
University of
Southampton

University of Southampton

Identification, Representation & Interaction

COMP3227 Web Architecture & Hypertext Technologies

Dr Heather Packer – hp3@ecs.soton.ac.uk

Uniform Resource Identifier

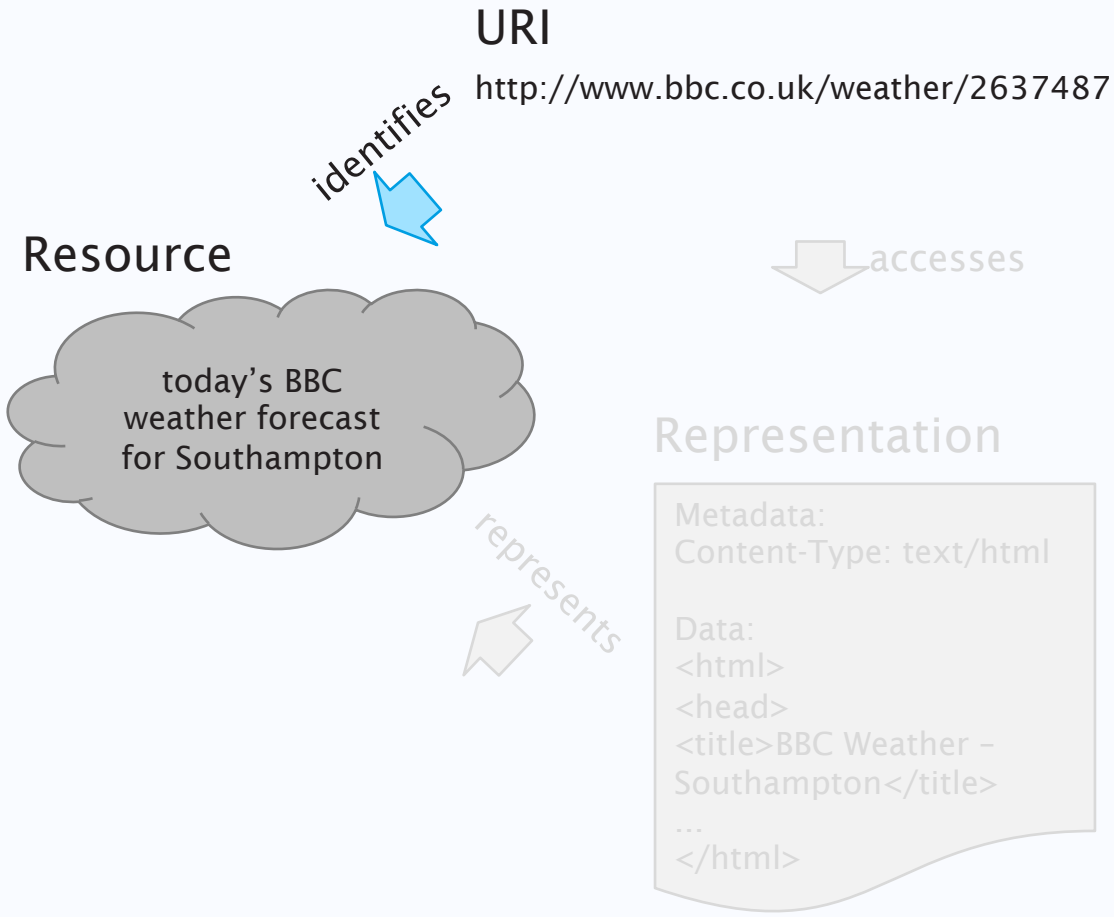
A “compact string of characters for identifying an **abstract** or **physical** resource”

Example:

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

General syntax:

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



University of Southampton

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

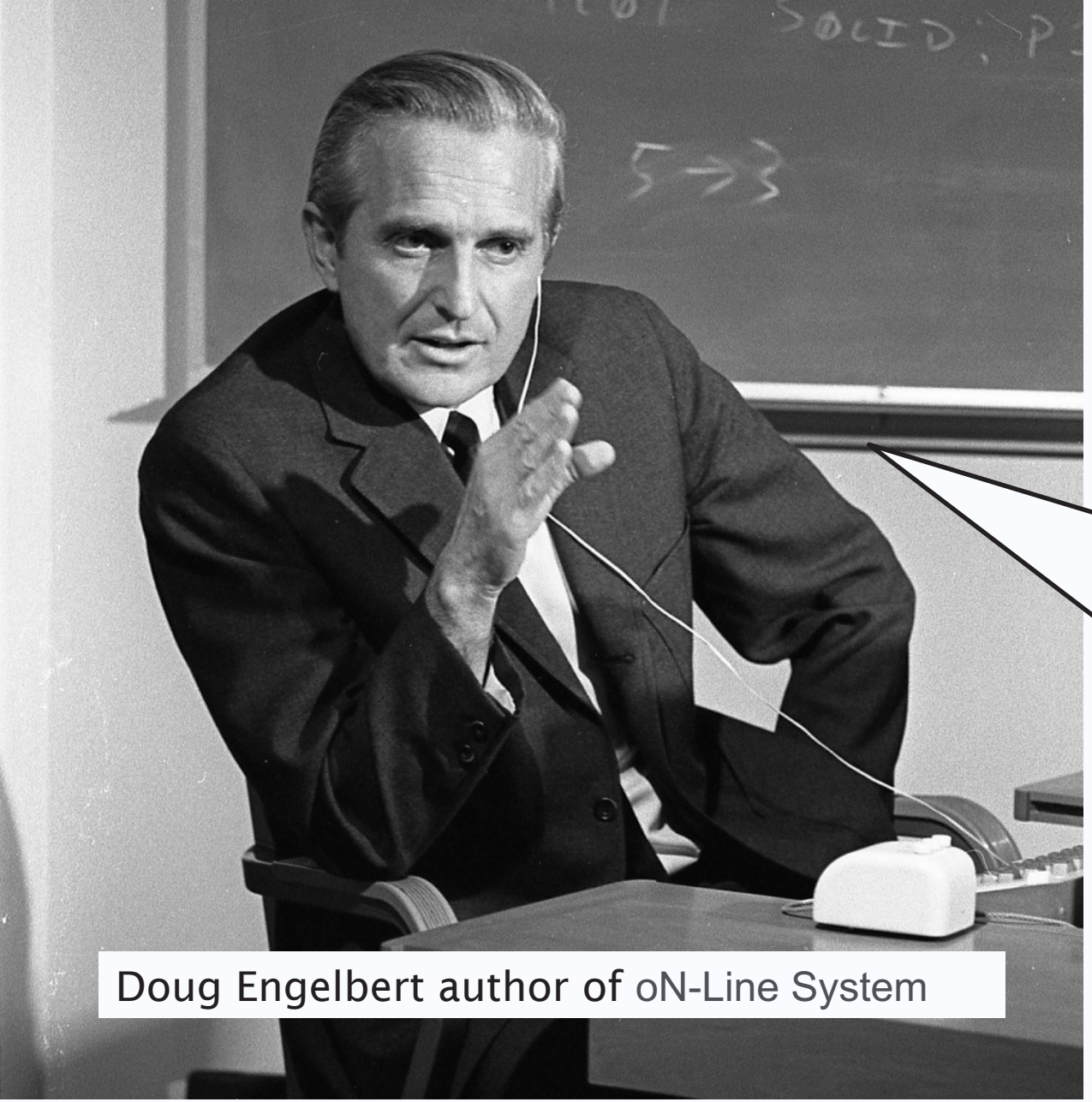
W3C's Identification Principles

1. Identifiers should be global

Global naming leads to global network effects.

We want to avoid creating walled gardens.

- 1980's CompuServe provided a bulletin board system



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.

W3C's 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.

Can't rely on context on the page

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.

W3C's 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

W3C's Identification Principles

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

- A canonical URL is the URL of the page that Google thinks is most representative from a set of duplicate pages on your site
  - Google Search Console Help
- `<link rel="canonical" href="https://www.website.com/page/" />`

