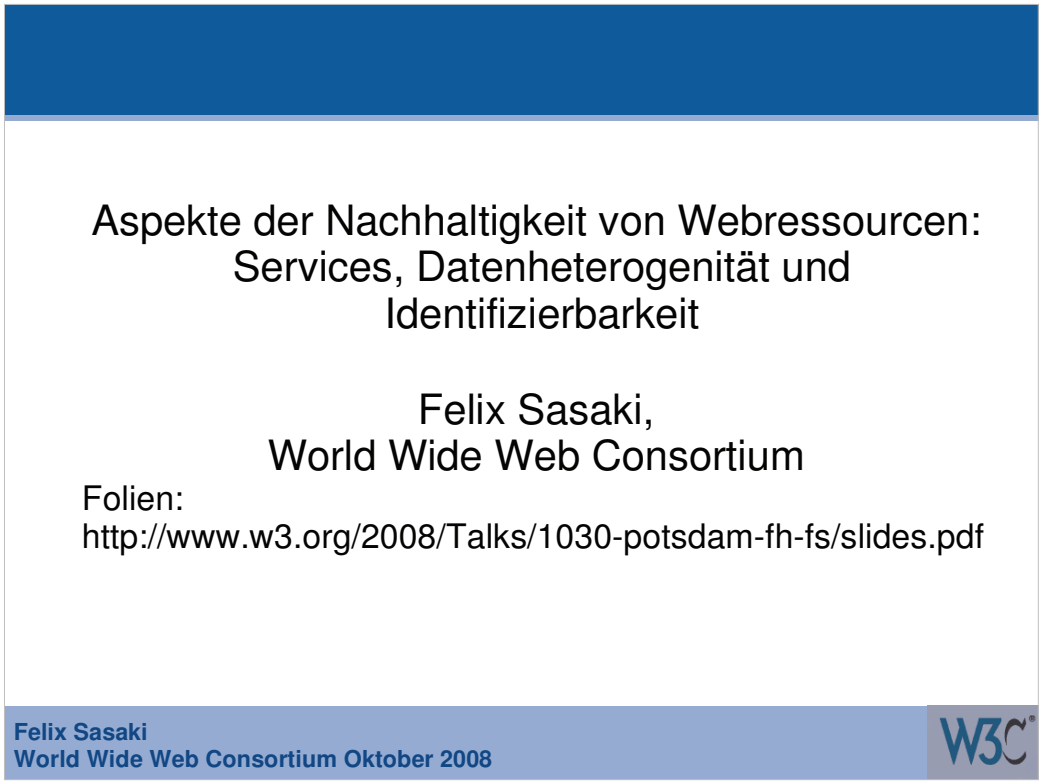


Dieser Vortrag beschreibt zentrale Aspekte, um die Nachhaltigkeit von Webressourcen zu sichern. Um die Spannung gleich vorweg zu nehmen, sei die Kernaussage schon jetzt verraten: Die Nachhaltigkeit von Webressourcen ist ein soziales Problem. Technologien sind ein Mittel, das zur Lösung eingesetzt werden kann – oder auch nicht.

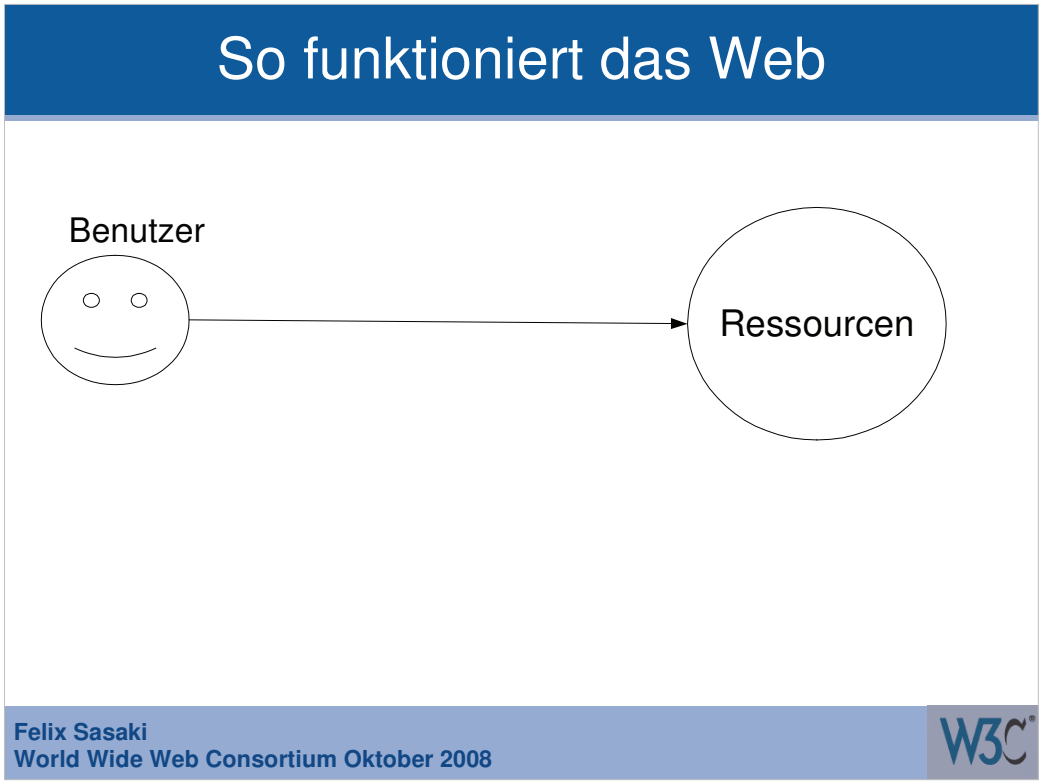
Falls Sie also vor allem an dieser Kernaussage interessiert sind und nicht an Details, dann können Sie jetzt schon nach Hause gehen und früh in den Feiertag starten.

