URN - a PURL, for instance; the second is the more precisely defined URN discussed earlier. Connolly comments that the first type is compatible with the deployed infrastructure; it could be improved, but it works as it is. However, ‘URNs in the 2nd definition require deployment of new
ubiquitous software. I haven't seen any schemes that justify that cost with technical improvements.' [i16].

For the W3C, the URL mechanisms would be more than sufficient, given some improved organisational support. Connolly suggests that the newer schemes propose [i16]:

‘(a) some new technology
(b) some new administrative infrastructure.’

and claims that ‘(b) is all you need, and (a) is a waste of time. … If we can kill two birds (naming and addressing) with one stone (URIs) then why not?’

In other words, on the Web a name is an address, and vice versa. This is the point made by W3C Director Tim Berners-Lee in his paper ‘The Myth of Names and Addresses’ [i5].

To this we could counter that the URL has some known deficiencies. For instance, it might locate a resource uniquely and efficiently, but how can it uniquely name a resource that exists at more than one location on the Web? This is something that the URN, for example, is designed to accommodate. Even so, the URN achieves this by presenting a list of URLs where the resource can be found; if we typed in a unique URL identifying the resource, and were presented with the same list, does this not achieve the same end? What added value does the URN bring?

Secondly, even if we accept that the URL is unique, we have seen that its capacity for persistence is somewhat suspect – remember the estimate that its average life is 44 days. But remember also that this is never the fault of the URL mechanism itself; as the PURL camp points out [i10]: ‘persistence is a function of organisations, not technology.’

Connolly [i16] echoes this view: ‘To get persistence, there are lots of contractual issues … but no technical ones.’ And also:

‘Witness the W3C technical reports archive:

    http://www.w3.org/TR/

    We handle all the issues of persistence, versioning, rights management, etc., all with URIs starting with http://....

    In other parts of our web site, we're less rigorous, and indeed, in the rest of the web in general, you get no such guarantees. ....

    But we're not the only site to administratively ensure persistence. Look at http://www.purl.org/ where OCLC is doing likewise.’

Indeed, the PURL initiative is designed to provide unique and persistent identifiers using the established URL mechanisms. Even so, the PURL people must see some value in the URN, since they see the PURL as an interim solution until the URN arrives.
Maybe the most significant characteristic of independent identifiers, such as the URN and the DOI, which sets them apart from the URL is that, unlike the URL, they deliberately do not incorporate any addressing or locating functionality. They are designed solely as a device for providing a name. The International DOI Foundation make this point – while a DOI might be used on the Web, it has an independent existence which makes it useable in other contexts. The URN camp also underlines this distinction [rfc1737]:

‘URNs are used for identification, … and URLs for locating or finding resources. … A URL identifies the location or a container for an instance of a resource identified by a URN. The resource identified by a URN may reside in one or more locations at any given time, may move, or may not be available at all. Of course, not all resources will move during their lifetimes, and not all resources, although identifiable and identified by a URN will be instantiated at any given time. As such a URL is identifying a place where a resource may reside, or a container, as distinct from the resource itself identified by the URN.’

The perceived complexity of effective identification might have been a major motivation for the development of specific schemes. The IDF point out [i1]:

‘… no single identifier is capable of serving all purposes; an identifier for a Package will have different requirements than the corresponding Work. A bookseller needs to separably distinguish the different ISBNs relating to a given title, whereas a reader may wish to consider all ISBNs of the same Work as indistinguishable; therefore E-commerce will require a network of related (linked) identifiers. The linkage between identifiers could be by direct referencing (resolving one identifier to another) or by metadata: for example, metadata which lists derivatives of the same Creation.’

The implication, perhaps, is that no existing scheme, such as URL, could deal with this functional complexity, hence the need for something new.

So it appears that we have persuasive arguments from the W3C that, at least on the Web, the URL - and more generally, the URI - can essentially meet all of the needs for naming and addressing. And on the other side, supporters of separate identification schemes are saying that these needs should be differentiated.

It would not be useful here to attempt to predict which approach will prevail, or indeed which is the best. In any case, as so often in the IT world, “best” does not always imply “most successful.” Whatever approach comes to predominate – if, indeed, a single approach does ultimately predominate – it will probably do so as a result of its success in large-scale deployment and in meeting the needs of the user community. This in turn would, no doubt, depend on the extent to which it captures the interest of the user community and of the major players in the Web arena – the software vendors, search engine providers and so on.
3.5.2 Collaboration

Despite the disagreements described above, two schemes appear to stand out - from the technology camp, the URN; and from the libraries and publishing camp, the DOI. And although they arise from different camps, and with different objectives, there is a degree of compatibility, and the two sides do seem to talk to, and about, each other. We do tend to find more discussion of URN in the work of the DOI camp than the converse, but perhaps this arises simply from the need to take account of the more infrastructural URN developments in order to provide a working solution on the Web.

The IDF still wants to differentiate the two initiatives, however [i1]:

‘Although the W3C ... is dealing with many issues of interest to the intellectual content world, it is broader in scope and has relatively few intellectual content members. The DOI initiative takes a different - but entirely complementary - stance from W3C’s focus on the much larger universe of Internet Resources.’

and

‘One of the aims of the DOI initiative was to put in place an identifier scheme for Objects which could be readily used by the content industries and which could be used as a URN; another aim was to facilitate digital trading (these two aims are not synonymous). The aim of the DOI system is however not to be only a URN system, but to be a system which is flexible enough to use in a variety of environments.’

So they do seem to be talking, but neither side sees the current co-operation as ideal. The IDF recognises the significance of URI and URN, but does not actively collaborate in these developments. Dan Connolly, at the W3C, comments ‘There are meetings and conferences where cross-pollination happens. I'd like to see the level of collaboration improve, but it's hard to get resources dedicated to such a low-level activity..' [i16]

By comparison, the W3C sees the overall URI scheme as being suitable for meeting at least some of the needs of the libraries and publishing camp, such as catalog numbers, ‘given sufficient socioeconomic infrastructure: rights management, payment assurance, privacy, digital signatures, etc.’ [i7]. But note that some of these infrastructural prerequisites are precisely those of a commercial flavour which the DOI scheme is designed to address.

3.5.3 Making it work

All the same, whatever scheme we are talking about, the agreed view is that the technology can be devised to deal with whatever is needed, and that the real obstacles lie in the practical aspects of its implementation. There are organisational and policy issues to agree, and many people and groups are involved in this. Controlling organisations must be established, and the software and services deployed. Only when these are sorted out, can the technology be put into place. More than that, this will all cost money.
The IDF recognises the tasks ahead for its own initiative: ‘There will be unavoidable costs for the DOI system. A standard on its own will not give us an implementation, and an implementation costs money to someone somewhere; the task is to make it efficient.’ [i1]

So while the schemes are moving forward, albeit rather slowly in some cases, nobody is pretending that the road will be easy. There will continue to be a need for collaboration and co-operation. And realisation of these schemes is going to need investment - in money as much as effort, and there is a long way to go before we are likely to see any of these schemes in a full global implementation.
4. DESCRIBING RESOURCES

4.1 The concepts

We commonly use descriptive information about things in our everyday lives. Your car has a registration number, which can lead to its identification as, say, a black Alfa Romeo built in 1995 and with a 2.5 litre engine. This is descriptive information retrieved from the car’s identification. Equally, if you wished to choose a brand new car, you might stipulate that you want a salmon-pink coupé with red leather seats, a 5 litre engine, four-wheel drive and air-conditioning. This could lead you to a short-list of models which meets that specification. Here you are using the descriptive information available on cars to locate models which are of interest to you.