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

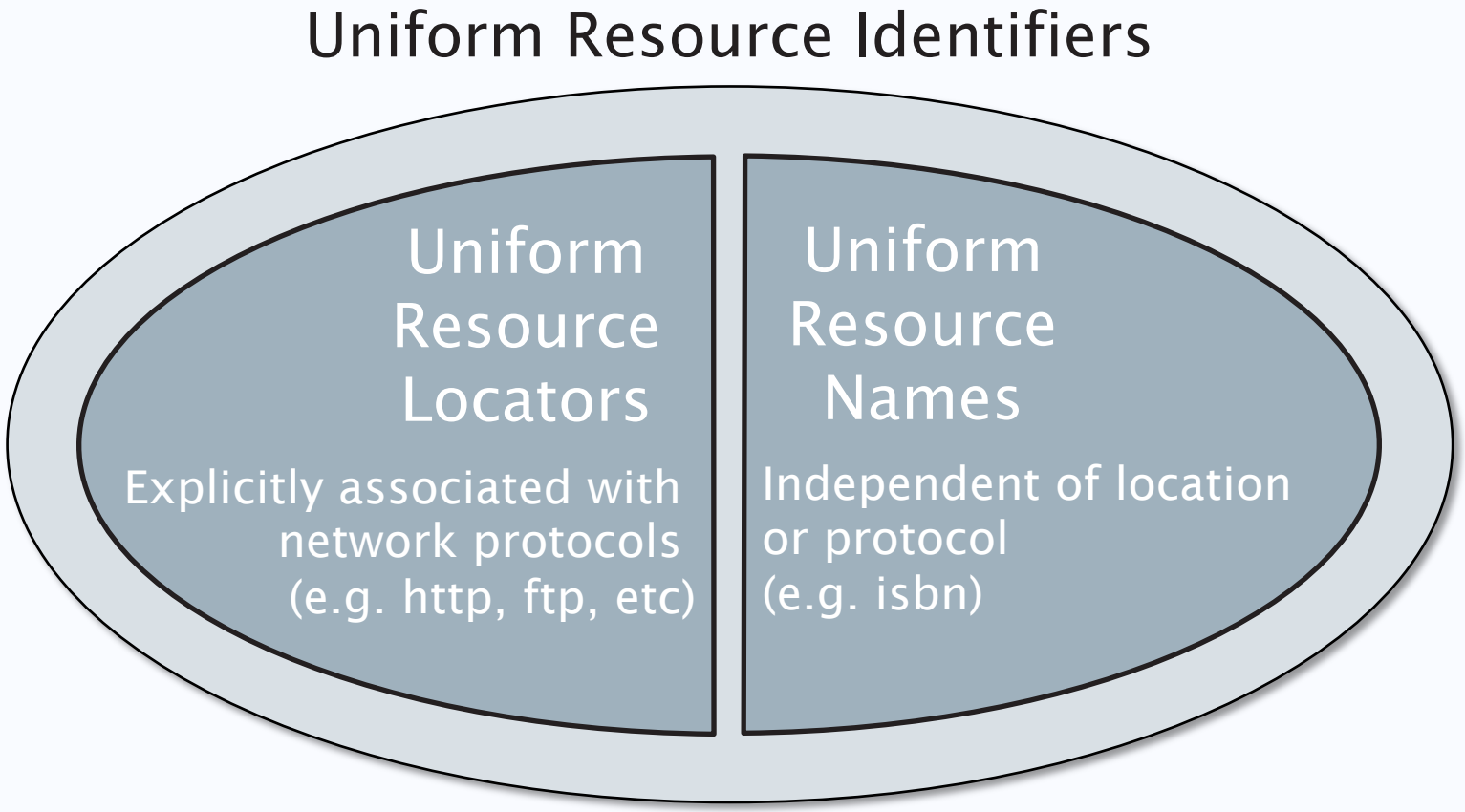
Early documents refer to document naming:

"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



Name resolution

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

- http://en.wikipedia.org/
- Hostname: en.wikipedia.org
- IP address: 91.198.174.192
- TCP socket port 80

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?
- urn:isbn:978-0471983125

Dynamic Delegation Discovery System

W3C and IETF worked on solving the URN resolution problem between 1998 and 2002

- Use part of URN to look up rule (via specific DNS records)
- Apply rule to URN to rewrite it into a URL
- Resolve URL in normal way

Did not gain traction

- Complex interaction, slow

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

University of Southampton

The Modern View

Classical view: URIs should either be URLs or URNs

Modern view: URIs should always be URLs

URL is a more widely accepted, understood, and supported

- "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

Use http/https schema (regardless of the type of resource) and ensure that it resolves to give "something useful"

- Simpler to resolve and cheaper than DDDS-like solutions
- Benefit from HTTP infrastructure such as caching and ubiquity of web servers
- Easy to produce and consume
- What does "something useful" mean when the URL refers to something that isn't on the web?