Prolog

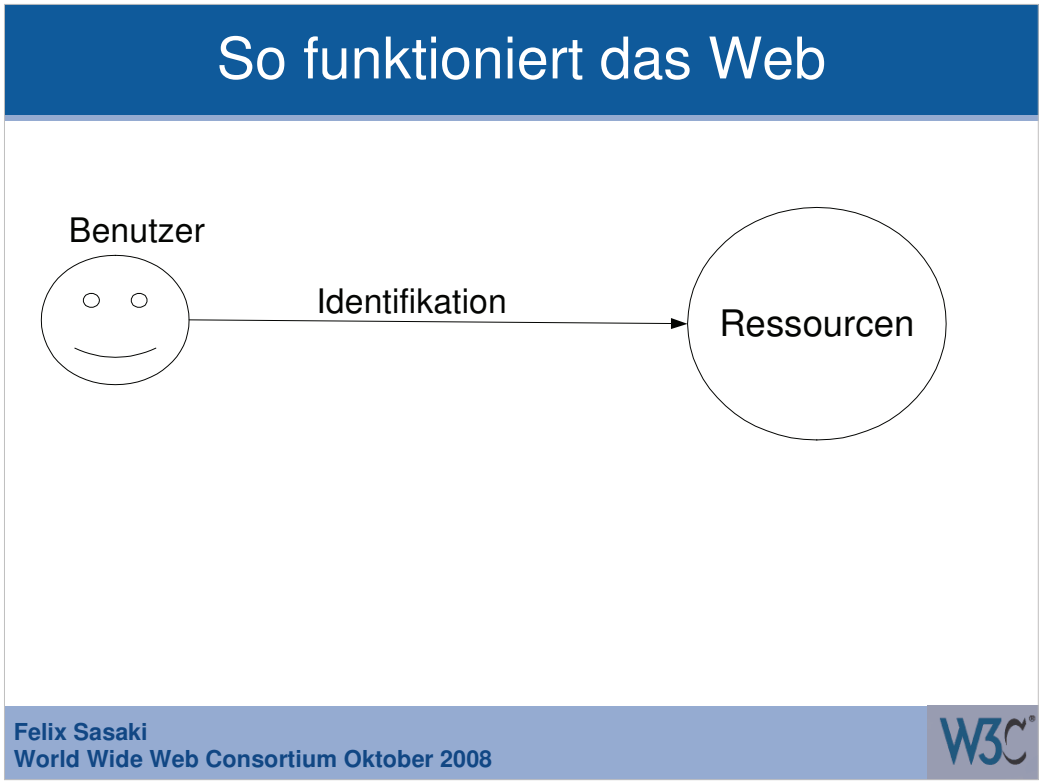
So funktioniert das Web

Felix Sasaki
World Wide Web Consortium Oktober 2008

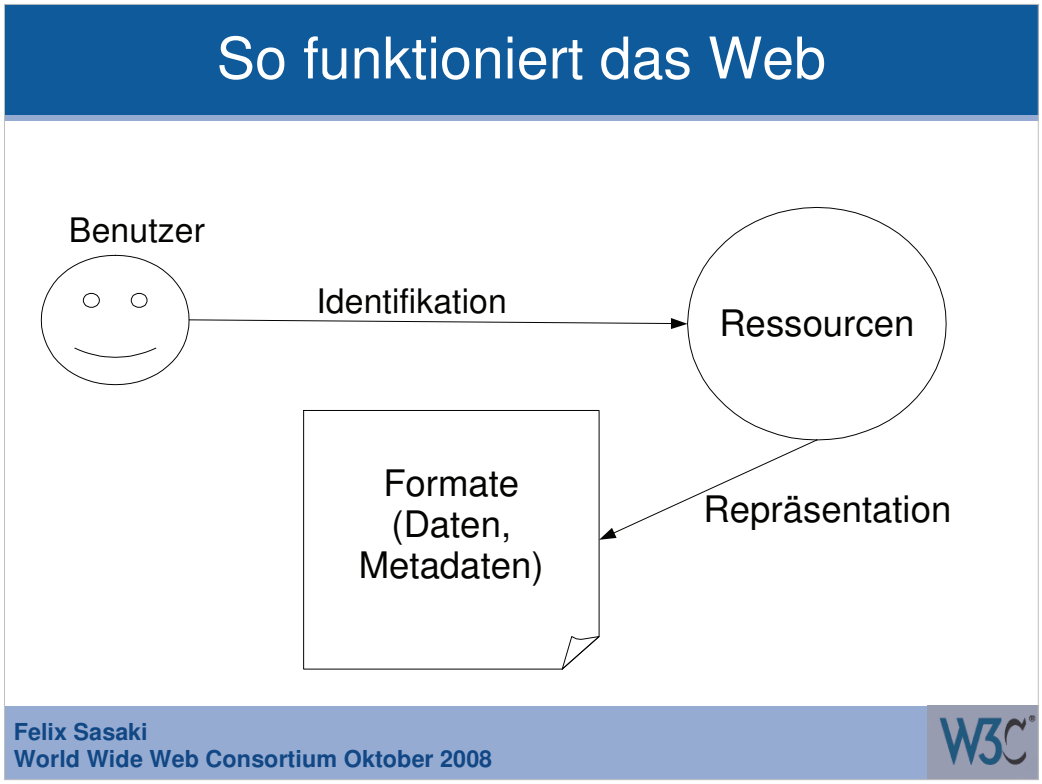
Dieser Vortrag verlangt kein technisches Vorverständnis. Stattdessen werde ich Ihnen in den nächsten Minuten eine Beschreibung der Funktionsweise des Webs geben. Anschließend zeige ich wie bei den jeweiligen Webfunktionen – es sind übrigens drei – Nachhaltigkeit mehr oder weniger gewährleistet werden kann.



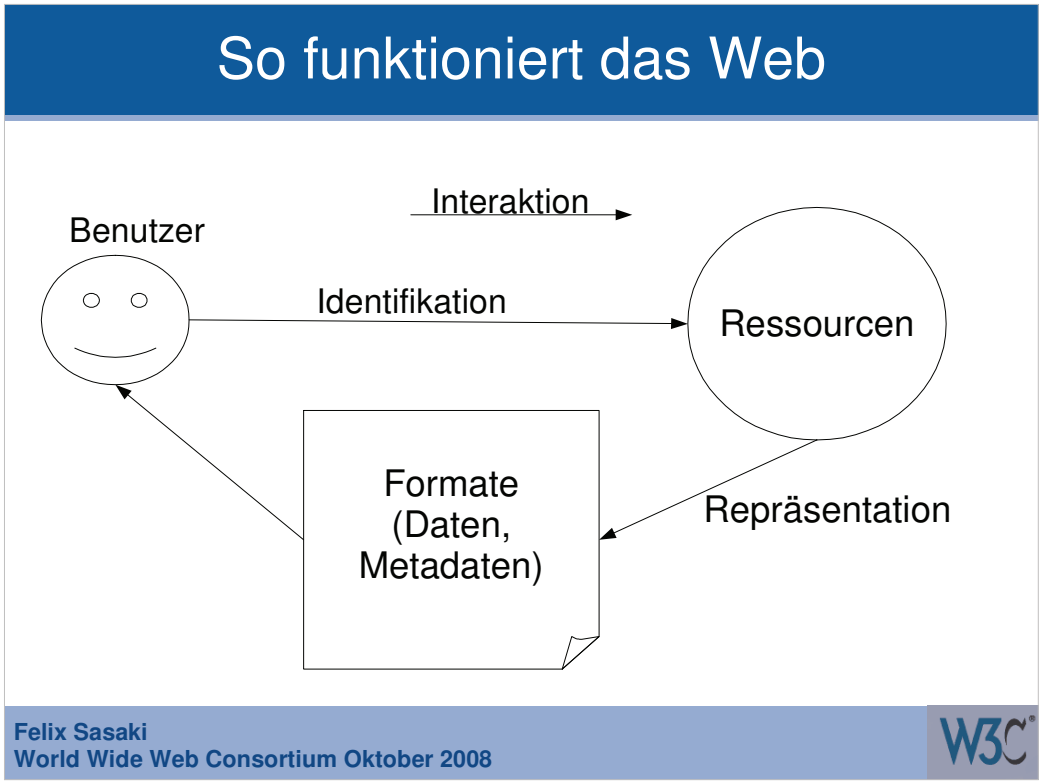
Am wichtigsten im Web sind die Nutzer. Hier sind sie schematisch dargestellt. Benutzer tun sehr verschiedene Dinge im Web - für das Thema dieses Vortrages ist es entscheidend, dass Benutzer Zugang zu Webressourcen haben möchten – zu textuellen Dokumenten, Bildern, Audio- oder Videomaterial oder Daten jedweder Art.



Um Zugang zu Ressourcen zu erlangen müssen diese identifizierbar sein. Zur Identifikation werden Identifikatoren benutzt: die „Adressen des Web“. Die Nachhaltigkeit dieser Identifikatoren ist ein Aspekt mit dem wir uns später befassen werden.