So it is with Web resources as well. We can assign a unique identifier to a resource, and we can define means to locate the resource from its identifier, but we also want provide information about that resource which can help us to understand what it is, to help us use it, or conversely to help us find resources which meet our needs. This information is the metadata - most frequently defined as "data about data" but nowadays in our Web context this smacks of a legacy definition. The term "metacontent" has also been suggested - ‘content about content’ - but here we are playing the name game again, and metadata remains the term in most common use. Think of metadata as, perhaps, information about information.

Whatever it is called, the concept is clear - metadata is a means to describe a resource, but is signifies more than that. Through this descriptive information we, and the machines which populate the Web, have a means to use the resource, because metadata is structured information and so can be processed, not just by us, but also by the computers that we want to set to work on the Web. A research paper available on the Web may have an Abstract - ‘This paper describes the findings of the ......’ and so on - which may well describe the document, but aside from probably offering a good selection of key words for consumption by a search engine, it is of little value in terms of making real use of the Web. What we need to be truly able to make use of resources on the Web is information that is both human-readable and machine-processable, and here metadata has fundamental roles to play:

- to help us use resources. From the identifier of a resource, its metadata can be retrieved, providing information about the resource, and can direct the user, or a computer, to the resource itself or to related resources or services.

- to help us find resources. Users or computers can search in the metadata provided for resources, to find the resource (or resources) that have a given set of attributes, in other words which correspond to particular criteria.

The first role has already been touched on with regard to URN and DOI. The second could open up new horizons for searching on the Web, given a Web well
populated with metadata. Making use of this inherently structured information, resource discovery could become more accurate and more efficient.

So if metadata is such a wonderful thing, why is it not there already? Two reasons, essentially.

Firstly, the Web has grown organically, euphemistically meaning that the growth has not been controlled. Although its technical foundations - HTTP, HTML and so on - were reasonably well-laid, the provision of information that it holds was pretty much open to all, with no real thought of organising that information.

Secondly, metadata needs standards in order to work effectively. It needs consistency in the ways that resources are described, so that we can understand what information is being provided and so that we are therefore able to use it. And good standards do not come easily, and they do not come fast.

Implementation of consistent metadata standards will have a significant implication for the World Wide Web. They will provide what will be, in effect, a filing and cataloguing system for the Web - can you imagine trying to use a library which did not employ a cataloguing system? While the technical infrastructure of the Web is essentially coherent, and now interoperates more or less without problem, the information which it contains and which gives the Web its value is essentially unstructured.

More to the point, there will never be a realistic way of applying a consistent format or structure to the diversity of the information resources themselves. The only way to provide consistency to the resources on the Web is to provide consistent means of naming and describing the resources. If the resources can be described in a structured and consistent way, then this information - the metadata - can be operated on, by us or by computers. It would facilitate resource discovery, cataloguing and rating of content, interrelationships between content, rights management, and so on.

4.2 What the technologists have been doing

While developments in the technology camp concerning identification schemes have largely been initiatives driven by the IETF, for metadata the W3C has been much more actively involved.

And while, as we will see, in the user community one scheme in particular has predominated, there has been a number of schemes in the technology camp. But let us start at the beginning.

4.2.1 Basic metadata in HTML

In a very simple sense, metadata can be embedded into a Web page using the HTML <META> tag, and typically this is currently used to provide basic information about the page, although not in any particularly structured or consistent way.
A simple example provided by UKOLN [org10] for their own home page (with the list of keywords much reduced for the sake of brevity):

```html
<META NAME = keywords CONTENT = national centre, network information support, library community, metadata, resource discovery >

<META NAME = description CONTENT = UKOLN is a national centre for support in network information management in the library and information communities. It provides awareness, research and information services >
```

Search engines can use this information, since they will index the keywords provided within it. Nevertheless, this basic approach is fairly crude, and there has also been some abuse of the keyword concept - less honourable Web page creators can overload their pages with keywords in the `<META>` fields to attempt to achieve a disproportionately high ranking in the engine’s hit list. A more consistent and standardised approach to metadata was seen to be required.

### 4.2.2 Uniform Resource Characteristics (URC)

The identification initiative in the technology camp is centred on URI/URN and, given the recognised importance of metadata, it is not surprising to find a corresponding metadata scheme, Uniform Resource Characteristics (sometimes called Uniform Resource Citation) - the URC. The W3C provides a succinct definition [i7]:

‘URC - Uniform Resource Citation, or Uniform Resource Characteristics. A set of attribute/value pairs describing a resource. Some of the values may be URIs of various kinds. Others may include, for example, authorship, publisher, datatype, date, copyright status and shoe size. Not normally discussed as a short string, but a set of fields and values with some defined free formatting.’

The original principle of the URC was to use it to retrieve the corresponding locator URLs for a URN identifier, as discussed earlier concerning the URN, but this quickly grew to embrace the provision of other attribute data, such as bibliographic information. The URC was designed to provide the framework within which particular metadata schemes could be used, through the use of a required 'URC subtype' to designate the scheme. Given that the scheme was recognised, the user or computer could then interrogate the metadata.

Thus, instead of specifying the attributes themselves - author, title, and so on - the URC provided the structure for storing metadata and the operations necessary to handle it. The (now expired) Internet Draft ‘URC Scenarios and Requirements’ [id3] provided some useful descriptions of the potential use of URCs, and Renato Ianella's paper ‘The User Interface of URNs and URCs’ [i9] offers some clarifying examples.
However, while the delivery of a working URN scheme appears to be a realistic possibility, even if only after a long wait, the gestation of the URC has been slower. Indeed in some quarters the URC is felt to be still immature; others have said that it is still not well specified, still others have seemed to suggest that it was already still-born. The URI Working Group of the IETF, formed in 1992, led to the URN Working Group (URN-WG) in the autumn of 1996, but a URC Working Group was never approved, and the definition of URCs that was adopted by the URN-WG was not precise. Unlike the URI and URN, work on URC has never reached the RFC stage.

But the spirit, at least, of the URC lives on. Its basic design philosophy is also found in RDF, which we will look at shortly.

4.2.3 Platform for Internet Content Selection (PICS)

The W3C’s first major venture into metadata was the Platform for Internet Content Selection (PICS) [m1], with the original objective of providing a form of rating system for Internet content. In a similar way to the ratings of films, which could be used to help parents, for example, to deny access by their children to unwelcome material such as pornography. This was in response to growing public concern about the accessibility of ‘unsuitable’ material on the Web. It was soon recognised that PICS had potential for much wider application in control of access to information, and for assisting in the search and retrieval of high-quality Web resources.

The PICS approach is based on the ‘labelling’ of the content. Its advantage is that it does not offer censorship of information, but merely provides it with a label, leaving the responsibility to the user (e.g. the parent) to decide what is appropriate to access. Necessarily, of course, to be effective this scheme requires that their providers appropriately rate pages, and that browsers support the system.

The W3C’s work on PICS is now complete with the establishment of PICS-1.1, and the focus has now moved to RDF, which will be able to embrace anything expressible by a PICS label; all future work on PICS will concentrate on its migration to the RDF format. The W3C state that ‘Software and Web content using PICS-1.1 will remain a supported W3C recommendation for as long as the market demands’ [m2]. Support for PICS in browsers, however, has only recently arrived - Microsoft’s Internet Explorer 4.0 integrates a PICS-compliant ratings system, while Netscape’s Communicator 4.06, announced in August 1998, provides support for two such schemes.

The source of the W3C home page [org3] shows an example of the use of PICS in an HTML page.

