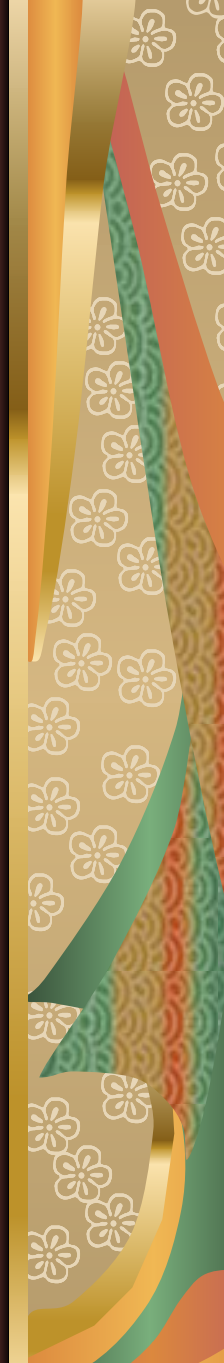
NewsMLでの外字表現を考える

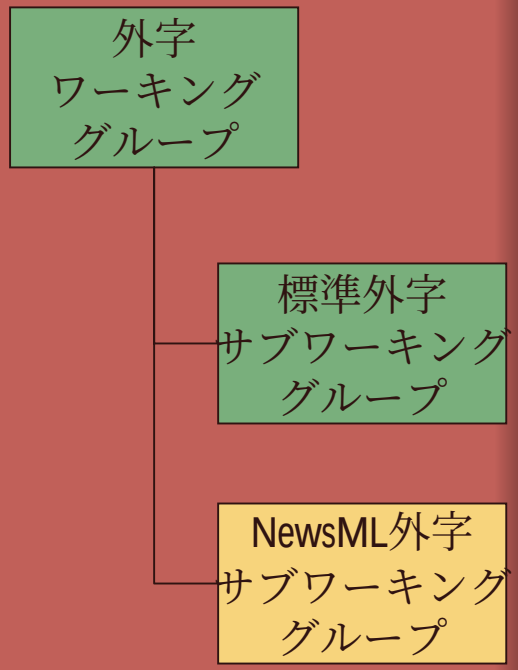
NewsML外字
サブワーキング・グループ

日本アイ･ビー･エム株式会社
ソフトウェア事業部
藤原　隆弘
fujiwat@jp.ibm.com



はじめに

- 外字ワーキング･グループの構成
  - 標準外字サブワーキング･グループ
    - 他からの参照モデルになるようなXMLの外字表現を提唱する
  - NewsML外字サブワーキンググループ
    - NewsMLで流通するデータでの現実的な外字の解決方法を提案する