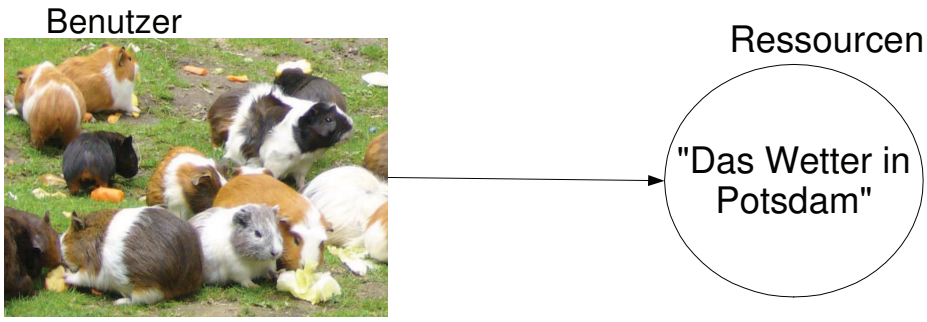


Mit Hilfe von Identifikatoren können Benutzer Zugang zu Ressourcen erlangen. Diese liegen in Formaten vor, zum Beispiel HTML für Textdokumente oder PNG für Grafiken. Formate umfassen Metadaten und Daten. Ein HTML-Dokument hat zum Beispiel einen Titel und eventuell Metainformationen zur Sprache (Deutsch, Englisch, Japanisch etc.) des Dokuments, und den eigentlichen Text.



Wir haben nun Benutzer, die Ressourcen mittels Identifikatoren identifizieren, und Formate, in denen die Ressourcen vorliegen. Was jetzt noch fehlt sind Regeln der Interaktion zwischen Benutzer, Ressourcen und Formaten. Diese Regeln werden oft als Protokolle bezeichnet. HTTP ist ein weit verbreitetes Protokoll, andere sind zum Beispiel FTP oder Mail.

So funktioniert das Web: Beispiel



Felix Sasaki
World Wide Web Consortium Oktober 2008

Nach dieser abstrakten Beschreibung von Webfunktionen – Identifikation von Ressourcen, Repräsentation von Formaten, Interaktionen mittels Protokollen – möchte ich Ihnen ein konkretes Beispiel zeigen. Hier sehen Sie eine Menge von Benutzer die den Tag gerne im Freien verbringen. Deshalb möchten sie natürlich etwas über das Wetter wissen, zum Beispiel das Wetter in Potsdam. Die Ressource um die es jetzt also geht ist der Wetterbericht in Potsdam. Offensichtlich ist die Interpretation dieser Ressource nicht beständig. Egal ob man den Wetterbericht im Web, in der Zeitung, im Fernsehen oder von einem befreundeten Meteorologen hört, man möchte den jeweils aktuellen Bericht bekommen. Dies spielt jedoch für die Frage der Nachhaltigkeit keine Rolle. Wichtig ist dass die drei beschriebenen Funktionen nachhaltigen Zugang zum Wetterbericht gewährleisten.

Identifikation

Benutzer

Ressourcen

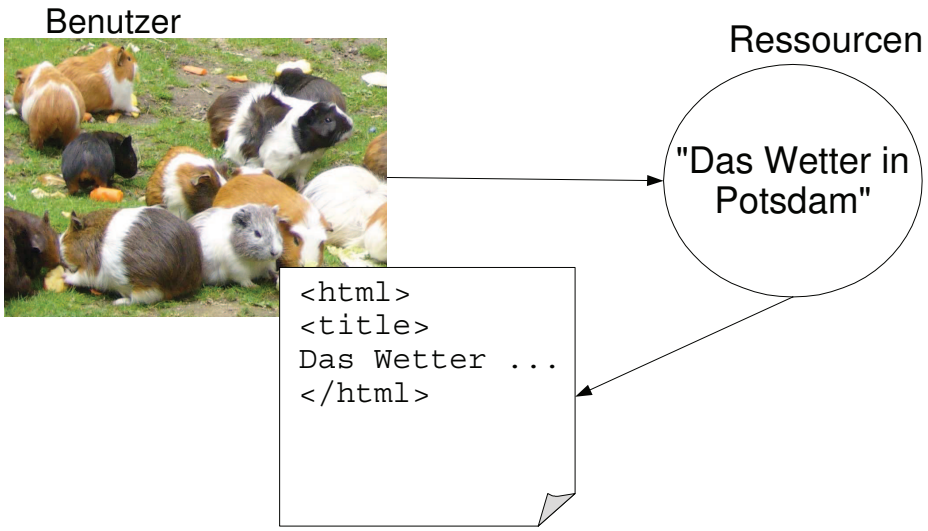
"Das Wetter in Potsdam"

`http://example.org/wetter-potsdam/`

URI (Universal Resource Identifier)

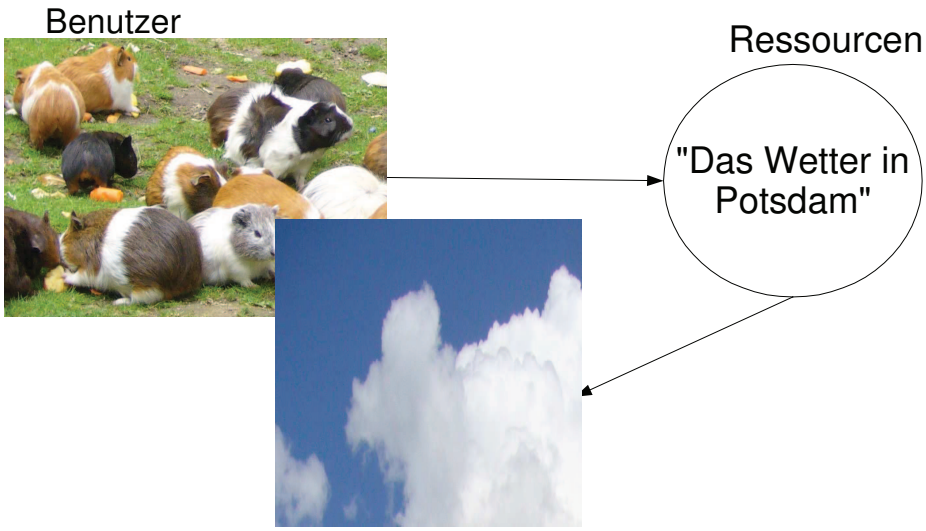
Zunächst zur Identifikation. Unsere Benutzer können den aktuellen Wetterbericht in Potsdam über die Webadresse, die so genannte URI, identifizieren. Hier habe ich eine exemplarische URI erfunden: http://example.org/wetter-potsdam/

Repräsentation (I): Dokument



Diese Webadresse, diese URI, führt die Benutzer zu einer Repräsentation des Wetterberichts, zum Beispiel ein HTML Dokument mit dem Titel "Das Wetter in Potsdam".

Repräsentation (II): Bild



Es wäre auch möglich dass die gleiche Webadresse die Benutzer zu einer anderen Repräsentation der Ressource führt. Auf der vorliegenden Folie habe ich ein Bild des Himmels von Potsdam aufgenommen. Ehrlich gesagt ist das Bild nicht von Heute – es wäre schön wenn heute so ein Wetter wäre –, aber ich denke es reicht um etwas deutlich zu machen: URIs, die Identifikatoren im Web, sind nicht zwangsweise mit Ressourcen in einem bestimmten Format verbunden.

Interaktion (z.B. HTTP Protokoll)

Request

```
GET /wetter-potsdam/
Host: example.org
Accept: text/html
```

```
<html>
<title>
Das Wetter ...
</html>
```

Response

http://example.org/wetter-potsdam/

Felix Sasaki
World Wide Web Consortium Oktober 2008

Woher wissen aber nun die Benutzer in welchem Format eine Ressource vorliegt? Nun, sie wissen es gar nicht, aber sie können ihre Wünsche in der Interaktion, z.B. mittels des HTTP Protokolls, ausdrücken. Hier ist der Request, der Beginn einer HTTP Interaktion dargestellt. Der Request drückt aus, dass die Benutzer, zum Beispiel Webbrowser, eine Ressource von der Domäne http:://example.org bekommen möchten. In dieser Domäne befindet sich die Ressource /wetter-potsdam/ . Die komplette Webadresse lautet also wie bereits gesehen http://example.org/wetter-potsdam/ .

Die Benutzer machen nun ihren Wunsch über das Format mit der Aussage
Accept: text/html
deutlich. Dies bedeutet dass die Benutzer Ressourcen ihm Format HTML erhalten möchten.