4.2.4 Channel Definition Format (CDF)

The Channel Definition Format [m3] was originally conceived by Microsoft as a mechanism to group Web resources together, creating a ‘Channel’ as a set of resources that can be manipulated as a unit. It saw a degree of interest when
push technology was very much in fashion a year or so ago, but that interest seems now to have waned.

Nevertheless, it is also capable of offering mechanisms to assist in searching, indexing and filtering content, and thus has made its contribution to metadata developments.

The W3C sees CDF as, at least, useful in providing a metadata vocabulary for push content. In addition, it is based upon Extensible Markup Language (XML, a significant development that we will discuss later), and the W3C feels that RDF could accommodate CDF by providing a framework to which a future revision of CDF could evolve [m2].

4.2.5 Meta Content Format (MCF)

Apple and, subsequently, Netscape were involved in the definition of Meta Content Framework (MCF), which is designed to offer a format for holding metadata externally to the content which it describes, as opposed to embedding the metadata in, for instance, an HTML page. In June 1997 Netscape submitted a proposal to the W3C on the use of XML for MCF.

As with CDF, however, RDF seems to have stolen the show, and by September of 1997 Netscape was putting its whole-hearted support behind RDF. Nevertheless, Netscape estimates that some 2000 sites have adopted MCF to describe the organisation of their content, so it maintains support for the scheme.

The document ‘An MCF Tutorial' [m4] provides an introduction to MCF.

4.2.6 XML Web Collections and XML-Data

These initiatives, in both of which Microsoft - again - has a leading role, both provide for metadata structures as applications of - again - XML.

Web Collections offers a syntax for describing the properties of a resource in a structured and hierarchical manner. Each collection uses a profile indicating a specific set of properties to be used by applications in treating the resource. For example, a collection used to describe a Web page might use a profile called WebPage which would identify the type of the resource - unsurprisingly, a Web page in this case - and indicate that there are properties (elements) such as author, date last modified, etc. A proposal for Web Collections [m5] was submitted to the W3C in March 1997.

XML-Data [m6] defines schemas, the formal specifications of element names that indicate which elements are allowed in an XML document and in what combinations. The XML-Data proposal was submitted to the W3C in January 1998.

Neither initiative appears to have had a significant impact in its own right, but both have been influential in the conception of RDF.
4.2.7 Resource Definition Framework (RDF)

And so we arrive at last to Resource Definition Framework, which although only initiated in 1997 now dominates the technology-based metadata scene.

Much as all roads lead to Rome, all metadata schemes seem to converge on RDF. RDF extends PICS, and draws on the concepts of MCF, Web Collections and XML-Data. Equally, its design has been influenced by the Dublin Core and Warwick Framework developments, discussed later, so there are many ingredients in the recipe. Moreover PICS, Web Collections, CDF and MCF can all be accommodated, although with the advent of RDF these are already being seen in some quarters as legacy approaches.

In essence, RDF is a language for expressing metadata, and is an application of - of course - XML. A fundamental objective of RDF is that is should provide for the exchange of machine-understandable information about Web resources, and thus provide the facilities for the automatic processing of these resources. Also, it is designed to be both vendor-neutral and operating system-independent.

Following the principle of the URC, it does not define the metadata vocabulary, which depends upon the type of resource. Different types of resource necessarily require different attributes, in order to describe them accurately and to provide the information which is useful and meaningful to the user. There is no realistic way for a complete metadata set to be defined which would encompass all of the required attributes. In general terms, for example, a book might have a title, author and publisher as metadata elements, which constitute part of the vocabulary in use. A song, on the other hand, might have a title, performer, and composer.

Clearly a reference point is needed for an application to understand what vocabulary is being used, therefore in RDF each vocabulary is uniquely identified. The way that this is done is through the definition of schemas, which group together for each metadata application the definition of the elements (properties), the values that these elements can have and the semantics of the elements. Each metadata scheme references a uniquely identified namespace wherein the elements are defined.

The W3C states: ‘RDF will allow different application communities to define the metadata property set that best suits the needs of each community. RDF will provide a uniform and interoperable means to exchange the metadata between programs and across the Web. Furthermore, RDF will provide a means for publishing both a human-readable and machine-understandable definition of the property set itself’ [m7].

The objective is that RDF should offer the following features [m2]:

- interoperability of different metadata schemes
- machine understandable semantics for metadata
- uniform query capability for resource discovery
• better precision in resource discovery than full text search
• a processing rules language for automated decision-making about Web resources
• a language for retrieving metadata from third parties
• future-proofing for applications as schemas evolve.

Great things are expected of RDF, finding application in areas such as:

• resource discovery, providing better search engine capability
• cataloguing, providing better description of content and content relationships for a Web site or page, or for a digital library
• use by intelligent agents to facilitate knowledge sharing and exchange
• in content rating for child and privacy protection
• for describing page collections that represent a single logical document
• for describing intellectual property rights of Web pages

The W3C sees RDF having a major role in enabling a range of new applications, such as the provision of bibliographic records and rights management. ‘Metadata will facilitate searching, helping authors to describe their documents in ways that search engines, browsers and Web crawlers can understand.’ [m8]

It also claims that RDF will aid in building the ‘Web of Trust’ necessary for electronic commerce, collaboration, and other applications. This hinges on the ability of RDF to accommodate digital signatures, which provide a means of authentication of resources, thus affording a degree of security – a prerequisite for reliable and bona fide Web transactions concerned with copyright, commercial and contractual matters.

Clearly much faith has been placed in this scheme, and indeed the major software manufacturers have by and large endorsed it. Netscape, in fact, are now quite bullish about RDF, pointing out recently that a number of major Web publishers, including CNET and the New York Times, have now deployed RDF-based site maps. A number of the leading search engine providers and on-line publishers have also stated their commitment to RDF.

Nevertheless, the ways in which RDF will work in practice, both in terms of the generation of metadata and its use by Web users and computers, is not yet clear. Much will depend upon the tools and services that will arrive to support it, and how the Web community decides to put it into practice. Moreover, much work remains to be done before RDF is fully realisable, and the W3C Metadata Project is expected to last until December 1999.

Still, it is fairly safe to say that it has a future. The UKOLN paper ‘Metadata for the Web - RDF and the Dublin Core’ [m9], which also provides a good introduction to RDF, notes: ‘What can be said is that RDF is likely to become the pervasive metadata architecture, implemented in servers, caches, browsers and other components that make up the Web infrastructure.’ How large a future depends upon the future for deployment of the metadata itself.
4.3  What the user community have been doing

The need for effective metadata was equally recognised outside of the technology camp, and again it was the library and publishing world which took the initiative. This is not surprising when one considers the essential cataloguing requirement of this domain - the most widespread cataloguing standard, MARC (Machine Readable Catalog), in essence is a metadata scheme for describing books and other publications.

With the advent of the Web and of the concept of digital libraries, the need for effective metadata for digital resources was soon realised, and the single predominant scheme is Dublin Core.

4.3.1  Dublin Core

Dublin Core (DC), or in its full title the Dublin Metadata Core Element Set, is so named due to its origins in a workshop hosted by OCLC and NCSA at the former’s home in Dublin, Ohio in March 1995. Its prime movers have largely been from the library world, and its main focus is on resource discovery on the Web via description of content generated by its author.

DC was never intended to be the all-singing, all-dancing solution to metadata needs. It seeks to define a minimal, but sufficient, core set of attributes, or elements, which can be used to provide a basic description of a resource.

