UNIVERSITY OF
Southampton

Identification

COMP3220 Web Infrastructure

Dr Nicholas Gibbins – nmg@ecs.soton.ac.uk

Uniform Resource Identifiers

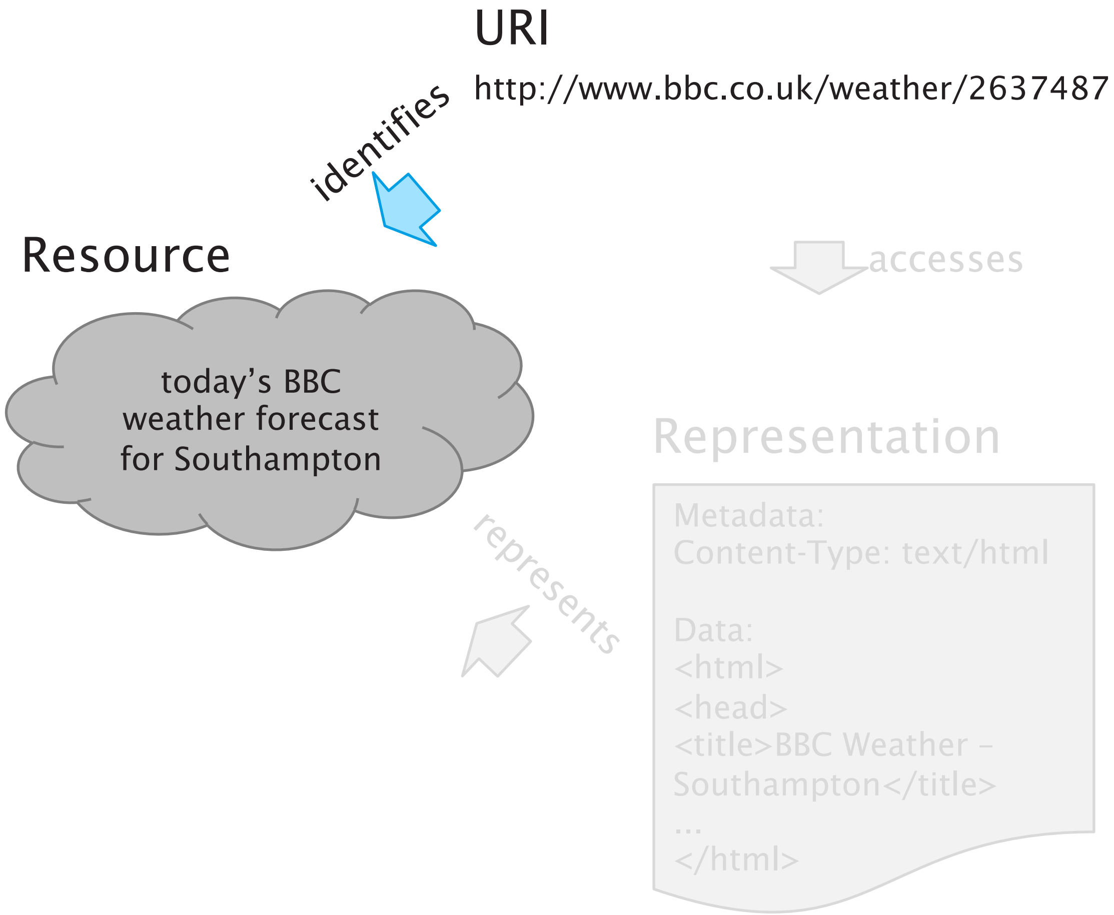
A “compact string of characters for identifying an abstract or physical resource”

Example:

`http://www.ecs.soton.ac.uk/`

General syntax:

<scheme>:*<hierarchical part>*?*<query>*#*<fragment>*



URI

http://www.bbc.co.uk/weather/2637487

identifies

accesses

Resource

today’s BBC
weather forecast
for Southampton

represents

Representation

Metadata:
Content-Type: text/html

Data:
<html>
<head>
<title>BBC Weather –
Southampton</title>
...
</html>

Berners-Lee, T. et al (2005) *Uniform Resource Identifier (URI): Generic Syntax*. RFC3986. Available at https://tools.ietf.org/html/rfc3986

URI Schemes and Examples

- http://www.example.org/aboutus#staff
- https://www.example.org/login
- mailto:joe@example.org
- ftp://example.org/aDirectory/aFile
- news:comp.infosystems.www
- tel:+1-816-555-1212
- ldap://ldap.example.org/c=GB?objectClass?one
- urn:oasis:names:tc:entity:xmlns:xml:catalog

