マルチモーダル記述仕様の策定とメディア生成への応用

伊藤　一成†　　曽我　真也††　　斎藤　博昭†

† 慶應義塾大学 大学院理工学研究科
†† 慶應義塾大学　〒 223–8522 神奈川県横浜市港北区日吉 3–14–1
E-mail: {k_ito,soga,hxs}@nak.ics.keio.ac.jp

あらまし　本稿は，複数のモーダル情報を一括して記述可能なメタデータ仕様を提案する．複数のモーダル情報を融合するために，統語情報を記述する GDA（Global Document Annotation）の属性を拡張した．この仕様により，各種モーダル情報の相互の対応付けが困難であるという既存の問題が解決される．また，適用事例として，提案したメタデータには談話や動作のモーダル情報が含まれているため，このメタデータからモーダル情報を加味したメディアコンテンツを生成するプロセスについて概説する．
キーワード　マルチモーダル，メタデータ，メディア生成

A Proposal for Multimodal Description Specification and Its Application to Media Generation.

Kazunari ITO†, Masaya SOGA††, and Hiroaki SAITO†

† Department of science and Technology, Keio University
†† Faculty of Science and Technology, Keio University
Hiyoshi 3–14–1, Kouhoku–ku, Yokohama, 223–8522 Japan
E-mail: {k_ito,soga,hxs}@nak.ics.keio.ac.jp

Abstract　This paper proposes a metadata specification that can describe multimodal information collectively. To unite the description about multi-modalities, we expand the attribute of the GDA（Global Document Annotation） tagset which mainly describes syntactic information. This specification enables explicit linkings of various modalities. As an application, we show the process of generating media contents from the proposed metadata which contain dialog and gesture modalities.
Key words　multimodal, metadata, media generation

1. はじめに

テレビドラマ，ニュースなどの動画・音声データはエンターテイメントや教育などにまつわる用途をそれ自体として持っているが，これに映像情報や意味内容に基づくインタラクティブな検索や掲示が出来るようになれば，実用的な価値を更に高めることができる．近年これらマルチメディア情報を効率よく検索したり要約する手段として，メタデータ技術が注目を集めている．これは，マルチメディアコンテンツからその内容に関する特徴を予め記述しておき，記述データを直接の処理対象とすることでマルチメディアコンテンツの処理を代替しようというものである．ここで，記述データはメタデータと呼ばれる．

マルチメディアコンテンツには，身振り，表情，音声など多種多様なモーダルを含み，それらは韻律，統語，談

話，社会的な関係等，多様な情報を内包している．それぞれは固有の分析単位をもち，記述方法も異なり，相互の対応付けが困難である．

また，マルチメディアコンテンツの意味内容に基づくメタデータの記述仕様として，MPEG-7 [1] が挙げられる．MPEG-7 はマルチメディアコンテンツの音声や映像に関するメタデータ仕様であるが，全てのモーダル情報に関してではなく，韻律情報など記述できない情報も存在する．

そこで，本稿では複数のモーダルに関する情報の統合を目的とした，メタデータ記述仕様について概説する．この記述仕様は，GDA（Global Document Annotation）[2] をベースに拡張し，さらにアノテーション汎用記述言語　MAML（Multimedia Annotation Markup Language）[3] を併用することによって，情報を記述する．

また，提案するメタデータ仕様は多くのモーダル情報を含んでいるため，様々な適用事例が考えられる．本稿では，その一例として，今回策定したメタデータを TVML（TV Program Making Language）[4] 形式に変換することでモーダル情報を含んだ動画生成を行なうプロセスについて概説する．

2. モーダル情報表現に関する既存の研究

2.1 韻　　律

発話の韻律は，発話内容に関する発話者の感情や内容の真意を同定する上で非常に重要な要素である．日本語の韻律ラベリングシステムである J_ToBI（Japanese Tones and Break Indices）[5] は，英語用の ToBI の原則に基づくもので，これによって日本語のイントネーションを表すことができる．J_ToBI にはローマ字表記により発話された言葉を単語ごとに区切った単語情報を記述する word 層，イントネーション情報を記述する tone 層がある．tone 層では音の高低を H%，L%等と記述する．また，アクセント辞書と話者による発話のアクセント型が異なる場合が存在する．このときは word 層にはアクセント辞書を用いたアクセント型に基づくものを記述し，発話アクセントは tone 層に記述する．図 1 に記述例を示す．