Its design objectives are:

- simplicity: to allow its use by non-specialists (remember that it is focused on author-generated metadata)
- semantic interoperability: DC aims to be the lingua franca for resource discovery on the Internet
- international consensus: to ensure global deployment. To date, there has been active participation from a good number of countries, including seven from Europe
- flexibility: to embrace more formal resource description applications than the original author-generated model.

The definitions of the elements are fairly loose - ‘Date’ is not rigidly defined, for instance, as the date of publication or date of creation. All elements are repeatable, for example to denote multiple authors.

No element is mandatory; and they may appear in any order. Note also that in many cases, the use of a controlled vocabulary is encouraged. Some lists of standard vocabularies already exist (such as RFC1766 for language tags), others are being developed.
### The Dublin Core elements

The current DC metadata set comprises 15 elements, in three categories:

#### Relating to the content of the resource

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>the name given to the resource, usually by the publisher or creator</td>
</tr>
<tr>
<td><strong>Subject</strong></td>
<td>simple keywords or terms, possibly from a controlled vocabulary</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>a textual description of the content</td>
</tr>
<tr>
<td><strong>Source</strong></td>
<td>the resource from which the current resource was derived. Not applicable if the present resource is in its original form</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>the language of the content. It is recommended that this element coincides with the language identification tags defined in RFC1766.</td>
</tr>
<tr>
<td><strong>Relation</strong></td>
<td>the identifier of a second resource, and its relationship to the current resource, for example to allow links between related resources</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>the temporal or spatial characteristics of the content (e.g. 18th century, or Uzbekistan)</td>
</tr>
</tbody>
</table>

#### Relating to the intellectual property aspects of the resource

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Creator</strong></td>
<td>the person or organisation primarily responsible for the intellectual content</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>defined as the body making the resource available in its present form</td>
</tr>
<tr>
<td><strong>Contributor</strong></td>
<td>not the Creator, but a secondary contributor who has made significant intellectual contributions</td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>a simple rights statement about the resource or an identifier linking to rights information</td>
</tr>
</tbody>
</table>

#### Relating to the instantiation (broadly, the particular instance) of the resource

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td>meaning a date associated with the resource’s creation or availability</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>the category of the resource (a Web page, an article, etc)</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>the data format of the resource, which is typically a MIME type, such as text/html. It is recommended that this element is taken from a standard list, since it is a key element in defining what tools are needed to operate on the resource</td>
</tr>
<tr>
<td><strong>Identifier</strong></td>
<td>which uniquely identifies the resource, and which could be a URL, URN, DOI, ISBN, or whatever</td>
</tr>
</tbody>
</table>
Inevitably, what starts out as very simple in concept becomes at least a little bit more complex when the practical implications begin to be thought out. DC was extended following the Workshop held in Canberra, Australia, via the use of qualifiers - sometimes known as the Canberra qualifiers.

**The Canberra qualifiers**

Controlled vocabulary lists, for example, are supported by the use of a SCHEME qualifier, for example:

\[
\text{Language SCHEME=RFC1766}
\]

The broad element definitions can be more precisely defined by the use of a TYPE qualifier, for example:

\[
\text{Relation TYPE=IsPartOf}
\]

Multilingual variances are also accommodated by the LANGUAGE qualifier:

\[
\text{Title LANGUAGE=fr}
\]

Although at the third DC Workshop there was further work on using DC for image description, this is essentially where the definition of Dublin Core stands today. It certainly looks promising as the solution to basic metadata needs - remember that RDF provides only the framework, not the metadata itself - but is it just standing, or is it moving forward?

The European Commission’s Libraries activity [org9] has hosted two workshops on metadata. The first, in December 1997 [m11], concentrated specifically on Dublin Core and, while welcoming its contribution to the development of metadata standards, found some shortcomings:

- the lack of formal responsibility for maintaining Dublin Core
- its currently unstable technical state
- the lack of guidelines for its use.

It was felt that the take-up of Dublin Core was slow, with a lack of critical mass. This latter point was seen a classic ‘chicken-and-egg’ scenario; publishers and other providers are not inclined to provide Dublin Core metadata if the ‘harvesters’ - search engines, for instance - do not use it, and the harvesters are not inclined to collect it if there is insufficient metadata available. At the second workshop in June 1998 [m12], the lack of progress on this was noted, and a number of actions were identified as necessary to promote Dublin Core:
• improvements in version control and maintenance

• pilot projects to develop experience

• the establishment of a European group of implementers

• liaison with other groups concerned with metadata, to help ensure applicability and interoperability.

Certainly, having no controlling body has not helped matters. Development of Dublin Core to date has taken place by means of a regular series of workshops, with e-mail discussion in the interim periods.

However, to be fair, the Dublin Core community itself recognises the need for a more formalised body to take some form of responsibility for standardisation, development and maintenance; an organisation is needed to own the Dublin Core. Established bodies such as IETF and ISO have been suggested, although there is no consensus yet.

Nevertheless, a more formal structure, with Policy and Technical Advisory Committees, has now been established; there is some progress in developing RFCs for describing the scheme and its means of implementation [id4]; and there is a long term aim to submit DC to ISO as an international standard.

In the meantime, there have been some DC projects, such as the Nordic Metadata Project [m13], which has also done some interesting work in the incorporation, but not the resolution, of URNs. And there have been some implementations of Dublin Core metadata embedded into HTML - two good examples can be seen by viewing the HTML source of the UKOLN Metadata home page [m14], or of IFLA's very comprehensive list of Metadata Resources [m15].

Nevertheless, <META> tags were never designed to manage complex metadata sets, so this is not a realistic approach for the longer term. RDF offers a better solution; the Dublin Core camp are involved in the RDF developments, and will probably be one of the first to adopt RDF as its preferred syntax.

Remember that DC does not aim to be an all-embracing metadata standard. Its objective is to provide a core set of elements which could be used by (just about) anyone for providing descriptive metadata. Its prominence in the metadata arena has led to a misperception that it offers much more than that, something that the DC camp have never claimed, and it is only the absence of other candidates that has given rise to its front-line position.

With the now close co-operation between the DC and RDF camps, it is possible - if not probable - that DC will find a level of application across the Web. Not by every content provider, but more likely from certain sectors - certain publishers, digital libraries, and so on. And other applications of DC are expected to evolve, for example archival management and its use for the description of non-Web resources and objects, even museum artefacts.
4.3.2 The Warwick Framework

Where DC will not find application is in the areas where more commercially oriented metadata is required, for instance for e-commerce. Applications of DOI, with their focus on trade, very much need this type of metadata. However, although it is not designed to meet these specific needs, DC can sit comfortably alongside such metadata, thanks to the Warwick Framework principles and their potential implementation in RDF.

The recognition in the DC community of the need to encompass more applications than resource discovery was addressed at the Warwick Workshop in April 1996, which formalised what is known - following their particular geographical nomenclature - as the Warwick Framework, described as a container architecture for sets of metadata.

Essentially the architecture embodies two components, the container and within the container, one or more metadata packages. The container is, in simple terms, a means of holding and transporting these metadata packages; they can be transient - constructed only to pass packages between systems - or persistent - stored on servers and accessed via an identifier. The package itself could be:

- a metadata set, such as Dublin Core, or one oriented to another application such as e-commerce
- an indirect reference to an external object, using a URI
- yet another, nested, container

The purpose of the Warwick Framework was to develop the DC principles to provide a more solid and useable formulation. Note that the framework can accommodate more than one variety of metadata; it is not restricted to Dublin Core. This was planned to allow different metadata schemes, offering different facilities, to be incorporated with DC in the metadata container.

