7A-7

HTML 及び CSS における IRI テスト環境の提供

米川 和宏　伊藤 一成　Martin J. DÜRST
青山学院大学理工学部

1　はじめに

2005 年 1 月，Internationalized Resource Identifiers（IRI）が RFC 3987 [1] として公開された．IRI は，Web アドレスである URI の国際化のための標準化提案（Proposed Standard）である．URI には，US–ASCII の文字しか扱えないという制約がある [2]．一方，IRI は世界全体の文字を扱える Unicode / ISO 10646 への拡張により，様々な言語を用いたリソースの記述と識別を可能にする．今後，IRI の仕様の確立にむけて，多くのテストが必要とされている．テストの実施には，テスト項目の検討や，文字コード，言語の種類の組み合わせからなる膨大な量のファイルが必要である．そのため，テスト環境を用意するには大きな手間がかかる．そこで本稿では，IRI のテスト環境を容易に構築するための拡張性の高い枠組みを提供する．テストでは，IRI が指定するリソースを取り寄せる作業（Resolution）が正しく行われるかどうか確認する．また，枠組みを利用して，HTML 及び CSS における IRI のテスト環境を構築する．

2　URI における問題を解決する IRI の長所

URI を使用してリソースを識別している技術仕様は，非常に多岐にわたる．例えば，XML，RDF，XHTML，SVG，Java などが挙げられる．また，一般的にファイル名として日本語などが広く使用されているが，URI ではそれらのファイルをそのまま指定できない．IRI は，利用可能な文字の拡張によって，コンテンツ制作者やユーザ自身が普段利用する文字を資源の識別に扱えるようにする．

IRI はすべての URI を包含する．よって，Web 上の情報を閲覧する場合，既存の URI に何か特別な作業を施す必要はない．また，全ての IRI は UTF–8 を経由して URI に変換できる [1]．

2.1　変換によって生まれる URI の曖昧さ

世界中の自然言語において，基本ラテン文字にあたる A〜Z の文字だけが使用される言語はごく一部である．よって，URI を構成するためには多くの言語を US–ASCII の文字に変換する必要がある．

例として，“ほうじ茶” という日本語を，ラテン文字に変換する場合について考える．すると，“houjicya”，“hojicha”，“hohzitya” など表記方法が多数存在し，結果的に URI が曖昧になるという問題が生じる．IRI は扱える文字を拡張し，問題を解決する．

2.2　可読性の高い言葉で構成される IRI

URI では，使用不可能な文字に対応したオクテットを表現するために，パーセントエンコーディングを用いる．任意の URI を紙に書き写して，持ち運ぶ場面を考える．もし，オクテットが含まれている場合，変換に使われた文字コードやオクテットの意味を理解するのは不可能である．例えば，“ほうじ茶” という言葉を変換すると，以下のようになる可能性がある．

`%E3%81%BB%E3%81%86%E3%81%98%E8%8C%B6`（UTF–8）
`%82%D9%82%A4%82%B6%92%83`（shift_jis）

IRI は，URI より可読性の高い識別子を提供する．さらに，UTF–8 によるパーセントエンコーディングを定め，URI の変換に使用する文字コードを明確にする．

3　IRI テスト環境の構築のための枠組み

HTML における IRI のテスト環境を例に，提供する枠組みについて述べる．

3.1　枠組み構築にあたって考慮した点

1 点目は，プログラムによるテストファイルの作成である．HTML の場合，文字コードと URI や IRI が使用される要素や属性の組み合わせについて，テストファイルを用意する．作成すべきファイルの数は膨大である．また，内容の正確性が非常に重要である．テスト内容を細かく反映し，ファイルを素早く作成するには，プログラムの利用が最も効率的である．

2 点目は，枠組みの拡張性である．今後，様々な仕様や，別の文字コードへの対応が必要となる場合が十分に考えられる．その際，わずかなデータの追加や修正によって，新しいテストが作成できる枠組みを目指した．この工夫により，最初からテストファイルを作成するよりも，はるかに手間が省ける．

3 点目は，提供するテスト環境の不変性である．IRI のテストは，どのブラウザを使用した場合でも，同じ条件の下で実施できなければならない．例えば，一昔前のブラウザは，デフォルトの背景色が現在では一般

Providing an environment for testing IRIs in HTML and CSS
Kazuhiro YONEKAWA, Kazunari ITO
and Martin J. DÜRST
College of Science and Engineering, Aoyama Gakuin University
5–10–1 Fuchinobe, Sagamihara, Kanagawa 229-8558, Japan
yone@sw.it.aoyama.ac.jp, {kaz, duerst}@it.aoyama.ac.jp