2.2 談　　話

会話や文章において，談話構造は，その内容や流れの理解を助ける役割を担う．談話情報を表すものとして，DAMSL [6] が有名である．表 1 に DAMSL のタグセットの一部を示す．また，DAMSL を用いたラベリング例を図 2 に示す．

表 1 において，Communicative Status はコミュニケーション（言葉の伝達）のレベル，Information Level は発話内容のレベル，Forward-Looking Function はそれ以降の対話に影響を及ぼす発話，Backward-Looking Function は先行対話に関連した発話を示す．

図 2 において発話はスラッシュユニットという単位で区切られており、スラッシュユニットごとに表 1 のタグセットを一つ記述する。

2.3 動　　作

人の動作は発話と別のコミュニケーション手段をなりうる．Anvil（Annotation of Video and Language）[7] は，人の様々な動作を転記することができるタグセットを定義している．Anvil の記述形式は XML（eXtensible Markup Language）[8] に基づいている．`attribute` タグのテキスト部に動作，位置，形などの値をそれぞれ記述することにより，細かい動作の指定が可能となる．ラ

```
    4.089916   -1  %L
    4.422615   -1  H-
    4.741055   -1  H*+L
    4.954933   -1  L%
    6.623179   -1  %L

   4.957513   -1 e-Qto
   6.894107   -1 kono'
   7.079452   -1 hito
   7.469185   -1 wa-
   8.619377   -1 kuro'me
```

図 1　J_ToBI のラベリング例（上:tone 層　下:word 層）

表 1　DAMSL の種類（抜粋）

クラス	種類	説明
Communicative Status	un	解釈不可能な発話
	aba	発話途中で放棄された発話
	st	独り言
Infomation Level	tsk	課題を遂行している発話
	tm	課題 mp 遂行に関するメタ的発話
	ol	その他（冗談，雑談等）
Forward-Looking Function	stt	主張を問える内容の発話
	asr	聞き手の信念の変更を意図した発話
	ras	言明の繰り返し
	ir	情報の要求
	ir(yn)	YES/NO 型疑問
	ir(wh)	WH 型の情報要求
	csf	発話者の将来の行動を拘束する発話
Backward-Looking Function	agr	承諾のタイプ
	acc	あきらかな承諾
	rr	言明の繰り返し
	ap	部分的な承諾
	mb	承諾の保留
	rp	部分的な拒否
	rj	拒否

```
A:
%
B:うん/<4.178>     ほかは/<1.855>
%  ack                ir(wh)

A:ほかはね<4.010>/  前髪が<0.123>掛ってない<1.926>/
%     rr                                 asr
B:          うん<6.350>/
%            ack
```

図 2　DAMSL のラベリング例

```
<attribute name="phase">
    stroke</attribute>
<attribute name="location-side">
    right</attribute>
<attribute name="handedness">
    right</attribute>
```
図 3　動作に関するラベリング例

ベリングの例を図 3 に示す.

2.4　表　　情

また，動作に限らす，特に顔の表情は多くの情報量を相手に提供できるモーダルである．FACS（Facial Action Coding System）は，顔の表情に関する分類規則を定めたものである．FACS は表情変化における顔の動きを 44 種の AU（Action Unit）に分解し，表情生成における基本動作とすることで，この組合せによりあらゆる表情を合成することができるとしている．今回使用する顔の表情は，FACS の基本 6 表情である驚き（surprise），恐怖（fear），嫌悪（disgust），怒り（anger），喜び（happiness），悲しみ（sadness）である．

3.　マルチモーダル記述仕様の策定

本章では，複数のモーダル情報を記述するメタデータの設計，及び記述仕様について述べる．

3.1　GDA の拡張

個々のモーダルの情報のタイムスタンプを基に，一つのタイムライン上に列挙する極めて単純な方法がまず思いつく．しかしながら，それでは，それぞれのモーダルの情報の包括関係が分かりにくい．そこで，タイムスタンプではなく発話テキストそのものを基軸として，音声に関するモーダル情報を融合する．

具体的には以下のとおりである．多言語間に共通の統語・意味に関する XML タグの標準を作成し普及させようというプロジェクトとして GDA がある．GDA タグは，情報として，係り受け，代名詞の指示対象，多義語の意味などの詳細情報まで定義されており，文全体の修辞構造による要約などが可能になってくる．

図 4 に GDA の記述例を示す．su タグは文を示す．文 n，np タグは名詞および名詞を主辞とする語句を表す．n タグは他の語句の係り受け対象となることができ，np タグはならない．他についても同様である．v タグ，vp タグは，終助詞以外の助詞，副詞，連体詞，接続詞，およびこれらの投射を表す．