Moreover, there is no constraint within the Warwick Framework of how containers are communicated, and it has been experimentally implemented using both MIME and SGML. Containers may also exist within another object which comprises both the data and its metadata; for example, containers might be incorporated into a Web page using the <META> tag.

The perceived benefits of the Warwick Framework are that:

- designers of metadata sets could focus on their specific requirements, without needing to incorporate a generalised set of elements
- the syntax of the sets could vary, depending on requirements
- management of and responsibility for the various sets could be undertaken independently
- interoperability is facilitated by allowing tools to access and manipulate packages individually
• access to the different sets related to the same object can be separately controlled
• future sets can be accommodated; existing sets or tools need not be changed

In view of this functionality, the Warwick Framework concepts have heavily influenced the development of RDF.

A comprehensive description of the Warwick Framework is provided in “The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata” [m16], but there are other, more concise, descriptions listed in the references.

4.3.3 The DOI and metadata

The DOI people equally accept the need for metadata to support their initiative, and this includes a form of descriptive metadata for discovering DOIs. The IDF recognises that, when a DOI is assigned, it is necessary to collect some metadata about the Creation denoted by the DOI; at the least, to have a minimum set of metadata which allows the identification of a DOI from the data available. Moreover, it sees the definition of that metadata set as an urgent task.

There may currently be this focus on the use of metadata as a search mechanism to locate a DOI, but the IDF also recognises that [i1]:

‘once such metadata has been established, it could serve many useful purposes, not least in linkage between related Creations, and that therefore there are many opportunities for devising added-value services and working with existing abstracting and information services.’

The current plans envisage a ‘bare-bones’ metadata set for DOI, e.g. assigner, Creation type and information identifier, enumeration (e.g. journal, volume, issue, etc), and format. But remember that the concept of the DOI is not only as a unique identifier of a digital object, but also as a gateway to a variety of services related to the object. To be useful in practice, it needs a consistent structure:

‘While it is true that the DOI implemented as a redirected link may be crafted to do just about anything we ask, solutions that will be acceptable to the publishing industry - especially those involving any sort of end user participation - require consistent, dependable responses.’ [m17]

What this implies is that, for the range of services that might be available for an object to be realisable, there must be a way to access the various sources of information about that object which are necessary for those services to function - in other words, the corresponding metadata sets. So DOI also has to address issues of interoperability, machine-understandable structure and metadata exchange formats. This has however been recognised within the DOI community, and work is continuing.

The focus in this community appears to remain on its minimal metadata set for the present, although working groups have been set up to study both descriptive and rights metadata, and it is in contact with the RDF initiative. Nevertheless, the European Commission’s DOI Workshop [i14] concluded that further work is still required to define the metadata elements underlying the DOI.
4.4 Conclusions

Although there is still a long way to go, there does seem to be a future for metadata on the Web. Dan Connolly of the W3C comments: ‘Metadata on the web has been coming very slowly since the dawn of the web. I think that RDF is just about to turn the corner and introduce what [Tim Berners-Lee] has been calling Web phase II.’ [i16]

And we appear to have two main candidates, complementing each other - RDF for the metadata framework, Dublin Core for the metadata itself.

4.4.1 Consensus and collaboration

This is all well and good, but it will only work if there is real collaboration, so how much are the two camps talking to each other?

The Dublin Core activists consistently affirm their commitment to RDF: ‘Representatives of the Dublin Core effort are actively involved in the development of [the RDF] architecture, bringing the digital library perspective to bear on this important component of the Web infrastructure.’ [m18]

Connolly agrees, pointing out that the DC community is now engaged at a nuts-and-bolts level. Indeed, at the fifth DC workshop, in Helsinki in October 1997, representatives of the W3C Metadata Project presented the RDF specifications and showed how it can accommodate DC, and there was much mutual praise from both sides about the cross-fertilisation of ideas and active participation.

This commitment is clearly important if metadata is to succeed, but not all agree that it is sufficient at present. The European Commission's second Metadata Workshop in June 1998 [m12], attended by many luminaries from the metadata world, concluded that ‘establishing widely accepted agreements is essential for the success of metadata’ and moreover that this consensus must be achieved across the interested parties - libraries, business, etc. Further, the formal and informal bodies involved ‘need to find effective ways of co-operation to ensure maximum acceptance of agreements and to avoid overlapping activities.’

So apparently, despite the collaboration claimed by all parties, some feel that more is needed - more collaboration, more co-operation, more liaison.

4.4.2 Making it work

Next question - presuming that we have or will have the necessary collaboration, and ultimately the necessary infrastructure, what else is needed to make metadata on the Web work in practice? The simple answer - we will need metadata on the Web. It is one thing to have the right framework and the right scheme in place, it is another thing to have people use it.

Connolly makes the point that RDF is very generalised and decentralised, but: ‘The question is whether the various metadata application designers and developers will accept the cost of this generalized approach, or whether they will do something specific to their application.’ [i16]
The view of the W3C is that RDF will provide ‘a uniform query capability for resource discovery’ with ‘better precision ... than full text search’ [m2]. Bear in mind that the URC camp were making similar statements: ‘Another benefit from using URCs is that WWW indexers (such as Lycos and Altavista) can then begin to index the metadata and not the whole document itself. This will improve the quality of the WWW indexed data and improve the retrieval of information via resource discovery tools.’ [i9]

This suggests that if we have metadata, resource discovery will be that much easier, but this of course depends on the metadata being provided in the first place. The Warwick Framework paper points out: ‘The simplicity of the Dublin Core was motivated by the desire to make it useable by the general class of non-professional authors and publishers so common on the World Wide Web. Yet, there is some empirical evidence that this class of providers will not even provide the simplest of descriptive information’ [m16], meaning that, at least at present, the people who are providing much of the content on the Web are maybe not interested in providing the supporting metadata. Connolly makes a similar point: ‘The social and economic issues are the hard part: getting people to actually put metadata in their documents. Getting them to trust it. Getting institutions to produce and consume it.’ [i16] So authors need to add it, and engines and applications need to make use of it – another chicken-and-egg scenario perhaps.

4.4.3 Dealing with diversity

The strength of RDF as an infrastructural metadata framework is that it allows for a diversity of different schemes, and all metadata activists appear to recognise that RDF offers real potential for the widespread deployment of metadata across the Web. As we saw above, whether it will be used is a different question. But even if it is, how will metadata be accessed and retrieved if a theoretically unlimited variety of different schemes can be accommodated?

We know that it would be impossible to achieve an all-embracing metadata set - ‘author’ or a broad equivalent such as ‘creator’ might almost always be found, but different fields of application have their own specific needs. A photographic collection might need elements such as ‘shutter speed’ and ‘aperture’ that would be redundant - and meaningless - for an collection of, say, photographic magazine articles. The Warwick Framework concept, embodied within RDF, is ‘horses for courses’ - different metadata sets, each best suited to its particular application needs, can be collected together to describe a given resource. The issues are recognised; the Warwick Framework people note:

‘Consider, for example, a common consumer of metadata - the “spider” or “crawler” that tries to gather descriptive metadata for networked objects and then compiles it into a searchable index. Designing this agent is difficult if descriptive cataloging metadata is contained, without concern for consistency, in a number of metadata sets for each object. What are the rules for assembling a usable index from such arbitrarily mixed metadata? What are the semantic transformations that can be made across the multiple metadata sets?’ [m16]
A common metadata set can, of course, be used for basic descriptive elements, and it is here that Dublin Core sees its role:

‘But most resources share a core set of attributes that are similar from one discipline to the next, but have different names simply because they have evolved independently and at different times. Promoting a commonly understood set of core descriptors will improve the prospects for cross-disciplinary search by unifying related attributes. ... The Dublin Core is intended to serve as this core element set.’ [m19]

Even so, the question of semantic interoperability still remains. A search engine, or any other automated process trying to operate on metadata, will need to recognise that, for instance, Author and Creator may be functionally identical. Let us say that on the Web, a photographer has collections of his photographs available, where the metadata scheme records him as the Creator, and he may also have articles which he has written for photographic magazines, where he is recorded as Author. If a search engine wishes to retrieve information on all of his works, it will need to understand the meaning and use of these elements, to recognise the correspondence between Author and Creator.