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)

http://💩.to/ → http://xn--ls8h.to/

Duerst, M. and Suignard, M. (2005) *Internationalized Resource Identifiers*. RFC3987. Available at: https://tools.ietf.org/html/rfc3987Klensin, J. (2010) *Internationalized Domain Names in Applications: Protocol*. RFC5891. Available at: https://tools.ietf.org/html/rfc5891

University of Southampton

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

University of Southampton

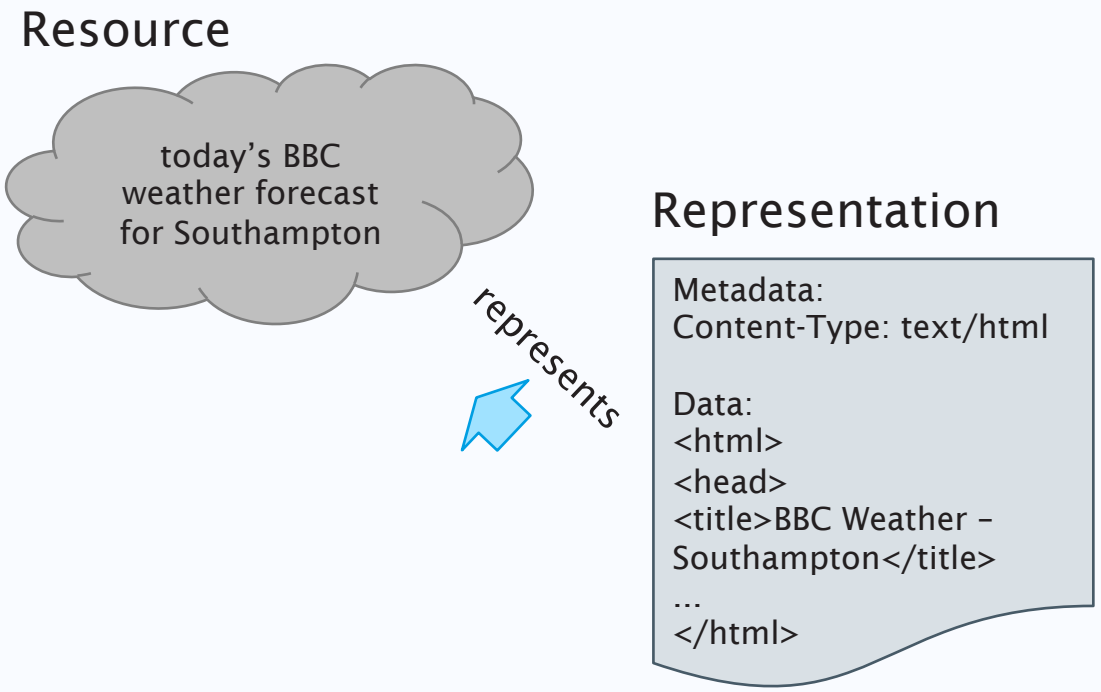
Representation

Representation

A representation is data that encodes information about resource state.

Representations have metadata

- When were they last modified?
- What format are they in?



University of Southampton

Internet Media Types

Hierarchical descriptions of data types (used originally in email - MIME)

Top-level types: `text`, `image`, `audio`, `video`, `application`
(also `multipart` and `message`)

Refinements of these top-level types:

- `text/plain`, `text/html`, `text/xml`, `text/csv`, …
- `image/jpeg`, `image/gif`, `image/png`, `image/tiff`, …
- `audio/mpeg`, `audio/ogg`, …
- `video/mp4`, `video/quicktime`, …
- `application/ecmascript`, `application/pdf`, `application/rdf+xml`, …

Registry of types maintained by Internet Assigned Numbers Authority (IANA)