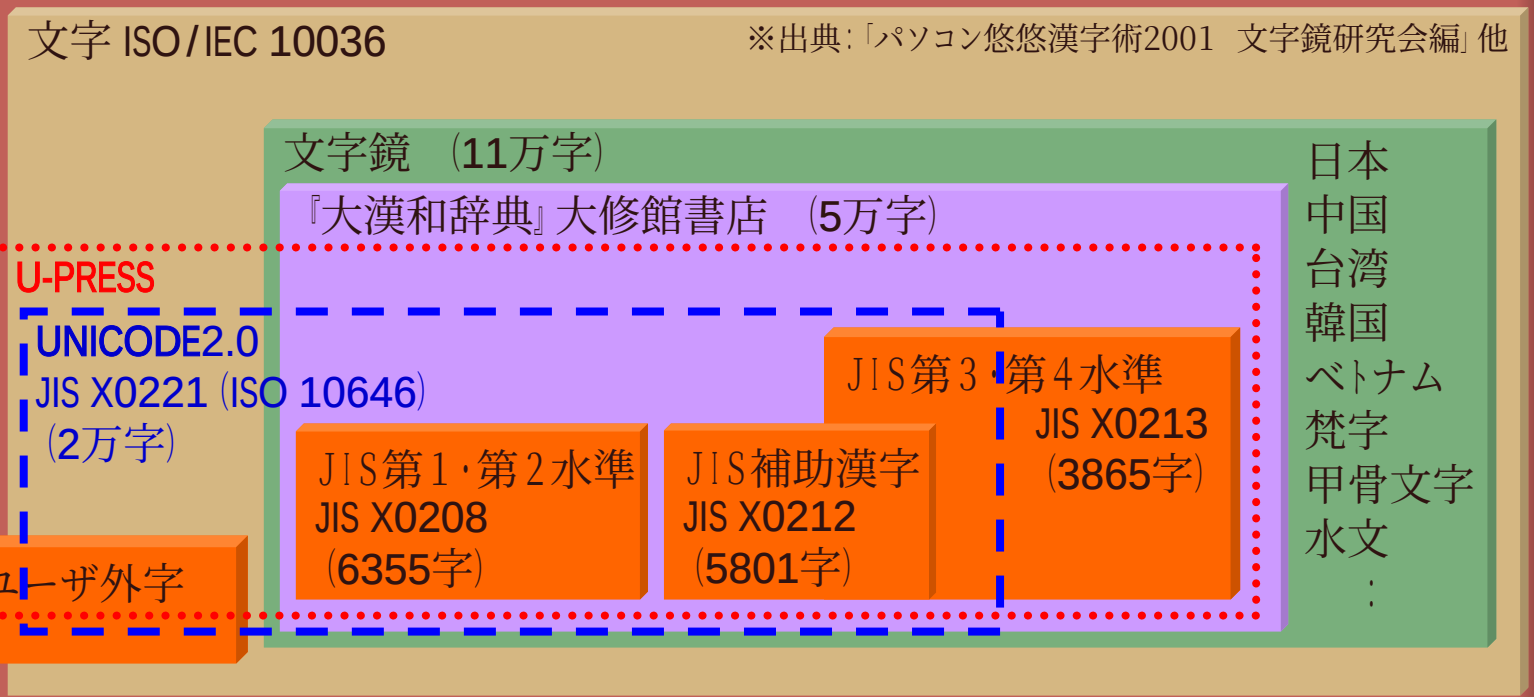
目的

- NewsMLでの外字表現方法を日本新聞協会へ提案
  - NewsMLは世界標準のニュース用のXMLフォーマット
- フォント環境ではなく、流通データの外字表現形式をターゲットとする
- 共同通信社の文字コード体系（U-PRESS）の文字も表現できるように考慮する
  - 日本新聞協会へ提案するために重要

概要説明
- 外字の定義 -

- ここで扱う外字とは、JIS第1・第2水準 (X0208:1997) 以外の文字で共通に使われる文字とする (後述)
  - 個人で作成する文字 (ユーザー外字) は今回の範囲外とする

下図は概念的なものであり、詳細は異なるものがあります



概要説明
- ISO/IEC 10036に準拠 -

- ISO/IEC 10036は国際規格
- 文字鏡の文字はすべてISO/IEC 10036に登録されている
- 文字鏡の範囲内ならばフォントが存在するので実用可能
  - 現在はISO/IEC 10036のグリフ番号下位7桁が文字鏡番号
  - 先頭2桁最初の10を文字鏡がリザーブしている。
    例： “ISO/IEC 10036/RA//Glyphs:100058562”

詳細については...

ISO/IEC10036 に関しては「日本規格協会」http://www.jsa.or.jp/
ISO/IEC 10036 Second edition 1996007-15, Information technology - Font information interchange - Procedures for registration of font-related identifiers
ISO/IEC 10180 font reference font specifications
http://www.net.intap.or.jp/oiia/cont1/index.html などを参照