Sie mögen sich fragen wer solche Interaktionen auslöst – wenn es nicht die Meerschweinchen sind . Der Täter ist oft der Browser, der einen HTTP Request an einen Webserver schickt und – wenn alles klappt – die HTML Seite als Antwort bekommt. Interaktionen wie auf dieser Folie beschrieben sind den Benutzern nicht unbedingt bewusst, finden aber im Web ständig statt.

Überblick: Nachhaltigkeit von ...

- ... Interaktionen: Services ("Gib mir das Wetter!" → Wetterdokument, Wetterbild)
- ... Identifikation: Identifikatoren ("Identifikator des Wetterberichts")
- ... Formaten (Bildformat, Textformat, ...)

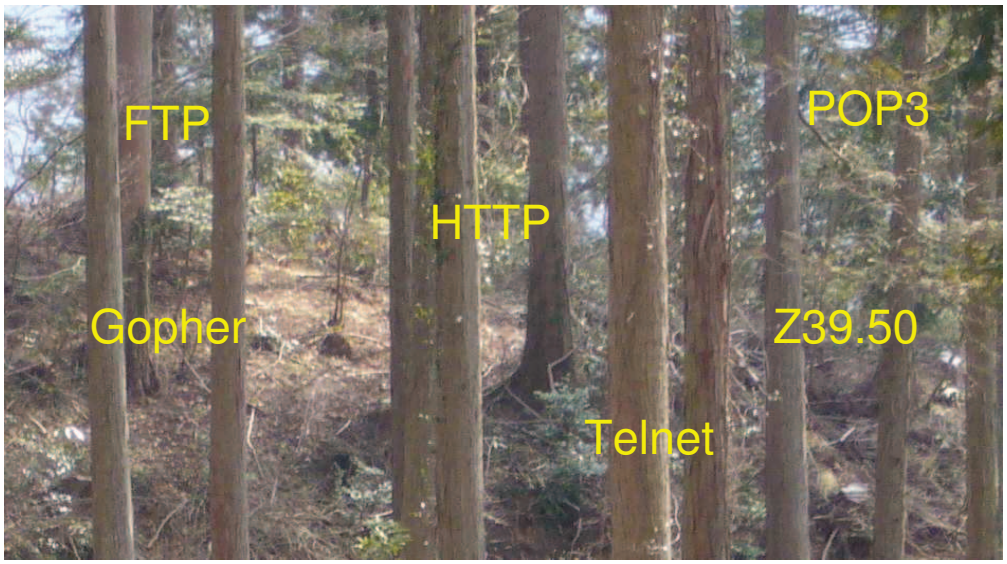
Die Nachhaltigkeit, um die es im Folgenden geht, betrifft also die drei Bereiche Interaktion, Identifikation, und Formate.

Interaktionen

Felix Sasaki
World Wide Web Consortium Oktober 2008

Interaktion bezeichne ich hier auch als "Services". Dies macht deutlich dass nicht nur die menschlichen oder "tierischen" Browserbenutzer mit Webressourcen interagieren, sondern auch Maschinen, also Computer im Web miteinander kommunizieren. Wenn Sie etwa ein Hotel per Internet reservieren wird dabei vom Computer des Reisebüros eine Verbindung zur Kreditkartenfirma aufgebaut, um die Gültigkeit Ihrer Karte zu überprüfen. Die folgenden Aspekte der Nachhaltigkeit umfassen auch solche Interaktionen.

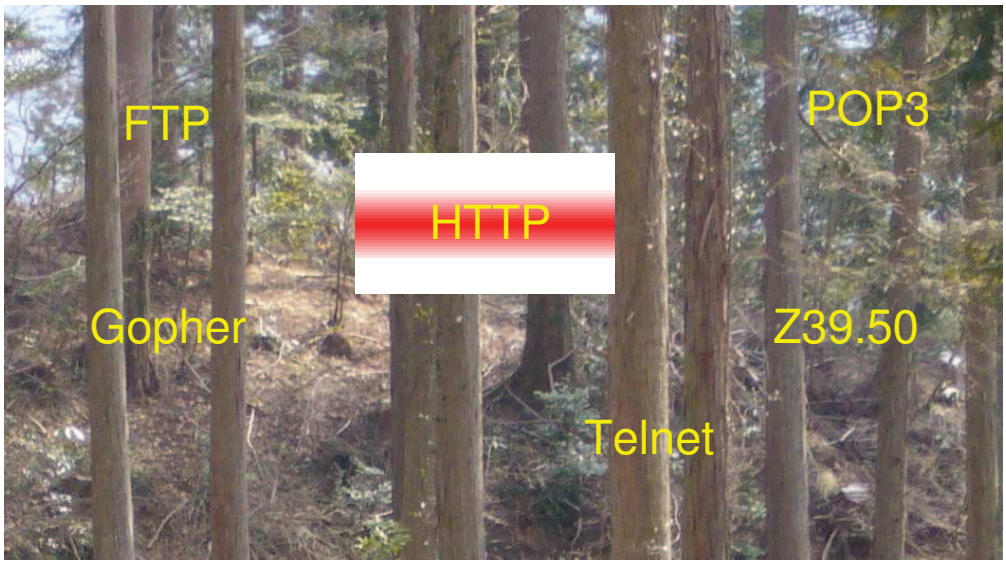
Der Protokollwald



Felix Sasaki
World Wide Web Consortium Oktober 2008

Interaktionen beruhen auf Protokollen, und wie gesagt gibt es eine Vielzahl von Protokollen im Web – ein regelrechter „Protokollwald“. Vor HTTP sind viele Protokolle entwickelt worden, z.B. das auf Text beschränkte Gopher oder Telnet. Zudem gibt es Community spezifische Protokolle wie Z39.50, das für bibliographische Daten bedeutsam ist, oder Protokolle für bestimme Aufgaben wie Transfer von Dateien (FTP), oder eine Art von Mailzugriff (POP3).

Die Lichtung



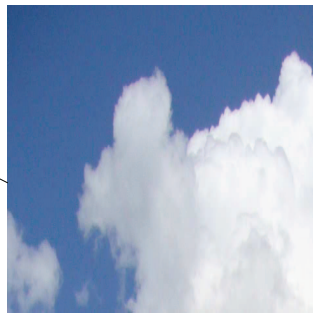
Felix Sasaki
World Wide Web Consortium Oktober 2008

Die Lichtung in diesem Wald für nachhaltige Services ist das HTTP Protokoll. Oder anders gesagt: nachhaltige Services sollten dieses Protokoll nutzen. Es ermöglicht die wichtigsten Operationen in einer Interaktion, z.B. mittels eines GET Requests den Benutzern (Browser) Daten und Metadaten zu übermitteln.

Services (I, empfehlenswert)

`http://example.org/wetter-potsdam/`

GET request



response

Neben der Wahl des Protokolls gibt es einen weiteren Aspekt von Services, der für Nachhaltigkeit bedeutsam ist. Er betrifft das Verhältnis von Ressourcen und Identifikatoren. Im vorliegenden Beispiel ist für eine Ressource, den Wetterbericht von Potsdam, genau ein Identifikator vergeben:

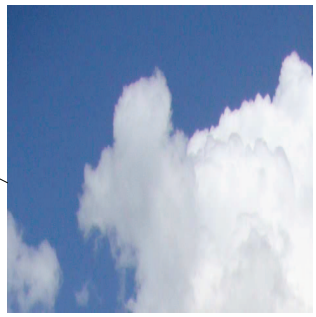
http://example.org/wetter-potsdam

Dieses Vorgehen ist insbesondere für nachhaltige Ressourcen empfehlenswert. Warum werden wir gleich beim nächsten, nicht empfehlenswerten Beispiel sehen.