https://www.iana.org/assignments/media-types/media-types.xhtml

University of Southampton

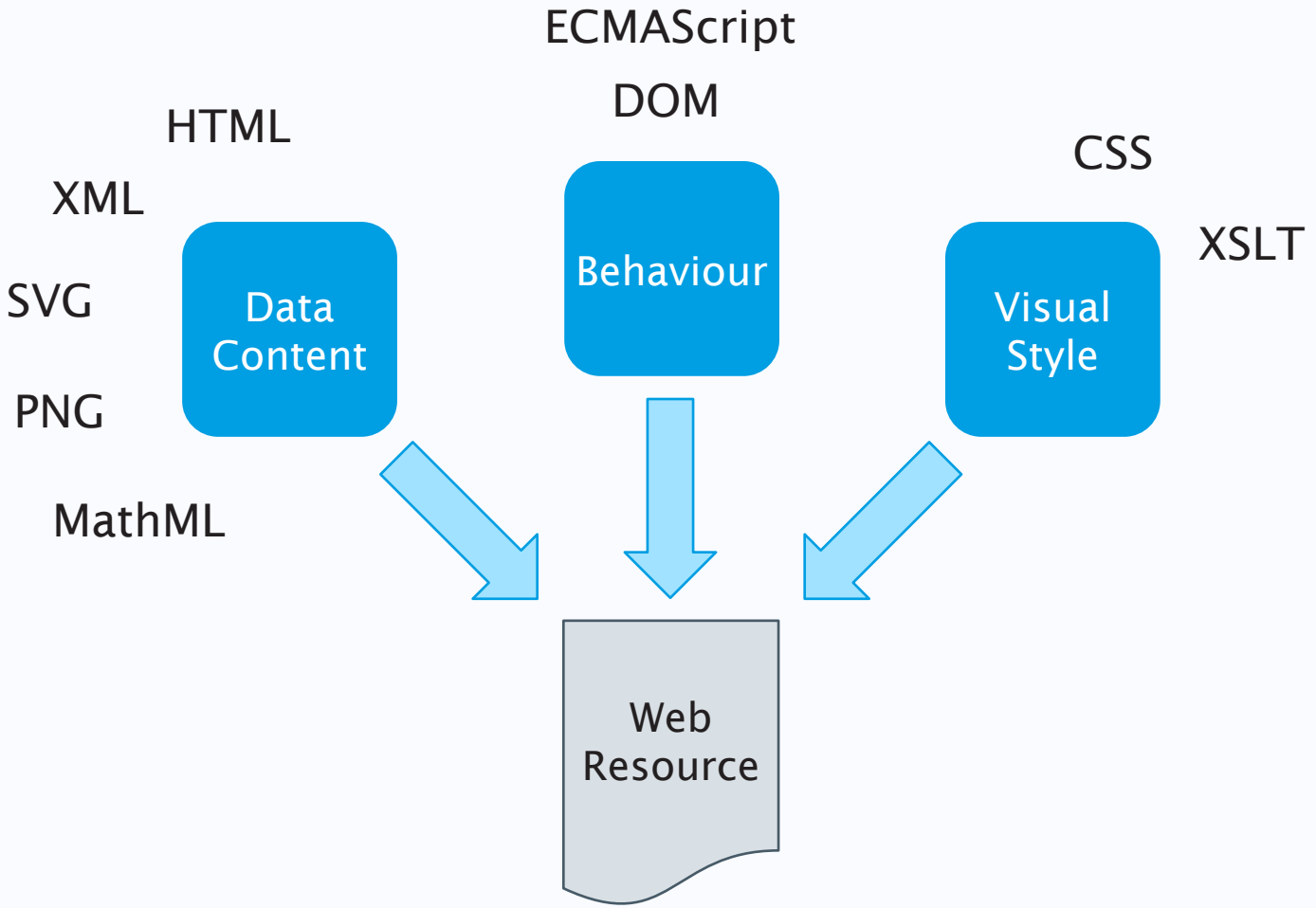
W3C's Representation Principles

1. Separate content, presentation and interaction

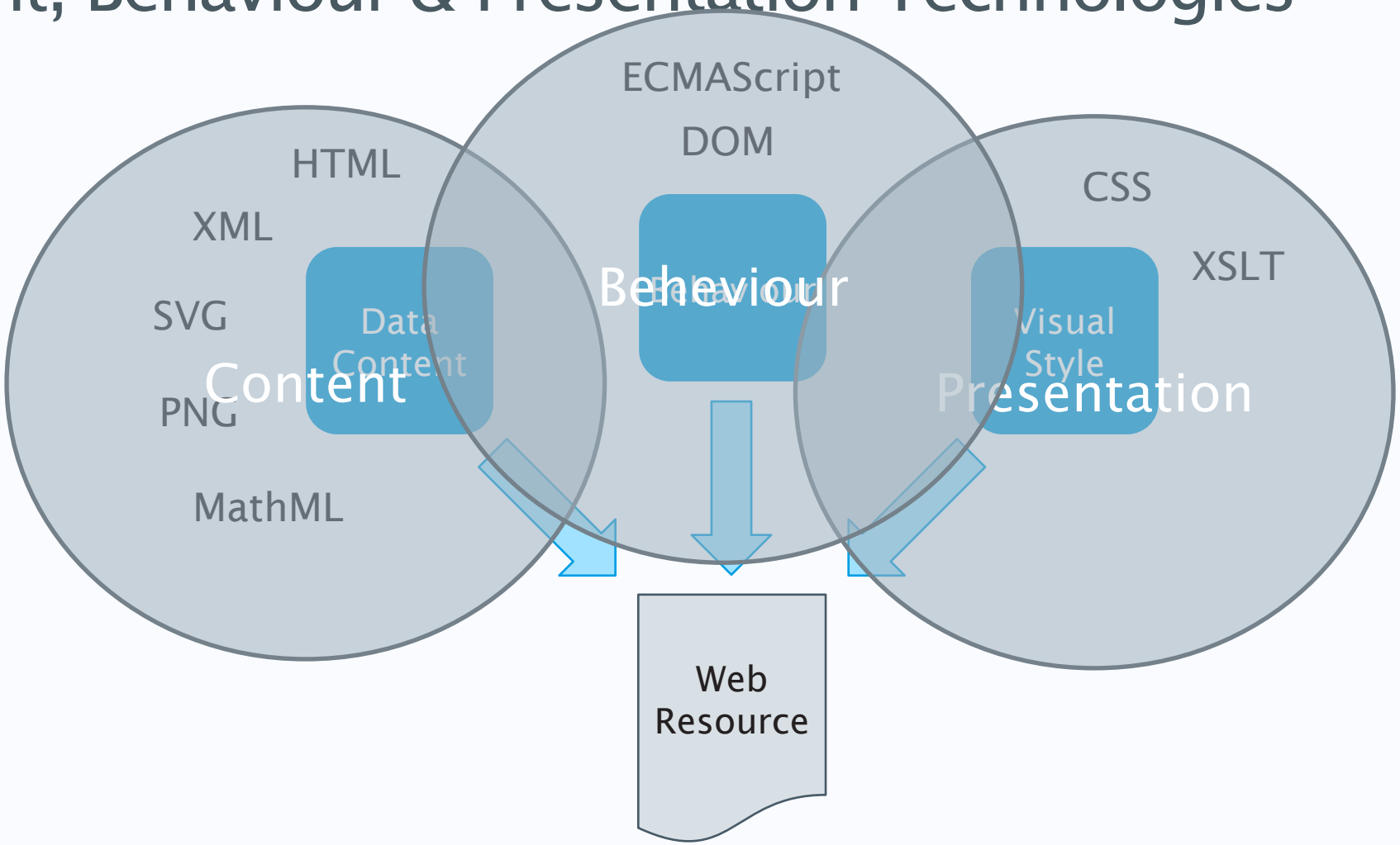
A representation format should allow authors to separate content from both presentation and interaction concerns.

How a resource is presented to a user (e.g. mobile versus desktop), and how the user interacts with that resource, are independent of the informational content of the resource.

Content, Behaviour & Presentation Technologies



Content, Behaviour & Presentation Technologies



W3C's Representation Principles

1. Separate content, presentation and interaction
2. Identify links

A representation format should provide ways to identify links to other resources, including to secondary resources (via fragment identifiers).

Html <a> tag

Some hypertext systems had no to embed links

W3C's Representation Principles

1. Separate content, presentation and interaction
2. Identify links
3. Links should be web-wide

A representation format should allow Web-wide linking, not just internal document linking.

(a corollary of global identifiers)

Identification – refer to anything

Representation – link to anything

W3C's Representation Principles

1. Separate content, presentation and interaction
2. Identify links
3. Links should be web-wide
4. Links should use generic identifiers

A representation format should allow content authors to use URIs without constraining them to a limited set of URI schemes.

Formats should be future-proof; we don't know what identifier types or protocols we'll be using in the future.

W3C's Representation Principles

1. Separate content, presentation and interaction
2. Identify links
3. Links should be web-wide
4. Links should use generic identifiers
5. Links should be navigable

A representation format should incorporate hypertext links if hypertext is the expected user interface paradigm.

We would like links between resources to be able to behave like any other hypertext links.