概要説明
- 確認用として文字鏡を採用 -

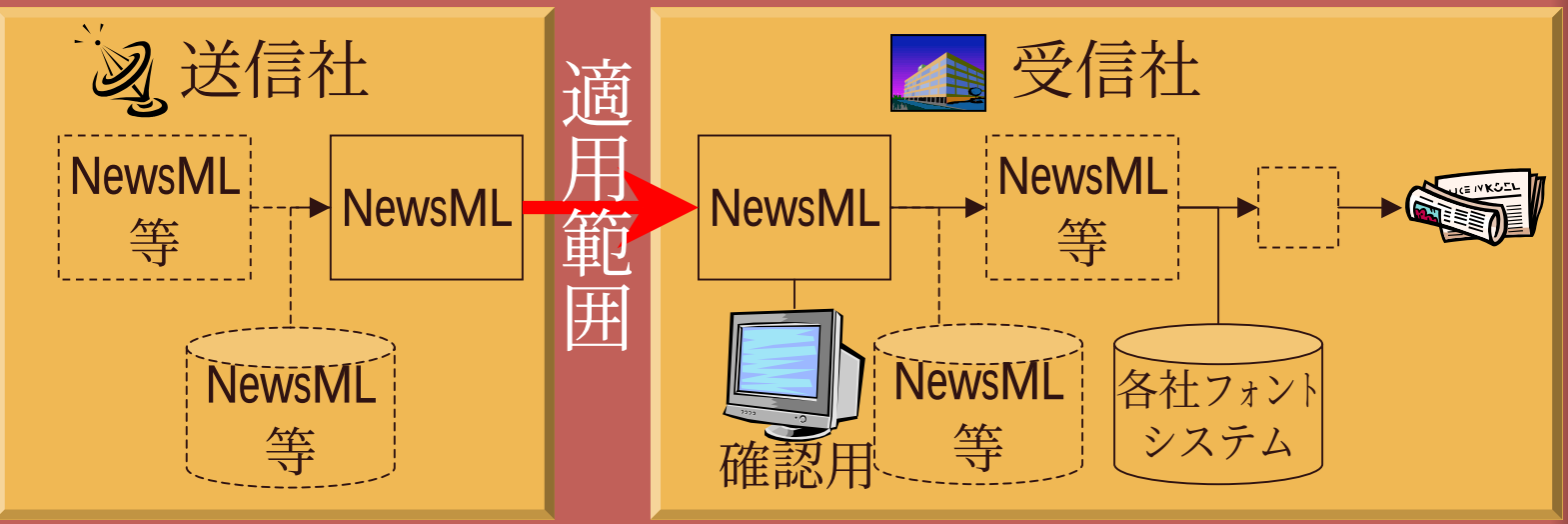
- 文字鏡のTrueTypeフォントは無料で配布している
  - 印刷フォントは各新聞社の既存システムを利用し、確認表示だけを文字鏡フォントで行う用途には最適
  - インターネットに接続しない環境でも利用可能
- 不足している外字を、文字鏡に登録する手段がある
  - 登録された文字鏡フォントを使えば専用システム以外でも表示可能
    - 野球を表す記号　ニュースを表す記号　など

詳細については...

文字鏡に関しては「パソコン悠悠漢字術2001　文字鏡研究会編」
http://www.mojikyo.com/ を参照

概要説明
- 適用範囲 -

- NewsMLでの外字表現方法の提案範囲は、NewsMLのファイルが社間で流通する部分とする
- 各社で印刷用に使うフォントは範囲外
  - 実際は各社のフォントシステムと接続される
- スタイルシートによる表示は確認用



NewsMLで外字表現をするにあたって

- NewsMLは
  - DataContent内は別にDTDを持つ
- 日本新聞協会では
  - DataContent内は送信側、受信側の調整で決める
  - DataContent内のDTDをFormatとして申請する
  - 基本となるテキスト記述用DTD（NskBasicText.dtd）を公開している

詳細については...

DataContent, NewsLines, NskBasicText.dtdは
日本新聞協会発行のNskNewsML解説書を参照
http://www.pressnet.or.jp/newsml/newsml.htm

NewsML外字サブWGの活動計画

- NewsMLでの外字表現方法を検討
  - XMLでの外字表現に詳しい人も参画
- 日本新聞協会へ検討結果を提案
  - DTDはフォーマットとして申請
  - 外字表現提案の説明会を実施
  - フォントのインストールと実演

- 不足している外字対象文字を文字鏡へ登録

日本新聞協会への提案に必要なステップ(1)

- NewsML内の外字記述方法を検討
  - DataContentに含める外字の表現方法

（以下は例）

<○○ xpr:name="ISO/IEC 10036/RA//Glyphs:100058562">置き換え文字</○○>

置き換え文字は通常1文字

  - 検討事項：
    NewsMLではXMLスキーマ、NameSpaceを使用しないが、どうするか
- DataContent用のDTDを作成
  - 日本新聞協会のNskBasicText.dtdを拡張

詳細については...