Services (II, abzuraten)

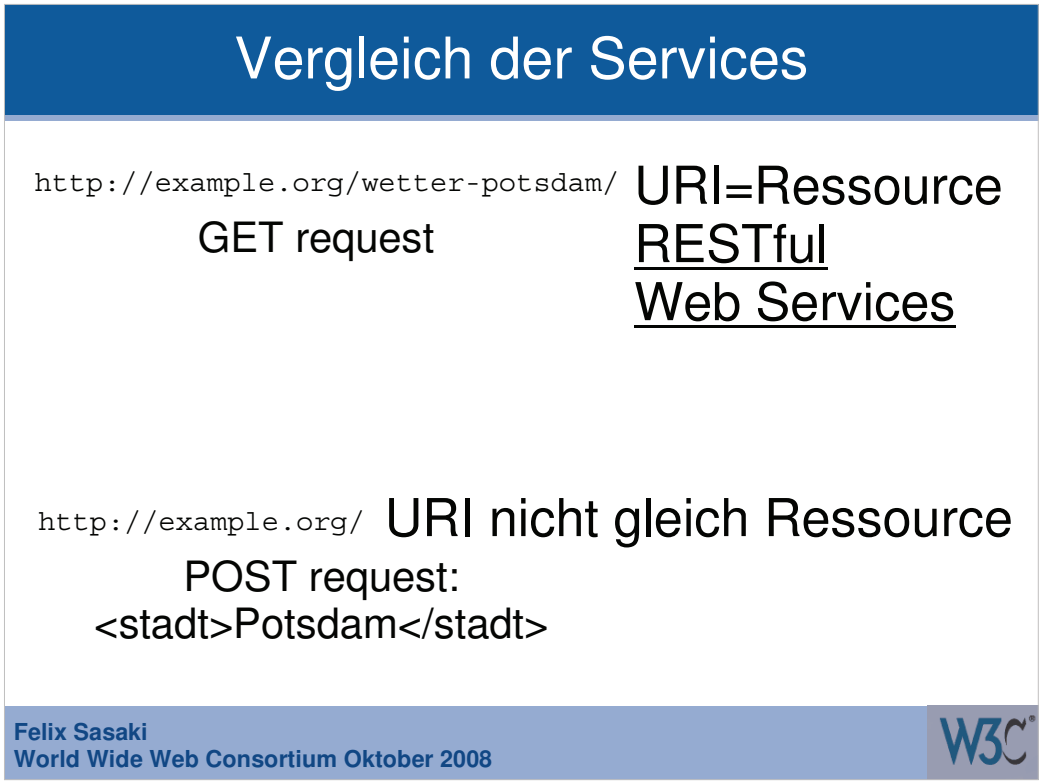
http://example.org/

POST request:
<stadt>Potsdam</stadt>

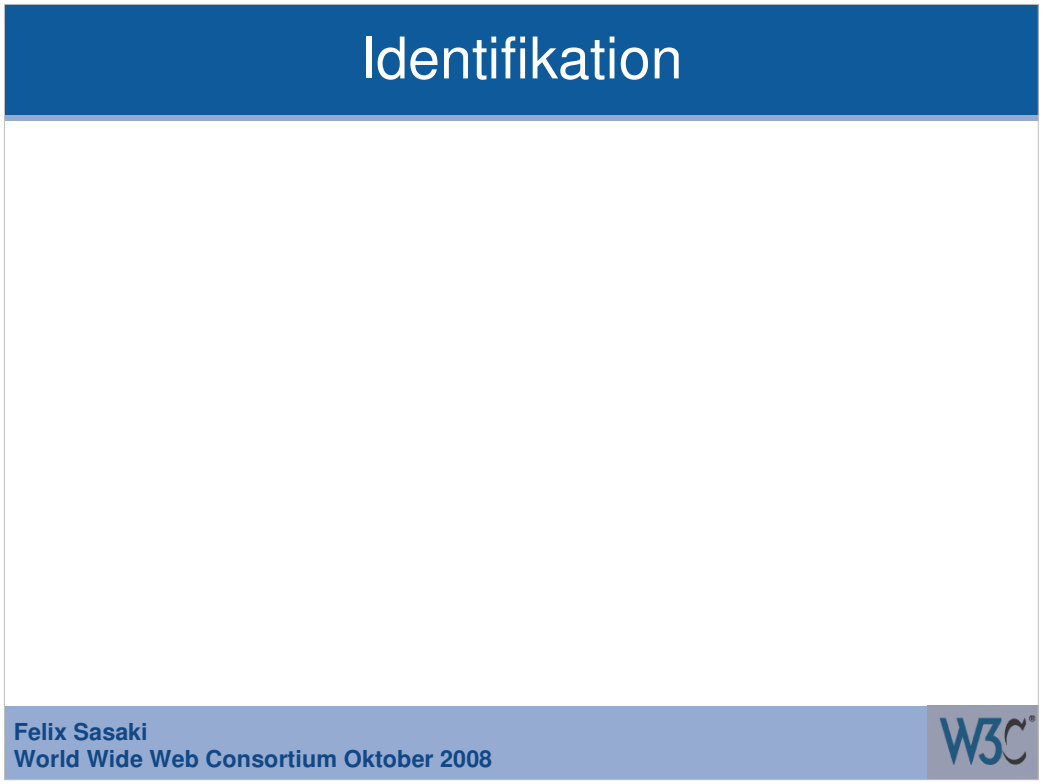


response

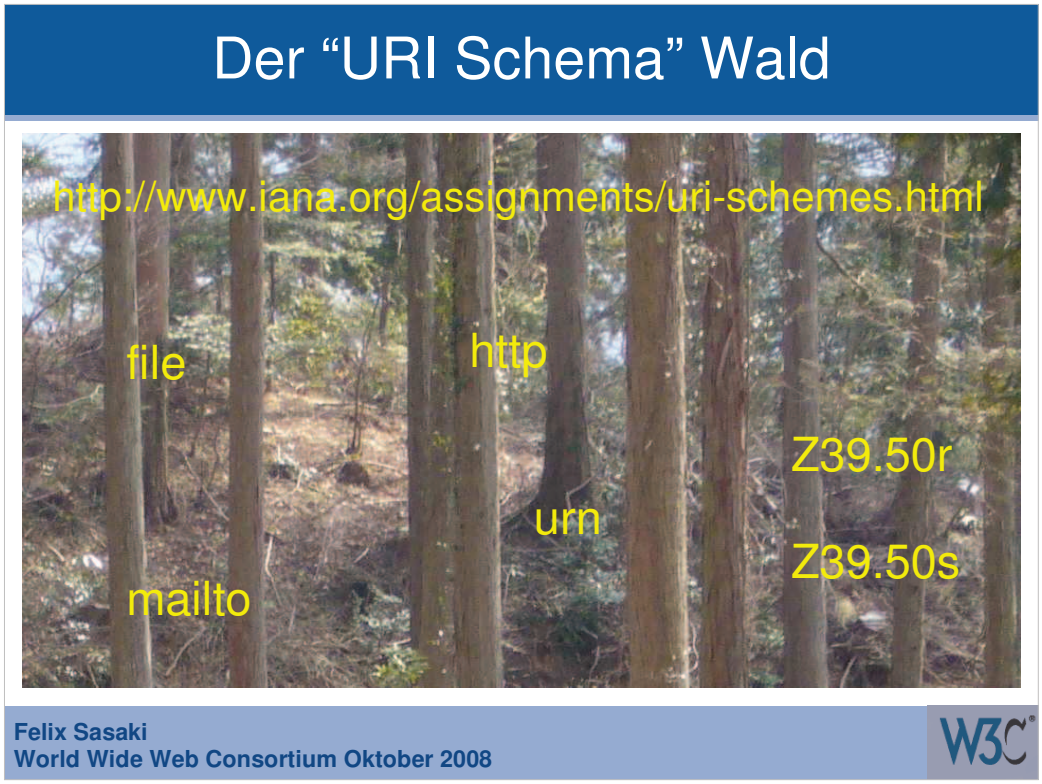
Die Aufgabe dieser Interaktion mittels der HTTP Methode „POST“ ist die gleiche wie zuvor: wir möchten den Wetterbericht von Potsdam bekommen. Allerdings ist der Identifikator den wir benutzen ein anderer: Er bezeichnet eine generelle Ressource
http://example.org/
In dieser Ressource ist der Wetterbericht "verborgen". Um ihn zu bekommen, wird als zusätzlicher Bestandteil die Information
<stadt>Potsdam</stadt>
übermittelt. Wie dieser Bestandteil aussieht ist nicht im HTTP Protokoll definiert. Eine erfolgreiche und nachhaltige Interaktion ist also nur möglich wenn die Teilnehmer zusätzliche Vereinbarungen getroffen haben, etwa hier die Information
<stadt>...</stadt>
als Teil des Requests zu übermitteln.