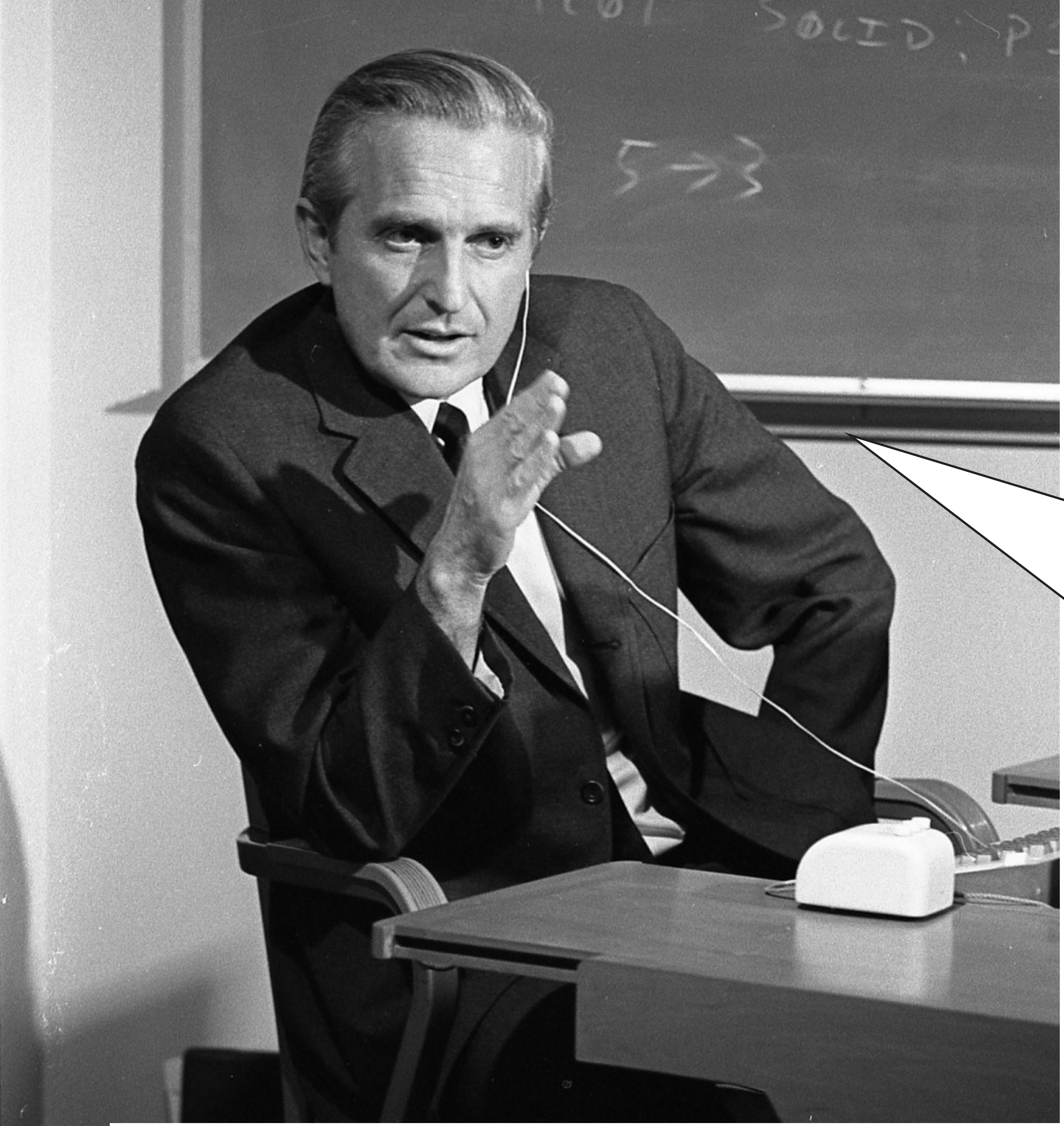
Identification Principles

1. Identifiers should be global

Global naming leads to global network effects.

We want to avoid creating walled gardens.

Every object should be addressable



In principle, every object that someone might validly want/need to cite should have an unambiguous address (capable of being portrayed in a manner as to be human readable and interpretable). (e.g., not acceptable to be unable to link to an object within a ‘frame’ or ‘card.’)

Englebart, D.C. (1990) Knowledge-Domain Interoperability and an Open Hyperdocument System. Proceedings of the Conference on Computer-Supported Collaborative Work.