DataContent, NewsLines, NskBasicText.dtdは
日本新聞協会発行のNskNewsML解説書を参照
http://www.pressnet.or.jp/newsml/newsml.htm

日本新聞協会への提案に必要なステップ(2)

- NewsLines内での外字記述方法を検討
  - NewsLines内はDTDが固定であり、Origin要素のみ利用可能

(以下は例)

```
<NewsLines>
 <HeadLine>ヘッド<Origin Href="....">置き換えA</Origin>ライン</HeadLine>
 <SubHeadLine>……</SubHeadLine>
 …
```

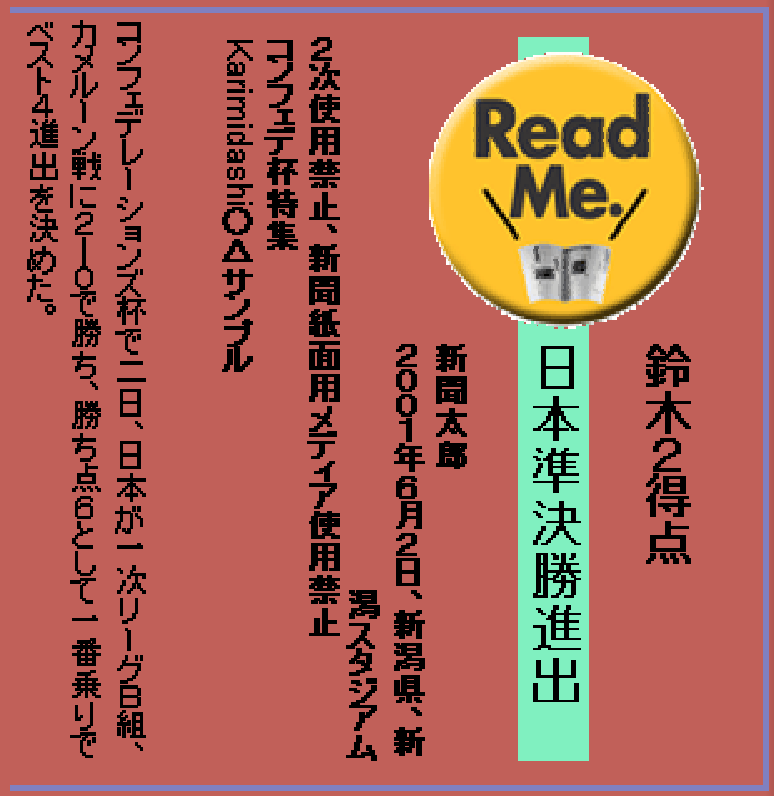
  - 上記、HeadLineの文字は「ヘッド置き換えAライン」であり簡単に取り出すことができる。
  - 「置き換えA」にはOriginで詳細説明がある。詳細説明はHref=がポイントする先にある。

詳細については...

NewsLines, Originは
日本新聞協会発行のNskNewsML解説書を参照
http://www.pressnet.or.jp/newsml/newsml.htm

日本新聞協会への提案に必要なステップ(3)

- スタイルシートの作成(以下は予定)
  - 外字文字に色をつけて通常の文字と区別して表示できるスタイルシートを作成
  - 原型となるNewsML表示スタイルシートがある