The solution might be to provide means of mapping between different metadata schemes, in order to identify semantic equivalence, but this will need some work. UKOLN maintains details of mappings between a number of established schemes [m20], which show that this is not always a straightforward task.

Nevertheless projects are under way to research these issues, such as that of the British Arts and Humanities Data Service. This project lists among its objectives the ability for users to ‘query a range of catalogue databases according to elements or fields such as creator, title, subject, which those databases commonly support, even if those databases use different names internally to refer to those elements or fields.’ [m21]

In the TURNIP project [s2], one claimed outcome was the successful use of different metadata schemes in the resolution services. An Australian service adopted the Dublin Core metadata set, while another, in Italy, used the IAFA template scheme. Both resolvers were independent of the URC scheme used, and were able to freely interoperate.

And we must not forget that the Web is multilingual; we need to equate, say, Author and Auteur. Again, this is recognised and some progress has been realised - the Dublin Core specification now includes language qualifiers, and the LANG qualifier is provided in HTML as direct support for DC.

The W3C also sees interoperability as a major objective. Tim Berners-Lee, W3C Director, in his keynote address to the September 1998 eBusiness trade show in Boston, predicted that eventually there will be schemas providing definitions for particular words and a means of converting various definitions so that they can be consistently understood by computers. This, note, was offered in the context of electronic commerce on the Web, and such commercial motivations may provide the impetus to ensure that the interests of the Web community are met.
4.4.4 In conclusion

The general view, therefore, is that metadata has a future, but that there is still more work to be done in:

- establishing well-defined and well-controlled standards
- developing its capability to offer the security and rights mechanisms that are important to the development of Web commerce
- establishing the tools and services to support it
- promoting the widespread provision of metadata for Web resources.

Given the vast amount of resources already on the Web, without metadata, this latter task will perhaps be the most difficult.
5. IMPLEMENTING THE SCHEMES

These are early days for all of the leading schemes that we have talked about here, but we can presume that at some point in the future they will be ready for deployment. And at that point, software will need to be ready to use them, and the supporting services will need to be ready on the Web to make them work. Since no scheme is yet close to that point, concrete work on software and services is still further away, so here we will only touch on the issues involved.

5.1 Software

For identifiers

The installed base of browsers is now enormous, therefore support for identification schemes in these software tools will inevitably be a key factor in the successful deployment of any scheme.

There is no immediate prospect on the horizon of software developments supporting any identification scheme, but this is to be expected given the early stages of development of the schemes themselves.

For the URN, the essential element is that browsers will inherently support it, which means at root that they will be able to parse a URN and connect to the required resolution service. Beyond that, there are a number of ways in which URNs could be handled by the browser, some of which depend themselves on what is done by the resolution service. RFC2276 talks of a "resolver" to indicate a service which translates URNs to URLs or URCs, but notes that some resolvers might directly link to resources being referenced [rfc2276].

The generally accepted scenario is that, given a URN, the service will return a list of URLs, from which the user can select, or will return the metadata for the resource (including, no doubt, that list of URLs) for the user to inspect before deciding what to do. To help in this respect, the browser could allow the user to specify his preferences on how this retrieved information should be handled. It might be configured to automatically select a URL for an HTML, or a PDF, version of the resource, or the URL for the location calculated to be the closest geographically and therefore - in principle - the fastest. Or the browser could work through the list of URLs until it retrieves the resource. In the future, the metadata that is returned may indicate, for example, that a payment is necessary to access the resource. So there are a number of different possibilities, and the organisations involved will have to deliver workable proposals. Otherwise, as often happens, the major software players may decide to go their own separate ways.

The IDF also recognises the need for browser support for the DOI, noting that the way forward will be 'making available plug-in browser support tools ... which provide improved capabilities for deployment of services beyond the http protocol and encouraging the development of further browser tools to give intelligent client capability for use with level 2 DOIs.' [i1]
For metadata

Metadata, too, will need its supporting software. It seems that Dublin Core toolkits are in development, notably by the academic and library community, and that these will probably become public domain (from the Nordic Metadata Project [m13], for example). It has been suggested that what is really needed is an automatic DC profile editor which is easy to operate and restricted to a subset of widely agreed elements, but this may not be not easy to achieve.

However, perhaps the real requirement is for an integrated approach to Web publishing; the generation of metadata should be an integral facility of the tools used to generate Web content, which means that the major software developers will have to commit themselves. Recourse to a separate tool to provide metadata will often mean that the separate tool will never be used. Moreover, in terms of supporting software for DC, maybe we have another chicken-and-egg situation - there is a need for workable tools to help create DC metadata, but there might not be market pressure to deliver these until DC is seen to be gaining ground.

Metadata in general, nevertheless, enjoys an advantage in that RDF, and by extension Dublin Core, will depend upon Extensible Markup Language (XML) [s6]. XML, like HTML, is a subset of Standard Generalised Markup Language (SGML). However, while HTML is a very limited subset concerned primarily with Web presentation, XML provides the means to present and process structured data, in other words machine-useable information, on the Web.

Here the future seems clearer. For some time now, XML has been the flavour of the month, every month, for a number of commercial developers of Web software, and looks to continue to be so. The level of commitment to XML has been impressive, and there are already tools available despite the fact that the W3C has not yet completed the full complement of specifications. The potential ability of XML to make the Web machine-useable, and thus ready for viable electronic commerce, is maybe helping to fuel this movement.

In the browser world, Microsoft already provide basic support for XML in Internet Explorer 4, and is promising more in IE5; Netscape have promised similar support in Communicator 5.

Even so, XML can only offer effective support for metadata when the various schemes which depend upon it, directly or indirectly, are themselves finalised and implementable. It was said that ‘XML solves how we are going to talk to each other. We still need to agree on what we are going to talk about.’ [s1]

5.2 Services

Resolution services are fundamental to the operation of the World Wide Web. The Domain Name System resolves the host name in a URL into the IP address required to locate, and route to, that host. Nevertheless, while this system is well established and, by and large, functions well, there has been a long debate, now perhaps drawing to a close, over its future structure and control. Maybe there is a lesson here - an infrastructural Web service needs careful planning, taking into
account its global scope, the implications of growth and the need for formal control. Even then, we cannot confidently predict the future, and there may come a time when a major, and previously unforeseen, overhaul is required.

The definition of a new resolution service should correspondingly be as considered and comprehensive as possible, and certainly the progress on such services for the new schemes discussed here has been slow and careful. Only two, the PURL and the DOI, have an operational resolution service in place.

The PURL service at OCLC is a relatively straightforward implementation, since it depends upon established mechanisms such as the URL and HTTP redirection, but as its proponents point out, it is not intended as anything more than an interim system until the URN is under way.

DOI is different, in that it has a long-term future planned for it, and so its resolution service needs to last. Although it does not necessarily require the Handle system that it currently employs, it is generally accepted that the use of this system will continue. Operational questions remain, however, for the future - should, for instance, a global DOI service be centralised or distributed?