Identification Principles

1. Identifiers should be global
2. Assign distinct identifiers to distinct resources

Using the same URI to directly identify different resources produces a URI collision.

Example: using `http://www.ecs.soton.ac.uk/` to refer to both a university department and a web page about that department

Collision often imposes a cost in communication due to the effort required to resolve ambiguities.

Identification Principles

1. Identifiers should be global
2. Assign distinct identifiers to distinct resources
3. Avoid aliases

A URI owner SHOULD NOT associate arbitrarily different URIs with the same resource.

Example: `http://www.soton.ac.uk/` and `http://www.southampton.ac.uk/` both refer to the same resource – but we can't tell that just by looking at the identifiers (URIs are opaque)

The value of a given resource can be measured by the number and value of the resources that link to it

The Early Web

Early documents refer to document naming:

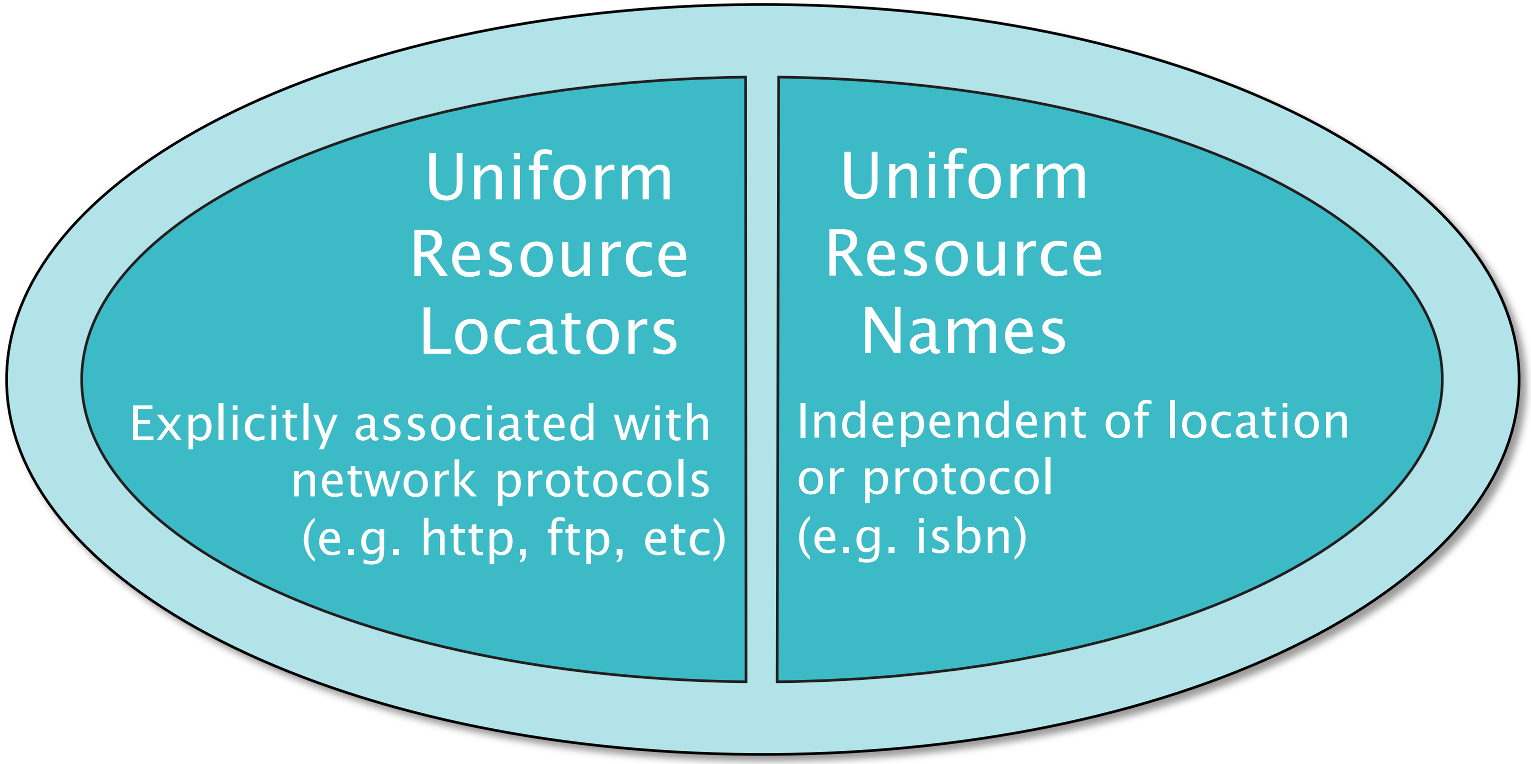
“As many protocols are currently used for information retrieval, the address must be capable of encompassing many protocols, access methods or, indeed, naming schemes”

“A hypertext link to a document ought to be specified using the most logical name as opposed to a physical address. This is (almost) the only way of getting over the problem of documents being physically moved. As the naming scheme becomes more abstract, resolving the name becomes less of a simple look-up and more of a search.”