原型に考えているスタイルシートでの表示

なぜ、ＪＩＳ第１・第２水準(X0208：1997)以外の文字を外字と扱うのか？

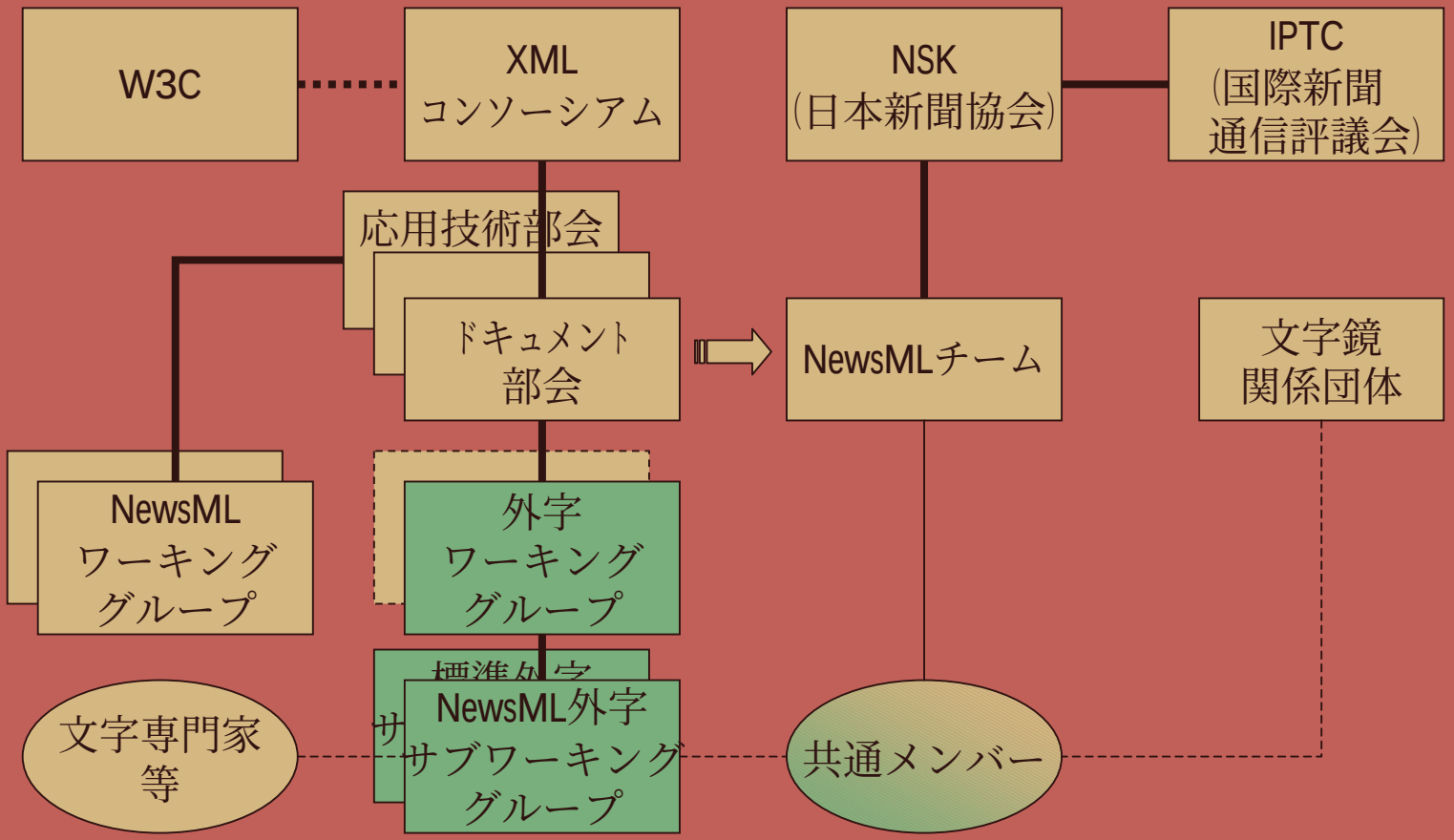
- JIS第1・第2水準文字以外を外字とする

  ＝置き換え文字はJIS第1・第2水準文字

- 検索条件にJIS第1・第2水準文字を指定して、外字の文字列を検索させることができる
  - 例：齋藤さんの齋も…
    「斉藤」で検索可能になる


```
<DataContent>
 <body xmlns:xpr="http://www.xml.gr.jp/PRE/Reference">
  …
  <○○ xpr:name="ISO/IEC 10036/RA//Glyphs:100058562" >斉< /○○ >藤
```