University of Southampton

Interaction

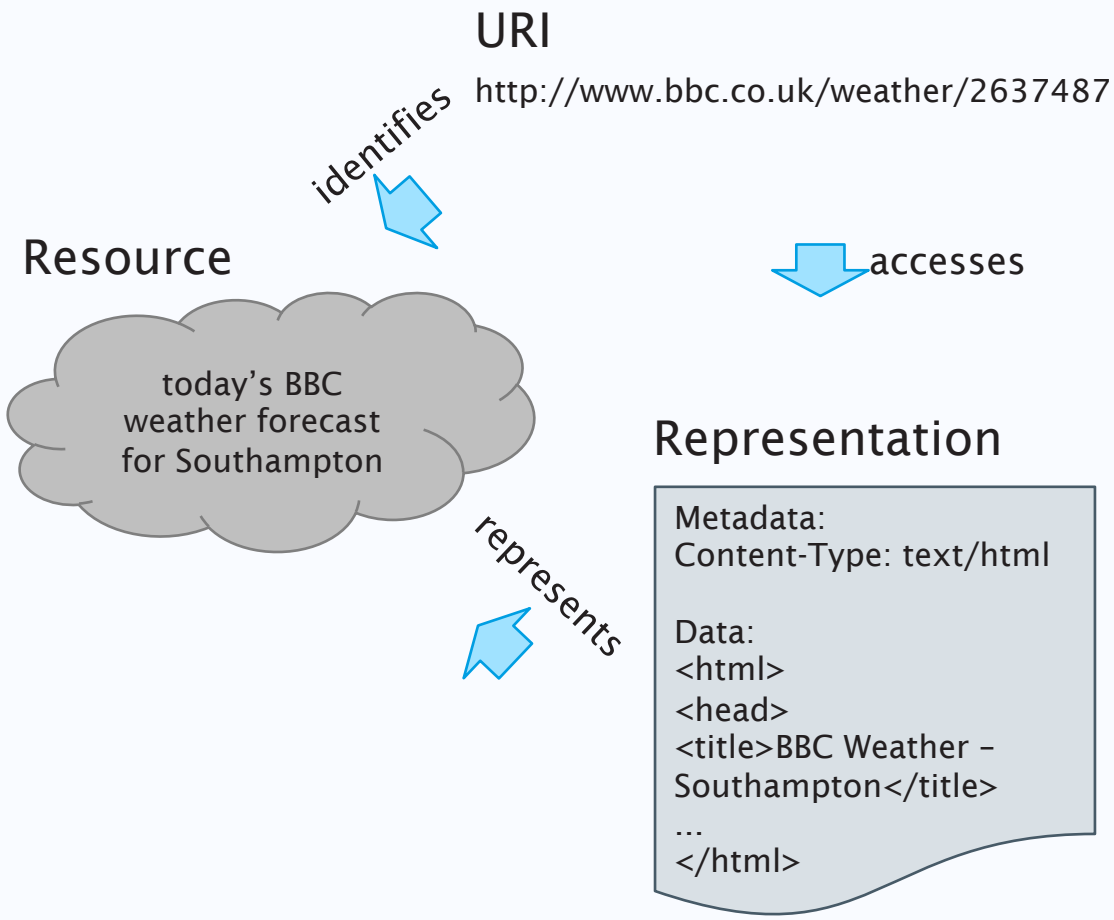
Interaction

Resource representations are transmitted using interaction protocols that specify the exchange of messages

- HTTP, FTP, SOAP, NNTP, SMTP, ...

Messages contain both:

- *data*
  (informational content of the message)
- *metadata*
  (description of the message or its content)