Berners-Lee, T. (1991) *Document Naming*. Available at: http://www.w3.org/DesignIssues/Naming

The Classical View

Uniform Resource Identifiers



Uniform Resource Locators

Explicitly associated with network protocols (e.g. http, ftp, etc)

Uniform Resource Names

Independent of location or protocol (e.g. isbn)

Name resolution

URL resolution is (usually) well-defined (use the listed protocol)

URNs don’t necessarily have well-defined resolution semantics

- Resolving names depends on context
- What does resolution mean for URIs which do not refer to network resources?

Significant development work in the 1998-2002 timeframe, culminating in the Dynamic Delegation Discovery System

- Uses DNS to store a database of rewrite rules for URIs in NAPTR records

Sollins, K. (1998) *Architectural Principles of Uniform Resource Name Resolution*. RFC2276. Available at: https://tools.ietf.org/html/rfc2276
Daniel, R. and Mealling, M. (1997) *Resolution of Uniform Resource Identifiers using the Domain Name System*. RFC 2168. Available at: https://tools.ietf.org/html/rfc2168
Mealling, M. and Daniel, R. (2000) *The Naming Authority Pointer (NAPTR) DNS Resource Record*. RFC 2915. Available at: https://tools.ietf.org/html/rfc2915
Mealling, M. (2002) *Dynamic Delegation Discovery System, Part One*. RFC3401. Available at https://tools.ietf.org/html/rfc3401

The Modern View

The URL/URN/URI debate has been a long-running issue for the Web

Formal URL/URN distinction is unhelpful, but URL is a useful informal concept

- "a URL is a type of URI that identifies a resource via a representation of its primary access mechanism"
- e.g. a http: URL identifies a resource whose representation can be retrieved using the HTTP protocol

Current practice when creating new URIs is to use http/https schema (regardless of the type of resource) and ensure that it resolves to give "something useful"

- Cheaper than DDDS-like solutions

W3C TAG (2002) *ISSUE-14: What is the range of the HTTP dereference function?* Available at: https://www.w3.org/2001/tag/group/track/issues/14
W3C/IETF Joint URI Planning Interest Group (2001) *URIs, URLs, and URNs: Clarifications and Recommendations 1.0*. W3C Note. Available at: http://www.w3.org/TR/uri-clarification/
Berners-Lee, T. (2002) *What do URIs identify?*. Available at: https://www.w3.org/DesignIssues/HTTP-URI.html

Internationalized Resource Identifiers

URIs as specified in RFC3986 use only US-ASCII characters

IRIs (defined in RFC3987) extend URIs by allowing Unicode characters:

- https://el.wikipedia.org/wiki/Αθήνα
- https://zh.wikipedia.org/wiki/北京市
- https://he.wikipedia.org/wiki/ירושלים
- Even http://💩.to/

Relies on internationalised domain names

Mapping from IRIs to URIs to support older tools (i.e. Punycode)

Duerst, M. and Suignard, M. (2005) *Internationalized Resource Identifiers*. RFC3987. Available at: https://tools.ietf.org/html/rfc3987
Klensin, J. (2010) *Internationalized Domain Names in Applications: Protocol*. RFC5891. Available at: https://tools.ietf.org/html/rfc5891

Cool URIs don't change

What makes a cool URI?
A cool URI is one which does not change.

What sorts of URI change?
URIs don't change: people change them.

Berners-Lee, T. (1998) *Cool URIs don't change*.
Available at: https://www.w3.org/Provider/Style/URI

Cool URIs don't change

Changing a resource’s URI breaks any pages that may have linked to the old URI

- `404 Not Found`
- Keep the old URI and redirect? (increased latency)

Changing the URI of a resource breaks the principle of avoiding aliases

Further Reading

Jacobs, I. and Walsh, N. (2004) *Architecture of the World Wide Web, Volume One*. W3C Recommendation.

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

Berners-Lee, T. et al (2005) *Uniform Resource Identifier (URI): Generic Syntax*. RFC3986

https://tools.ietf.org/html/rfc3986

Next Lecture: Representation