URN is a different matter. Since the URN is still far from the delivery of an operational implementation, there remains much work to be done on its resolution service. The IETF discussion on the topic can be found in ‘Architectural Principles of Uniform Resource Name Resolution’ [rfc2276].

Nevertheless there have been some experimental implementations of URN resolution services, such as the TURNIP project [s2]. The *mu* project at the Computer Science Laboratory at the Massachusetts Institute of Technology (MIT) is dealing with some of the issues [s4] and, in particular, there is a good overview on the requirements for such services [s5]. Another, although older discussion, can be found on the Australian BURNS project [s6].

Progress will inevitably be slow, because the URN, like the Domain Name System, will need to be part of the fundamental infrastructure. Accordingly many bodies will need to be involved in the definition of the resolution service and its long term future support and control, and some will need to make a concrete investment in time and money - governments, institutions, publishers, libraries, software vendors, and so on. The conclusions of the Nordic Metadata Project [m13] suggest that ‘establishment of the URN resolution infrastructure will quite likely require a few years.’

The consensus view remains that the necessary technology is easy to define and implement; the hurdles to clear are commercial, organisational and social. Dan Connolly of the W3C [i16] accepts this, but still sounds a note of warning: ‘On second thought deployment of new, scalable, ubiquitous resolution services is a technical challenge that I'd call a “major problem area.” Beware of anything that relies on that.’
6. LOOKING AHEAD

So to summarise, we have had a variety of schemes for identifying and describing Web resources, out of which only a few remain in the limelight. In all cases, the road to implementation is slow and careful. First comes the process of defining the scheme, then agreeing that definition. Once that is done, the necessary organisational infrastructure must be set up, and the technical foundations laid - the supporting services and software. All of this requires a considerable investment in time and money before implementation and deployment can take place. And while this is all happening, the scheme may well be concurrently evolving further, leading to a continuous cycle of development and implementation.

But even then, however powerful the forces that are behind the scheme, nobody can reliably predict its take-up - whether the world will actually use it. If the market or the user community are not attracted to it, then it will not happen, or at best it finds only a niche market. The history of IT offers many examples of Good Ideas which never found favour, even when promoted heavily by the major players. To try to offer a perspective, we can ask the same questions about any of these schemes - Will it happen? When? And, do we really need it?

For identifiers

Here the newer schemes are URN and DOI. While URI is an interesting development, it is not strictly a scheme in itself, of course; rather, it is a mechanism for encompassing all identification schemes within a unifying syntax.

And in the context of the Web, we could also include the trusty URL as an identification scheme. Moreover, although it has known limitations at present, it is already happening, and on the whole does its job.

Equally, although these are early days, the DOI is also already happening. Not yet deployed in a fully operational sense, perhaps, but there are supporting services in place, software tools (browser plug-ins) are available, and the system can be seen working at the IDF Web site [org1].

Full operational deployment of the DOI will take time, but progress is being made. Do we really need it? Probably yes. It is not something that aims to become a universal identifier, and so not something that the whole world will use, but it is designed to offer specific functional benefits for particular application areas, largely within the publishing and libraries domain, and as such should have a real future.

The URN can claim to be a new universal identifier, but here progress is slow and we are a long way from implementation. It is even hard to say if it will ever happen - in some quarters, people are growing impatient, meaning that there could be pressure for a more immediately realisable solution (the PURL, perhaps?), and the dissent about its real value may have a influence in the long term. In concept, the URN is sound - it offers a robust mechanism for identifying
resources and for handling the inherent complexity. But what some, notably the W3C, are saying is that this is all well and good, but that the URN does not add sufficient value to make it worth the time and money required to bring it to implementation. With these forces against it, it may be that it will at least have a difficult gestation.

For metadata

The picture seems clearer for metadata. There is more consensus and collaboration, and the two most advanced schemes, RDF and Dublin Core, are complementary - the former providing the framework and the latter the resource description within that framework. RDF seems pretty certain to happen; there is sufficient weight behind it to move it forward, and already some operational Web sites are employing it. Dublin Core also looks to have a promising future, since it provides the basic metadata set to support most needs for resource description.

There remains much work to be done to provide metadata facilities for more specific applications, most significantly in e-commerce and rights management, but commercial motivations may accelerate that process, and RDF will be able to accommodate any new scheme and provide the necessary security.

Large-scale deployment of metadata, as with all the other schemes that we have covered here, will take time, but it is fairly safe to say that it will happen, because we do need metadata on the Web. If the Web is to evolve from being a largely amorphous mass of information to a coherent and organised information repository, then metadata probably offers its best chance to do so.
REFERENCES AND LINKS TO OTHER RESOURCES

Relating to identification schemes

[i1]
The DOI Initiative; current position and view forward
International DOI Foundation - August 1998
Available in PDF format from http://www.doi.org

[i2]
A Beginner's Guide to URLs
http://www.ncsa.uiuc.edu/demoweb/url-primer.html

[i3]
Identification – Related resources
BIBLINK, hosted by UKOLN
http://hosted.ukoln.ac.uk/biblink/wp2/links.html

[i4]
W3C and Digital Libraries
James Miller, W3C - Dlib Magazine, November 1996
http://www.dlib.org/dlib/november96/11miller.html

[i5]
The Myth of Names and Addresses
Tim Berners-Lee, W3C - December 1996
http://www.w3.org/DesignIssues/NameMyth.html

[i6]
Unique Identifiers: a brief introduction
Brian Green and Mark Bide - September 1996, revised March 1997
www.bic.org.uk/bic/uniquid.html

[i7]
Naming and Addressing: URIs
http://www.w3.org/Addressing/Addressing.html

[i8]
Uniform Resource Identifiers (URIs)
Ron Daniel, Advanced Computing Laboratory, Los Alamos National Laboratory
http://www.acl.lanl.gov/URI/uri.html

[i9]
The User Interface of URNs and URCs
Renato Iannella, DSTC Pty - August 1998
[i10]  
Introduction to Persistent Uniform Resource Locators  
OCLC - 1996  
http://purl.oclc.org/OCLC/PURL/INET96

[i11]  
PURLs: Persistent Uniform Resource Locators  
OCLC - updated August 1998  
http://purl.oclc.org/OCLC/PURL/SUMMARY

[i12]  
Introduction to the Digital Object Identifier  
International DOI Foundation - updated April 1998  
http://www.doi.org/introduction.html

[i13]  
The Handle System  
http://www.handle.net

[i14]  
European Commission DGXIII/E4 - May 1998  
http://www2.echo.lu/libraries/en/doiwks.html

[i15]  
IETF WEBDAV Working Group  
http://www.ics.uci.edu/pub/ietf/webdav

[i16]  
Dan Connolly, Architecture Domain head, W3C  
Communications to the author, July 1998

[i17]  
Advanced Network Architecture, Current Projects  
MIT Laboratory for Computer Science - 1998  

[i18]  
The Real Name system  
http://www.centraal.com

Others

W3C - The Architecture Domain  
http://www.w3.org/Architecture/

PADI - Preserving Access to Digital Information  
National Library of Australia  
Identifiers and Their Role in Networked Information Applications
Clifford Lynch
http://www.arl.org/newsltr/194/identifier.html

Digital Object Identifiers (DOIs) and Clifford Lynch’s five questions on identifiers
William Y. Arms - October 1997
http://www.arl.org/newsltr/194/arms.html

In Search of the Unicorn: the DOI from a user perspective
Mark Bide - February 1998
http://www.bic.org.uk/bic/unicorn2.pdf

Information Identifiers
Norman Paskin
http://www.elsevier.nl/homepage/about/infoident/