Der für nachhaltige Webressourcen zu bevorzugende Ansatz sind so genannte RESTful Web Services. Die Haupteigenschaft von RESTful Web Services ist dass jede Ressource eine eigene URI erhält. Dem gegenüber stehen Web Services welche Ressourcen nicht mittels URIs identifizieren, sondern in anderen Bestandteilen der Interaktion. Als Beispiel hatten wir die HTTP Methode POST mit der Information <stadt>Potsdam</stadt>
gesehen. Da der Aufbau dieser Bestandteile, z.B. der Name <stadt>, im Gegensatz zum Aufbau der HTTP URIs nicht standardisiert ist, kann Nachhaltigkeit nur schwer erzielt werden.



Bei der Diskussion von Services ging es vor allem um das Protokoll (HTTP) und die richtige Gestaltung von Identifikatoren (eine Ressource = ein Identifikator). Beide Aspekte haben viel mit den standardisierten Schemata für Identifikatoren, den so genannten URI ("Universal Resource Identifier") zu tun. Um diese wird es nun gehen.



Wie bei den Protokollen gibt es auch bei den URI Schemata einen Wald – und dieser Wald ist noch dichter. Einige Schemata sind eng mit Protokollen verknüpft, z.B. das „http“ Schema, oder die Schemata „Z39.50r“ sowie „Z39.50s“. Andere sind mit Absicht nicht an Protokolle gebunden, z.B. „urn". Manche Schema umfassen mehr oder wenige explizite Nachhaltigkeitsversprechen, vgl. „urn": „Universal Resource Name".

Die Fragen lauten nun: Welches Schema soll man wählen? Sollte man ein neues, für die jeweilige Community spezifisches Schema erfinden?

Die Lichtung: Kein neues Schema

Felix Sasaki
World Wide Web Consortium Oktober 2008

Die Antwort ist einfach und wahrscheinlich zunächst unbefriedigend: das „http" Schema ist ausreichend um nachhaltig Ressourcen zu identifizieren. Ich möchte hier die Position vertreten dass die soziale Bereitschaft, sich auf eine Form von URIs in einem existierenden Schema zu nutzen, entscheidend ist.

Nachhaltigkeit von Identifikatoren

- Ein soziales Problem - soziale Lösungen:

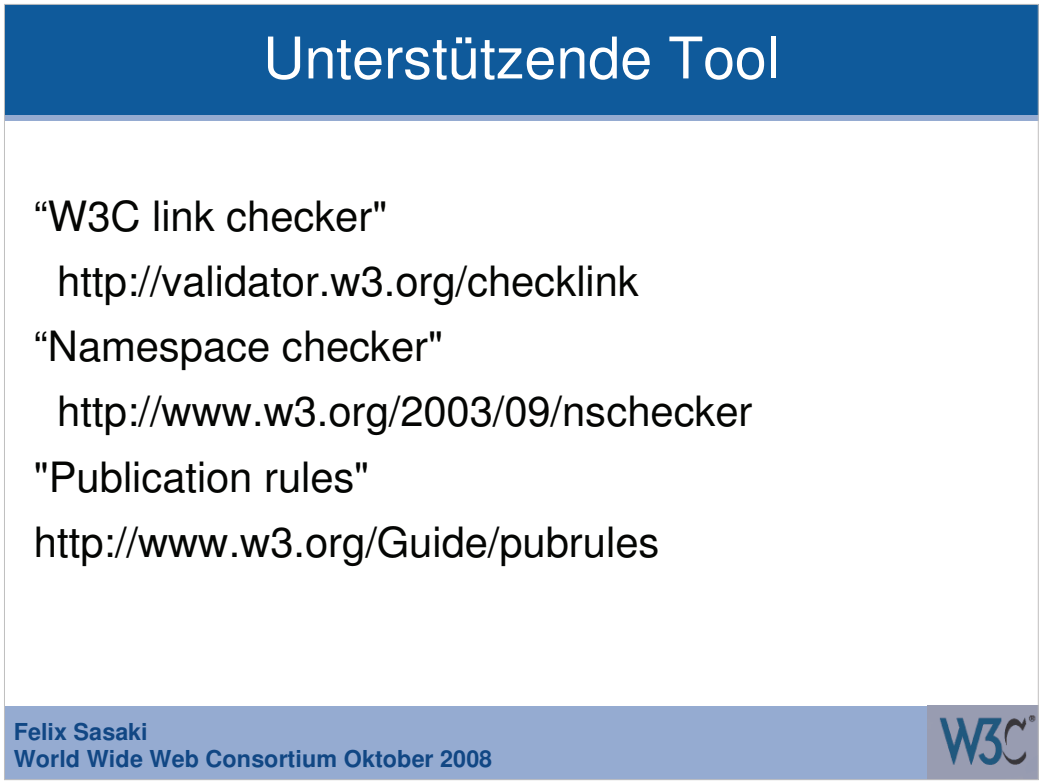
“Referencing electronic documents from W3C spec."
http://www.w3.org/2001/06/manual/#ref-section

“URIs for W3C Namespaces"
http://www.w3.org/2005/07/13-nsuri

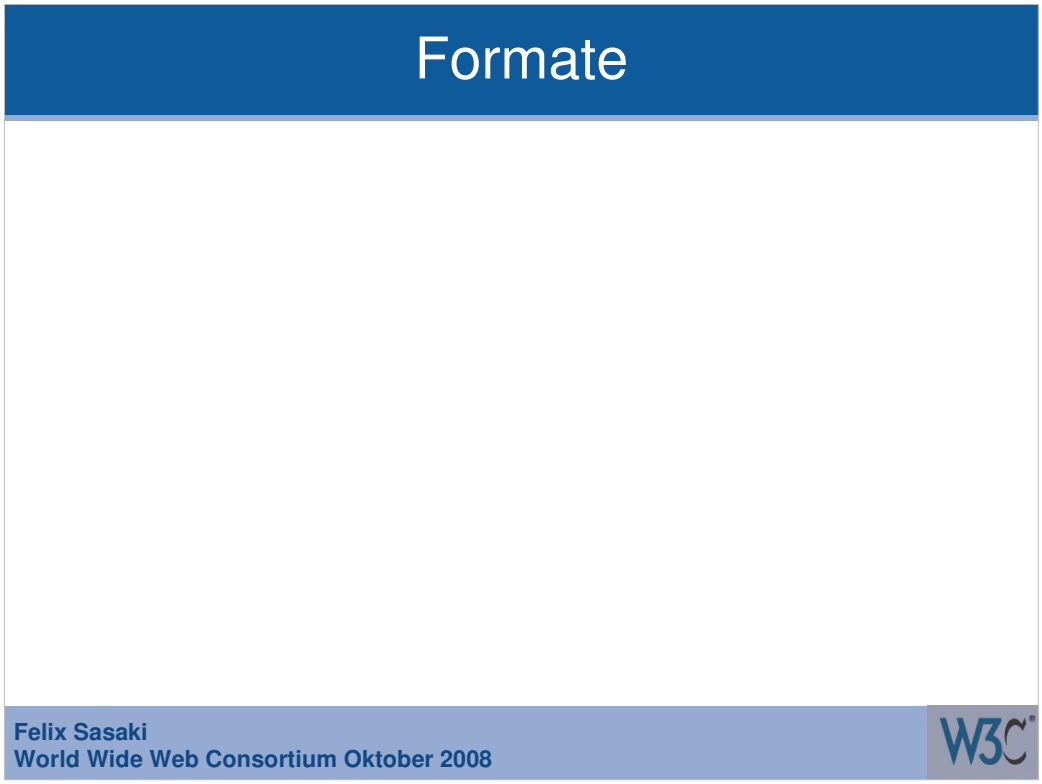
...