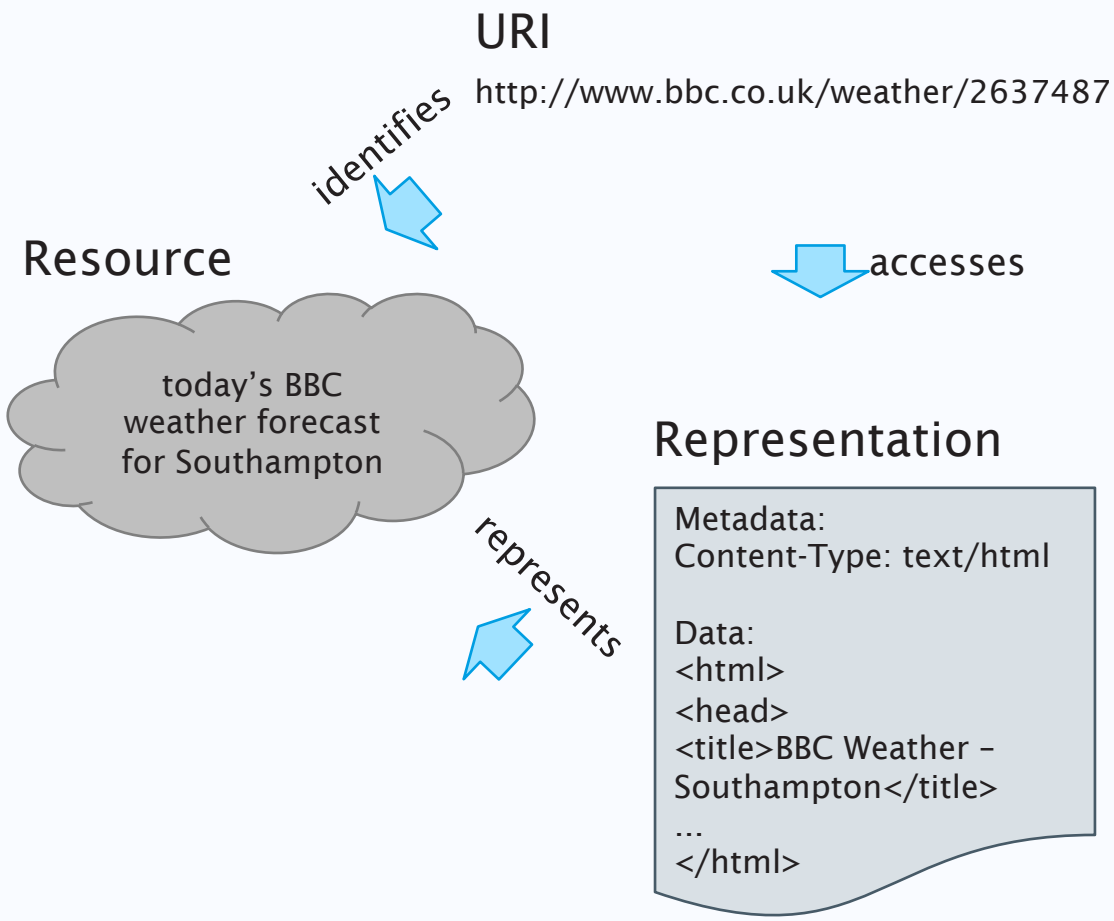
Dereferencing URIs

URIs used to identify resources may indicate protocols that can be used to access those resources

- Though not always: caches, proxies, name resolution services (DNS)
- Many URI schemes define a default interaction protocol

Resource access takes several forms:

- Retrieving a representation of the resource
- Adding or modifying a representation of the resource
- Deleting some or all representations of the resource



University of Southampton

W3C's Interaction Principles

1. Reuse representation formats

Any new protocols created for the Web should transmit representations as octet streams typed by Internet media types.

https://www.iana.org/assignments/media-types/application/octet-stream

W3C's Interaction Principles

1. Reuse representation formats
2. Provide representations

A URI owner should provide representations of the resource it identifies.

There is a general expectation that it should be possible to retrieve a representation of any resource.

W3C's Interaction Principles

1. Reuse representation formats
2. Provide representations
3. Retrieval should be safe

Agents do not incur obligations by retrieving a representation.

Put another way, the act of retrieving a representation of a resource should not have any significant side-effects (for example, deleting the resource or changing its state).

W3C's Interaction Principles

1. Reuse representation formats
2. Provide representations
3. Retrieval should be safe
4. Reference does not imply dereference

An application developer or specification author should not require networked retrieval of representations each time they are referenced.

Just because you *can* retrieve a representation of a resource, doesn't mean that you *must.*

Example: URIs used to identify document schemas:
`http://www.w3.org/TR/html4/strict.dtd`

W3C's Interaction Principles

1. Reuse representation formats
2. Provide representations
3. Retrieval should be safe
4. Reference does not imply dereference
5. Representations should be consistent

A URI owner should provide representations of the identified resource consistently and predictably.

We want our identifiers to be persistent: once an identifier has been associated with that resource, it should continue to refer to that resource indefinitely.

(a matter of policy, not technology)

University of Southampton

Further Reading

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

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

Fielding, R.T. (2000) *Architectural Styles and the Design of Network-based Software Architectures*. PhD Thesis. University of California at Irvine. Chapter 5.

http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

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

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

Fielding, R.T. and Reschke, J. (2014) *Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content*. RFC7231.

https://tools.ietf.org/html/rfc7231

University of Southampton

Next Lecture: HTTP