(参考)
関連組織・団体


W3C
XML
コンソーシアム
NSK
(日本新聞協会)
IPTC
(国際新聞
通信評議会)
応用技術部会
ドキュメント
部会
NewsMLチーム
文字鏡
関係団体
NewsML
ワーキング
グループ
外字
ワーキング
グループ
NewsML外字
サブワーキング
グループ
文字専門家
等
共通メンバー

(参考)
日本新聞協会(NSK)

- 新聞倫理綱領を制定し実践する自主組織1946年7月に創立した社団法人
- 新聞112、通信4、放送39（ラジオ単営8、テレビ単営25、ラ・テ兼営6）計155社
- 各種フォーマット策定のためベンダー参加
- 各社の代表者で構成する総会、理事会のもとに、各種の委員会、専門部会が設置

詳細については...

日本新聞協会WEBサイト
http://www.pressnet.or.jp/

(参考)
NewsMLチーム

- 日本新聞協会内で2000年10月に発足した、NewsMLを日本で使っていくために必要な事項を検討する集まり
- 新聞協会加盟新聞・通信・放送26社46人、外国通信社・メーカー・ベンダー・プロバイダー等18社29人

詳細については...

日本新聞協会NewsML関連WEBサイト

http://www.pressnet.or.jp/newsml/newsml.htm

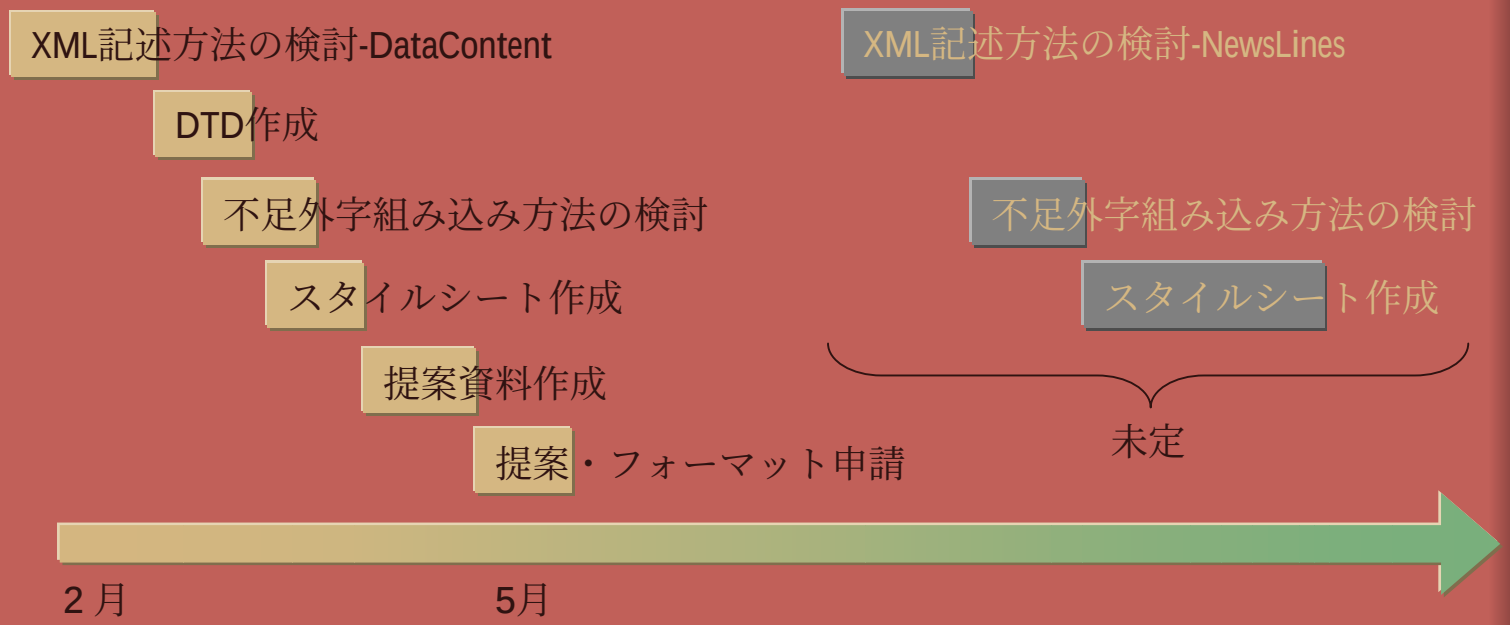
(参考)
NewsML

- XMLを用いた世界標準のニュース管理フォーマット
- IPTCが2000年10月に発表
  - IPTC:（International Press Telecommunication Council）
    国際新聞通信評議会
  - 世界約55社の通信社･新聞社･ベンダーが標準化と開発に参加（ニューヨークタイムズ、ＡＰ、共同通信など）

詳細については...

IPTC　http://www.iptc.org/
NewsML　http://www.newsml.org/

全体スケジュール案



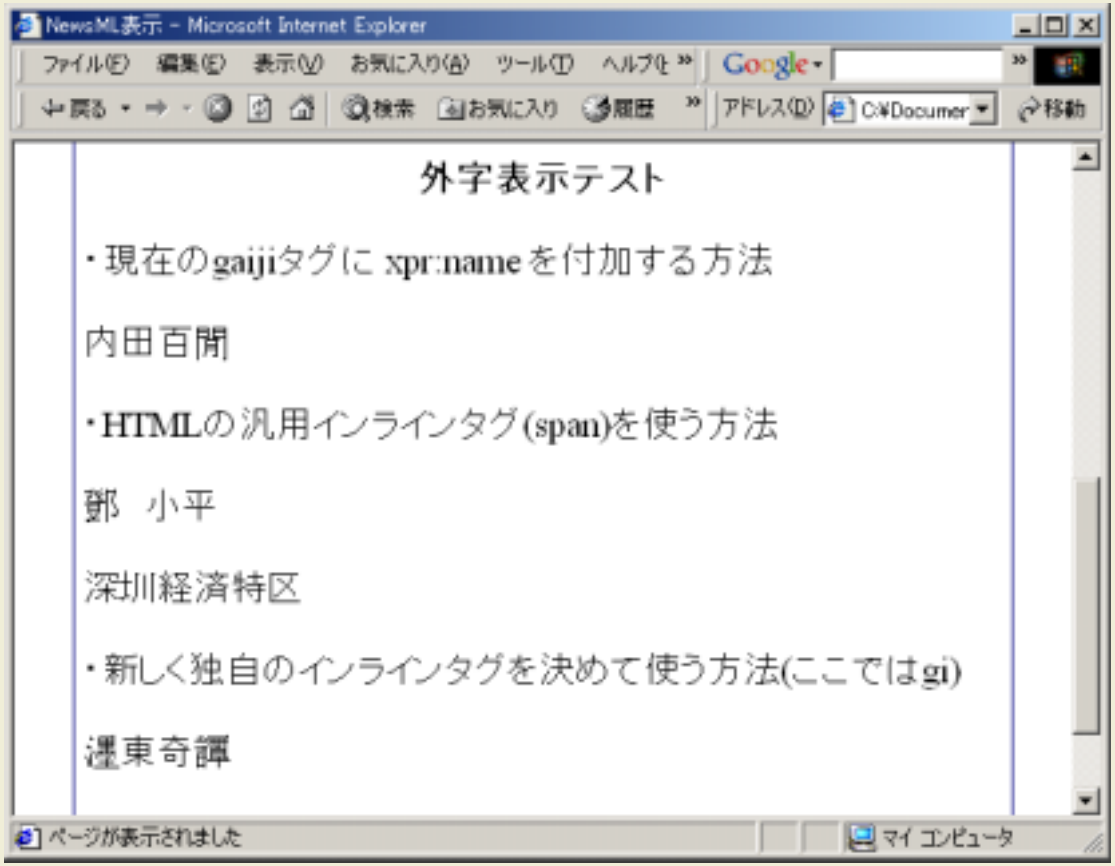
詳細については...