- Herausforderung “Community Consensus"

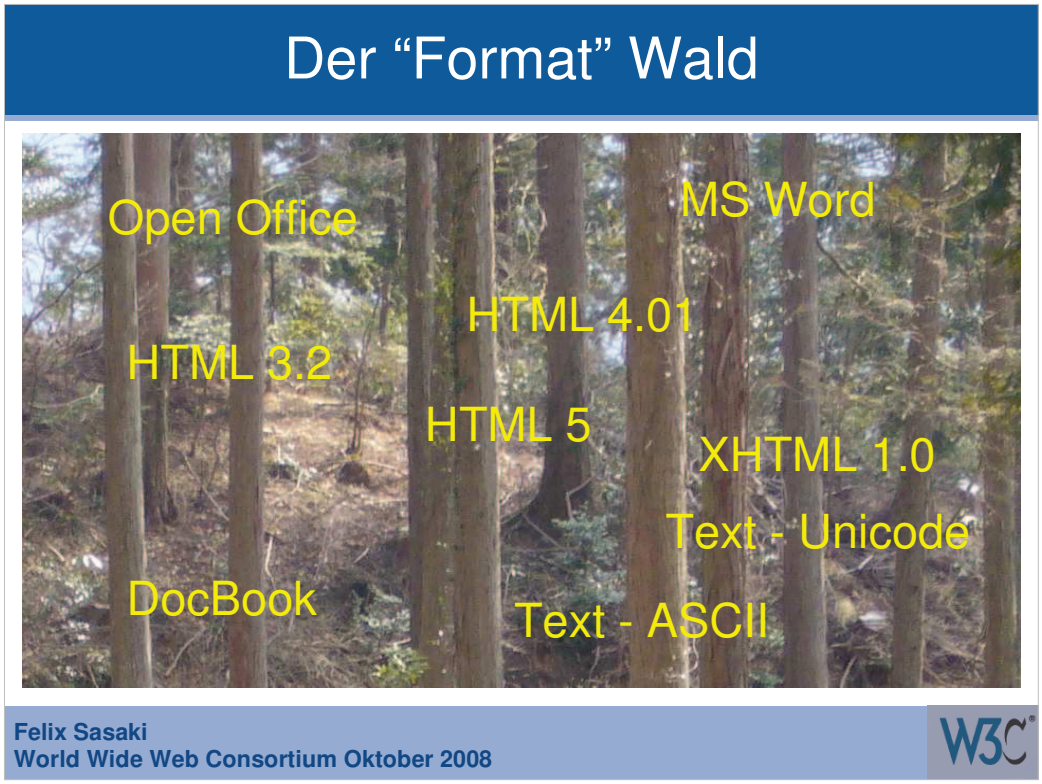
Das W3C hat für seine Community verschiedene Dokumente entwickelt, um diese soziale Bereitschaft zu fördern. Ich habe die wichtigsten auf dieser Folie aufgelistet. Die Herausforderung ist die Zustimmung, den "Consensus" der Community zu erlangen und zu bewahren.



Im Falle des W3C gibt es verschiedene Tools um die Nachhaltigkeit von URIs mit dem HTTP Schema zu überprüfen. Der "W3C link checker" überprüft HTTP Links in Webdokumenten. Der "Namespace checker" überprüft sichtbare HTTP Links. Die "Publication rules" schließlich umfassen diese beiden und weitere Tools um Nachhaltigkeit von Identifikatoren und andere Aspekte technischer Dokumente zu überprüfen.



Kommen wir schließlich zu den Formaten.

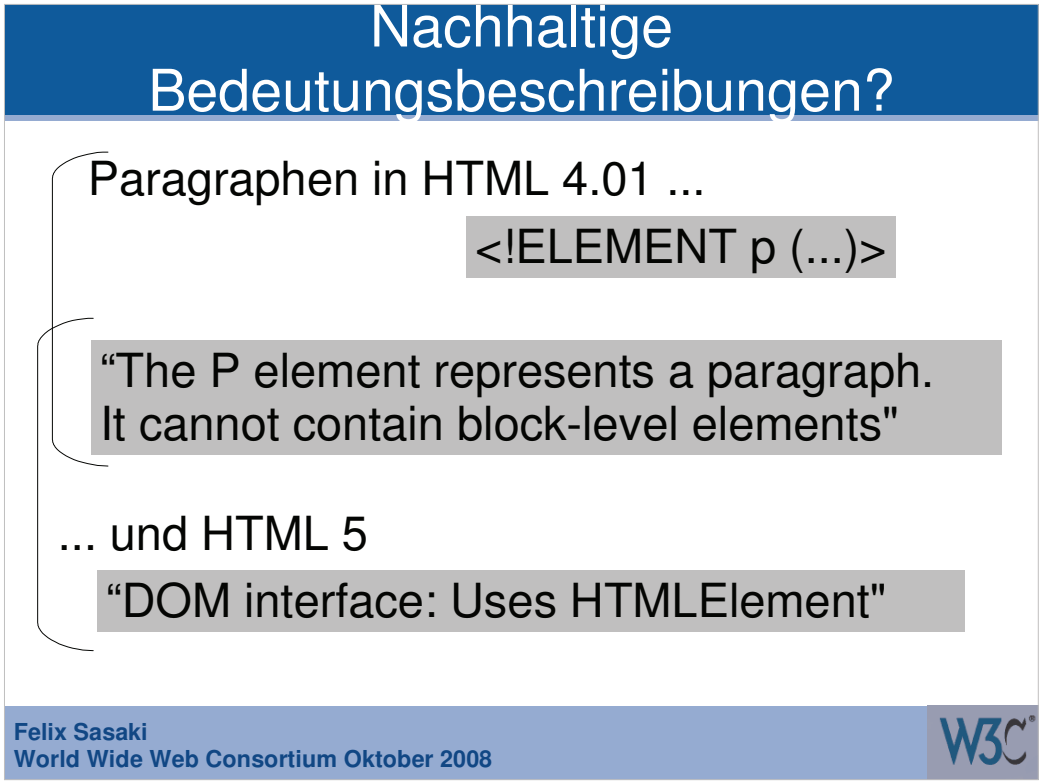


Wie bei den Services bzw. Protokollen und URI- / Identifikationsschemata gibt es auch bei Formaten einen Wald. Und er ist noch dichter als die beiden anderen. Selbst wenn man sich auf textuelle Dokumente beschränkt findet im heutigen Web eine Vielzahl von mehr oder weniger komplexen Formaten. Einige sind spezifisch für Browser (HTML 3.2, HTML 5, ...), andere für Textsorten wie technische Dokumentationen (DocBook). Wieder andere haben einen universalen Anspruch (Open Office, MS Word), oder sind „nur Text“ – und auch hier gibt es Varianten (ASCII, Unicode etc.).