Unique Identifiers in a Digital World
BIC workshop report, Andy Powell - April 1997
http://www.ariadne.ac.uk/issue8/unique-identifiers

Relating to metadata schemes

[m1] W3C - Platform for Internet Content Selection (PICS)
http://www.w3.org/PICS/

[m2] W3C - Frequently Asked Questions about RDF
http://www.w3.org/RDF/FAQ

[m3] Channel Definition Format (CDF)
Microsoft - April 1998
http://www.microsoft.com/standards/cdf/default.asp

[m4] An MCF Tutorial
June 1997
http://www.textuality.com/mcf/MCF-tutorial.html

[m5] Web Collections using XML
Microsoft - March 1997
http://www.w3.org/TR/NOTE-XMLsubmit.html

[m6] XML-Data
Microsoft - December 1997
http://www.microsoft.com/standards/xml/default.asp
[m7]
W3C - Resource Description Framework (RDF)
http://www.w3.org/RDF/

[m8]
W3C - Metadata Activity
http://www.w3.org/Metadata/Activity.html

[m9]
Metadata for the Web - RDF and the Dublin Core
Andy Powell, UKOLN - 1998
http://www.ukoln.ac.uk/metadata/presentations/ukolug98/paper/intro.html

[m10]
Description of Dublin Core Elements
http://purl.org/metadata/dublin_core_elements

[m11]
Metadata Workshop, Luxembourg - 1-2 December 1997
European Commission DGXIII/E4 - March 1998
http://www2.echo.lu/libraries/en/metadata.html

[m12]
Report of the Metadata Workshop held in Luxembourg, 26 June 1998
European Commission DGXIII/E4 - August 1998
http://www2.echo.lu/libraries/en/metadata2.html

[m13]
Nordic Metadata Project
June 1998
http://renki.lib.helsinki.fi/meta/

[m14]
Metadata
UKOLN - March 1998
http://www.ukoln.ac.uk/metadata/

[m15]
Metadata Resources
http://www.nlc-bnc.ca/ifla/II/metadata.htm
An enormous list, maintained by IFLA, of metadata resources that are available on the Web

[m16]
The Warwick Framework:
A Container Architecture for Aggregating Sets of Metadata
[m17]
Requirements for DOI-Based Applications and Services
John Erickson - January 1998
http://www.ybp.com/yrm/papers/NISODOIServices.htm

[m18]
Dublin Core Metadata
OCLC - November 1997
http://purl.org/dc

[m19]
The Evolving Metadata Architecture for the World Wide Web: Bringing Together the Semantics, Structure and Syntax of Resource Description
Stuart Weibel, OCLC (International Symposium on Research, Development and Practice in Digital Libraries - Ibaraki, Japan) - November 1997
http://www.dl.ulis.ac.jp/ISDL97/proceedings/weibe.html

[m20]
Metadata - Mapping between metadata formats
UKOLN - last update August 1998
http://www.ukoln.ac.uk/metadata/interoperability

[m21]
Discovering Online Resources, Metadata and Middleware: A Systems Architecture for Cross-Domain Discovery
http://ahds.ac.uk/public/metadata/disc_06.html

which is part of

Discovering Online Resources Across the Humanities - A Practical Implementation of the Dublin Core
Arts and Humanities Data Service/UKOLN - November 1997
http://ahds.ac.uk/metadata/discovery.html

Others

The Warwick Framework: A Container Architecture for Diverse Sets of Metadata
An abbreviated version of [m16]
http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

The Warwick Metadata Workshop
Lorean Dempsey, Stuart Weibel - D-Lib July 1996
http://www.dlib.org/dlib/july96/07weibel.html

The Warwick Framework - A quick overview
UKOLN
http://homes.ukoln.ac.uk/~lisap/BIBLINK/wf.html
What is... RDF?
Rachel Heery, Ariadne - March/July 1998
http://www.ariadne.ac.uk/issue14/what-is/
A useful introduction to RDF

URC working web - Uniform Resource Characteristics (URCs)
Ron Daniel - November 1995
http://www.acl.lanl.gov/URC/

Netscape Metadata pages

Metadata for the masses
Paul Miller - Ariadne, September 1996
http://www.ariadne.ac.uk/issue5/metadata-masses/

Relating to supporting software and services

[s1]
Matthew Fuchs, CNGroup, quoted in
XML and the New Industry Order
Wired - Dec 1997

[s2]
TURNIP (The URN Interoperability Project)
Renato Ianella, D-Lib Magazine - March 1996
http://www.dlib.org/dlib/march96/briefings/03turnip.html

[s3]
u:\: Multicast URN resolution
MIT Laboratory for Computer Science - July 1998
http://phobos.lcs.mit.edu/

[s4]
Requirements for URN Resolution Systems
Laboratory for Computer Science, MIT. February 1998
http://phobos.lcs.mit.edu/requirements.html

[s5]
BURNS: Basic URN Service Resolution for the Internet
Renato Ianella, Hoylen Sue, Danny Leong- August 1996

[s6]
W3C - Extensible Markup Language (XML)
http://www.w3.org/XML/
IETF REQUESTS FOR COMMENT

[rfc1630]
Universal Resource Identifiers in WWW (June 1994)

[rfc1737]
Functional Requirements for Uniform Resource Names (December 1994)

[rfc2141]
URN Syntax (May 1997)

[rfc2276]
Architectural Principles of Uniform Resource Name Resolution (January 1998)

[rfc2396]
Uniform Resource Identifiers (URI): Generic Syntax (March 1998)

RFCs in general

where nnnn is the RFC number

IETF INTERNET DRAFTS

[id1]
Internet Draft: URN Namespace Definition Mechanisms

[id2]
Persistent Document Identifiers
James Mallery – November 1997
http://info.internet.isi.edu:80/in-drafts/files/draft-mallery-urn-pdi-00.txt

[id3]
URC Scenarios and Requirements
Ron Daniel Jr., Michael Mealling - June 1995
http://www.acl.lanl.gov/URI/Scenarios/

[id4]
Dublin Core Metadata for Resource Discovery
http://info.internet.isi.edu:80/in-drafts/files/draft-kunze-dc-03.txt
Internet Drafts in general

http://info.internet.isi.edu/0/in-drafts/files
This provides a comprehensive (and very long) list of Internet Drafts

http://www.ietf.org/ID.html
This provides keyword search and other facilities for Internet Drafts
ORGANISATIONS

[org1]
The International DOI Foundation (IDF)
http://www.doi.org

[org2]
Internet Engineering Task Force (IETF)
http://www.ietf.org

[org3]
World Wide Web Consortium (W3C)
http://www.w3.org

[org4]
Internet Assigned Number Authority (IANA)
http://www.iana.org

[org5]
Online Computer Library Center (OCLC)
http://www.oclc.org

[org6]
PURL Home Page
http://purl.oclc.org

[org8]
Corporation for National Research Initiatives (CNRI)
http://www.cnri.reston.va.us

[org9]
Electronic Publishing and Libraries Unit, European Commission DGXIII/E4
http://www.echo.lu/ibs

[org10]
UK Office for Library and Information Networking (UKOLN)
http://www.ukoln.ac.uk/

Others

The Ariadne newsletter
http://www.ariadne.ac.uk/

The D-lib program (D-Lib)
http://www.dlib.org/

Association of Research Libraries (ARL)
http://www.arl.org/
Coalition for Networked Information (CNI)
http://www.cni.org/

International Federation of Library Associations and Institutions (IFLA)
http://www.ifla.org/

EDItEUR
http://editeur.org/

Book Industry Communication
http://www.bic.org.uk