今後ドキュメント部会・NewsML外字ワーキンググループで決定する

主なイベント

- 01年12月　NSKNewsMLチーム有志で文字鏡の調査
- 02年1月　ピーデー川俣氏、イースト渋谷氏と打ち合わせ
- 2月4日　日本新聞協会で経過報告予定
- 2月6日　XMLコンソーシアム成果発表会で経過報告
- 2月7日　Page2002併設セミナー：
  XMLコンソーシアム・ドキュメント部会で
  経過報告、今後の予定検討

(デモ)
外字表示テスト



- スタイルシートの作成はイースト株式会社の渋谷氏に全面的に協力いただきました。
  http://www.est.co.jp

(デモ解説)　内部案1.
gaijiタグ案にxpr:name追加

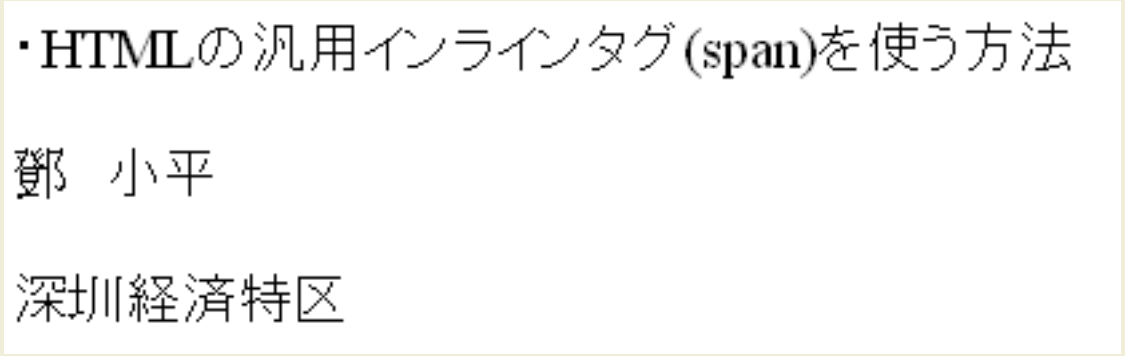
```
…
<DataContent>
    …
    <p>・gaijiタグ案に xpr:name を付加する方法</p>
    <p>内田百
        <gaiji xpr:name="ISO/IEC 10036/RA//Glyphs:100077620">
            <gaiji.alt>〓</gaiji.alt>
            <gp>
                <sgp> (</sgp>
                <yomi>ケン</yomi>
                <gs>:</gs>
                <jikai>門がまえに月</jikai>
                <egp>) </egp>
            </gp>
        </gaiji>
    </p>
    …
```

・現在のgaijiタグに xpr:name を付加する方法

内田百閒

(デモ解説)　内部案2.
ＨＴＭＬ汎用タグにxpr:name追加

```
…
<DataContent>
    …
    <p>・HTMLの汎用インラインタグ(span)を使う方法</p>
    <p>
        <span xpr:name="ISO/IEC 10036/RA//Glyphs:100039630">トウ</span>
        小平</p>
    <p>深
        <span xpr:name="ISO/IEC 10036/RA//Glyphs:100004894">セン</span>
        経済特区
    </p>
    …
```



(デモ解説)　内部案3.
ＨＴＭＬ汎用タグにxpr:name追加

```
…
<DataContent>
    …
    <p>・新しく独自のインラインタグを決めて使う方法(ここではgi)</p>
    <p>
        <gi xpr:name="ISO/IEC 10036/RA//Glyphs:100050021"/>
        東奇
        <gi xpr:name="ISO/IEC 10036/RA//Glyphs:100035978"/>
    </p>
```

・新しく独自のインラインタグを決めて使う方法(ここではgi)

澫東奇譚

(デモ解説)
スタイルシート

```
<xsl:template match="*[@xpr:name]">
```

- 上記ですべての要素の属性で xpr:name がある場合の処理を記述可能

予想質問

- 置き換え文字の決まりはあるのですか？
- 今後追加する文字鏡番号もISO/IEC 10036の下位6桁が同じになるのですか？