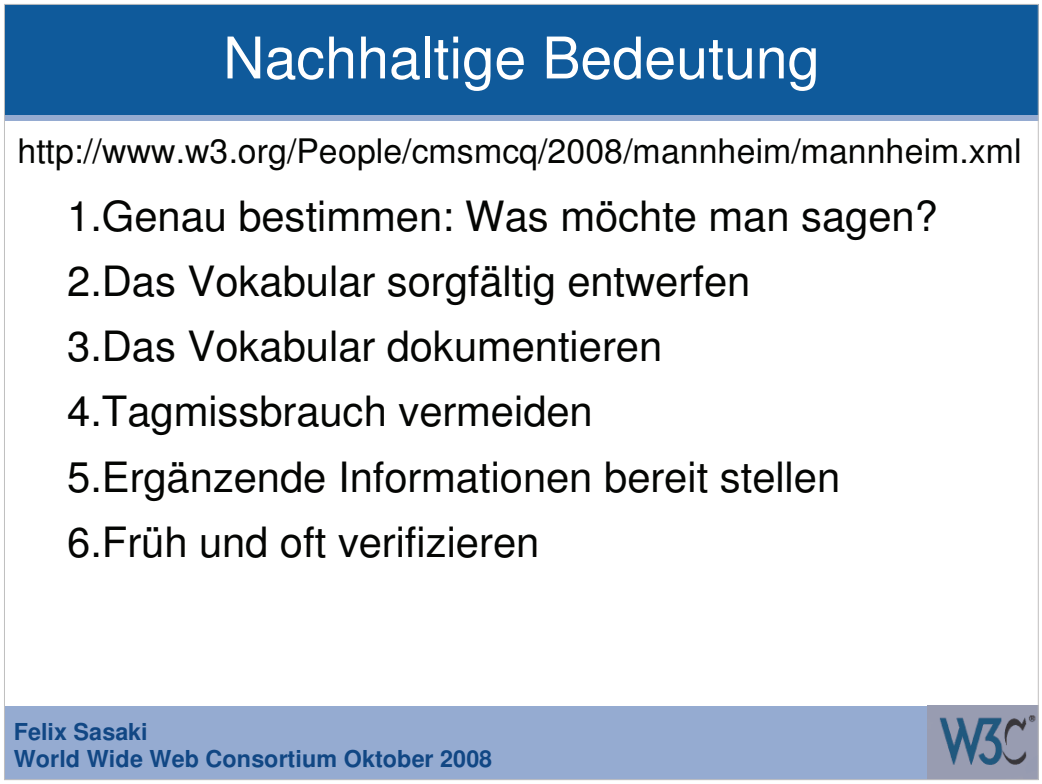
Die Lichtung:

Felix Sasaki
World Wide Web Consortium Oktober 2008

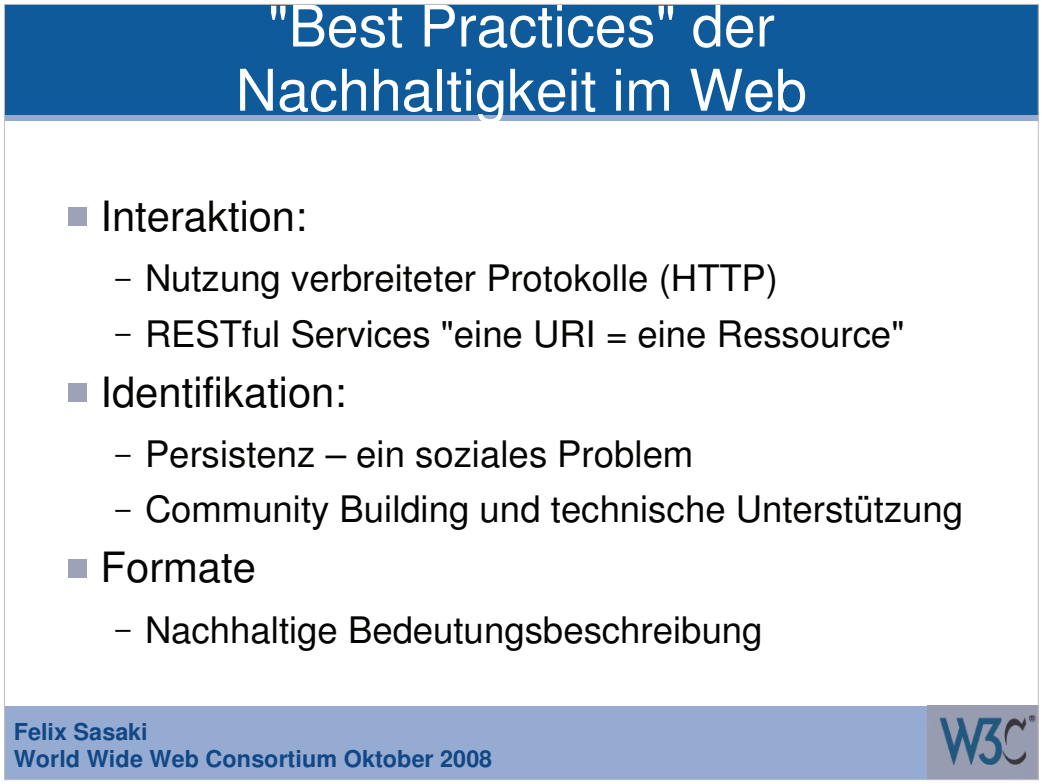
Das traurige am Formatwald ist dass eine Lösung des Problems der Nachhaltigkeit nicht in Sicht ist. Im folgenden sei der Grund hierfür am Beispiel des HTML Tags <p> demonstriert.



In der schon älteren, aber immer noch weit verbreiteten HTML Version 4.01 ist <p> wie auf der Folie dargestellt definiert:
<!ELEMENT p (...)>
Diese formale Definition bedeutet grob gesagt "ein <p> Tag besteht aus den Tags (...)".
In HTML 5 gibt es diese Definition nicht. Der Grund ist dass HTML 5 nicht die Struktur von <p> oder anderen Tags, sondern ihre Verarbeitung mit Programmiersprachen wie JavaScript ins Zentrum stellt. Für diese Verarbeitung ist die hier angegebene Information zum „DOM interface" wichtiger als der Inhalt von <p>.
Glücklicherweise haben beide Versionen von HTML eines gemeinsam: die Beschreibung des <p> Tags für den menschlichen Leser. HTML 4.01 umfasst jedoch noch weitere Dokumentation die insbesondere für das Editieren von HTML Dokumenten wichtig ist. HTML 5, dessen Standardisierung noch nicht abgeschlossen ist, wird diese möglicherweise in einer zukünftigen Fassung bekommen.



Dieses Beispiel zeigt wie schwierig es ist selbst bei einem, weit verbreiteten (Web)Format nachhaltige Bedeutungsbeschreibung zu erlangen. Auf Grund der fortgeschrittenen Zeit verweise ich hier nur auf die Arbeit meines Kollegen Michael Sperberg-McQueen, der kürzlich in einem Vortrag sechs Regel vorgeschlagen hat, die dem Problem wenigstens etwas Abhilfe schaffen können.



Zum Schluss noch ein Zusammenfassung von "Best Practices" der Nachhaltigkeit.

10000 Foot View

- Nachhaltigkeit von Webressourcen ist vor allem ein soziales Problem
- Nutzung von Protokollen, Identifikatoren und Formaten kann nachhaltig sein – muss aber nicht
- Wissen um Möglichkeiten und Grenzen der Technologie in der eigenen Community ist die beste Gewährleistung für Nachhaltigkeit

Dankeschön!

Felix Sasaki
World Wide Web Consortium Oktober 2008

Felix Sasaki
World Wide Web Consortium Oktober 2008

Formate

- Tag Soup
- Bedeutung – Was ist das?
- Existiert etwas wenn Google es nicht sieht?

Verwandte Ressourcen (Beispiel)

```
http://example.org/wetter-potsdam/

GET /wetter-potsdam/
Host: example.org
Accept: text/html
Accept-Language: de, en
```