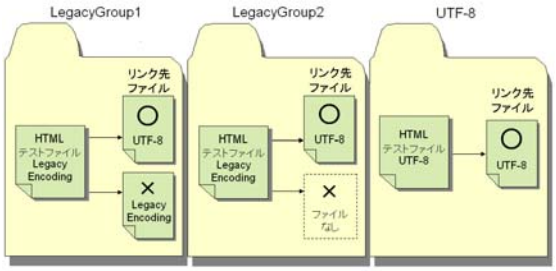

図 1: 出力ディレクトリの種類と内容

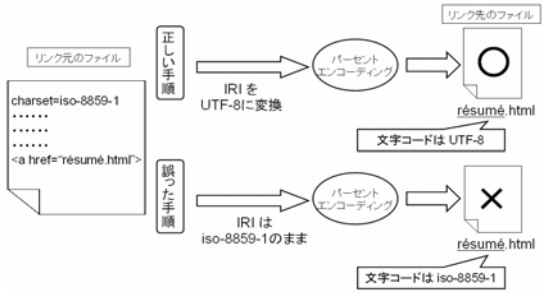

図 2: IRI を URI に変換する手順の例

的な白色ではなく灰色である．そのため，背景色を白，文字色を黒に指定する記述を CSS に追加し，環境を整える必要がある．また，テスト環境を Web 上で扱うために，Apache の設定ファイルを利用し，ブラウザがファイル内容の文字コードを間違えないようにする．

4 点目は，ファイル名が様々な文字コードからなるファイルを扱う環境である．Windows や Mac OS において，ファイル名は文字として扱われるため，任意の文字コードに指定できない．これに対して UNIX や Linux 環境では，ファイル名はバイト列として扱われるため，細かい設定が可能である．よって，本稿の枠組みを利用した IRI テストに適している．

3.2 出力されるディレクトリとファイル

プログラムを実行すると，URI や IRI を使用するそれぞれの要素と属性について，図 1 に示す 3 種類のディレクトリが出力される．さらに，その中に文字コード別のディレクトリ，HTML ファイルなどのテストファイルが階層的に出力される．それぞれの HTML ファイルには，文字コードの種類とその文字コード特有の文字を含む単語が表示される．HTML ファイルの中身を確認せずに，目視で文字化けが起こっているか否か確認するためである．例えば，iso–8859–1 の場合は，résumé という単語が該当する．それぞれのディレクトリの違いを以下に示す．

LegacyGroup1 では，レガシーエンコーディングを使用して IRI をテストする．レガシーエンコーディングとは，国や地域ごとに使い分ける必要がある文字コードを総称した呼び名である．IRI を URI に変換する手順の例と，テスト結果の対応を図 2 に示す．IRI は，パーセントエンコーディングを行う前に，UTF–8 に変換しなければならない [1]．しかし，この変換が実施されず，リンク元のファイルで使用された文字コードのままでパーセントエンコーディングが行われると，IRI で指定するリソースを正しく取り寄せられない．また，一般的にリンクを作成する場合，リンク元の Web アドレスに対して，リンク先のファイルは 1 つしか存在しない．しかし，IRI をテストするには，リンク先のファイルとして，ファイル名がレガシーエンコーディング及び UTF–8 で記述された両方のファイルを使用する．テストの失敗を視覚的に確認するためである．

LegacyGroup2 は，機械がリンクチェックを実施するために用意してある．LegacyGroup1 では，テストの結果に関わらず，必ずファイルが返される．しかし，ファイルの内容を確認できる人間と異なり，機械はファイルが成功を示すか否か理解できない．よって，ファイル名が UTF–8 で記述されたファイルのみリンク先として用意してある．成功の場合のみファイルを返し，機械にテスト結果を知らせるためである．

出力ディレクトリ UTF–8 には，ファイル名や内容が全て UTF–8 で記述されたファイルが出力される．リンク元のファイルが UTF–8 で記述されている場合，パーセントエンコーディングに使われる文字コードも UTF–8 である．そのため，テストは成功するはずである．しかし，何らかの理由で 8 ビット目が欠けている場合，テストは失敗する．

4 まとめと今後の課題

本稿では，IRI のテスト環境を構築するための拡張性の高い枠組みを提供し，実装を行った．今後，様々な技術仕様に対して，さらに多くのテストが必要とされている [3]．今後の課題は，多くの技術仕様に対するテストの実施，及びそれに伴う IRI 仕様の確立である．

HTML 及び CSS における IRI のテスト環境は，以下の Web ページで公開している．ぜひ参照していただきたい．

http://www.sw.it.aoyama.ac.jp/2005/iritest/

参考文献

[1] Martin J. Dürst, M. Suignard, Internationalized Resource Identifiers (IRIs), RFC 3987 (January 2005)

[2] T. Berners-Lee, R. Fielding, L. Masinter, Uniform Resource Identifier (URI): Generic Syntax, RFC 3986, STD 66 (January 2005)

[3] Martin J. Dürst, François Yergeau, Richard Ishida, Misha Wolf, Tex Texin, Character Model for the World Wide Web 1.0: Resource Identifiers, W3C Candidate Recommendation (November 2004) http://www.w3.org/TR/2004/CR-charmod-resid-20041122/