前章で解説したモーダルの内，韻律や談話のモーダル情報を含めるように GDA を拡張する方式を採用する．GDA タグを基にすることにより，GDA タグが持つ統語

```
<su syn="fc">
 <n>解答</n>
 <v agt="B" obj="X">し</v>
 <v>ます</v>
</su> <su syn="fc">
```
図 4　GDA の記述例

```
<su syn="fc" sem="com">
 <n prn="kaito-" tone="%L H-"
       begin="68.349" end="68.792">
  解答
 </n>
 <v agt="B" obj="X" prn="sh#i" tone=""
       begin="68.792" end="68.946">
  し
 </v>
 <v prn="ma'su" tone="L%"
       begin="68.946" end="69.255">
  ます
 </v>
</su>
```
図 5　拡張 GDA の記述

情報を反映させる。そして GDA タグの属性に DAMSL のタグセットを組み込むことによって，GDA タグのエレメントが DAMSL のスラッシュユニットを意味する．図 1 のように既存の J_ToBI タグではタイムスタンプとタグが付与されているだけなので，GDA タグの属性に J_ToBI 情報とタイムスタンプを含めることによって，発話テキストとの対応を分かりやすくする．

以上のことを考慮して拡張を行った GDA の記述例を図 5 に示す。n タグや v タグなどの GDA タグの既存の sem 属性値や arg 属性値に，表 1 の DAMSL のタグセットを記述する．その際，DAMSL のスラッシュユニットの区切りと GDA タグによる区切りが異なる場合，スラッシュユニットの区切りに合わせるために GDA タグの追加を行なう．その他に，J_ToBI を用いた韻律情報及び発話時間を記述する．prn 属性はローマ字表記により表された単語情報を表す word 層，tone 属性は”%H”などのアクセント情報を表す tone 層の情報を記述する．また時間情報について，begin 属性で発話開始時間，end 属性で発話終了時間を表す.

3.2　MAML との融合

動作や表情などの映像に関する情報は，必ずしも発話テキストと対応できるわけではなく，前節の方式だけでは，映像の情報を記述することが出来ない．そこで，汎用アノテーション記述言語 MAML を適用する．MAML は，人間が理解・記述しやすい文章中心の表現構造と，メディアの種類やフォーマットに依存しない統一的な記述

```
<maml>
 <media type="movie"
   duration="01:42:56">
  <element id="1" begin="4.954" end="10.421">
       :
  </element>
       :
  <element id="22" begin="68.340" end="69.260">
   <audio>
    <utterance id="B">
     解答します
    </utterance>
   </audio>
   <visual>
    <character id="B">
    </character>
   </visual>
  </element>
  <element id="23" begin="69.260" end="70.304">
       :
  </element>
       :
 </media>
</maml>
```

図 6　MAML 記述例

仕様を用いたアノテーション記述言語である．MAML の記述例を図 6 に示す．

MAML では，最上位に maml タグ，その下層にメディア情報を表す media タグを、さらにその下層にアノテーションの基本単位となる element タグを記述する。element タグの下層に音声情報を記述する場合は audio タグ，映像情報を記述する場合は visual タグを列挙していく．さらにその下層により詳細な分類を示すためのクラスタグを記述する．

モーダルの情報は音声，映像などの情報に分類されるが，MAML を用いることにより，それらの情報を一元的に記述処理可能である．音声に関するモーダル情報に関しては，拡張した GDA を MAML の発話（utterance）タグの下層に記述する．一方動作，表情などを含めた映像情報は，人物（character）タグの下層に記述する．動作を表す Anvil は XML 形式であるので，図 3 の Anvil の記述仕様をそのまま下層に用いる．表情を表す FACS は，facs タグを用い，expression 属性に表情を指定することにより表情を記述する．図 7 に拡張した GDA を用いた MAML の記述例を示す。また，拡張を行った部分の DTD を図 8 に示す．intrasentential は GDA タグを表している。

```
<maml>
 <media>
  <element id="1" begin="4.954" end="10.421"
     :
  </element>
     :
  <element id="22" begin="68.340" end="69.260">
   <audio>
    <utterance who="B">
     <su syn="fc" sem="com">
      <n prn="kaito-" tone="%L H-"
            begin="68.349" end="23.659">
       解答
      </n>
      <v agt="B" prn="shi" tone=""
                 begin="23.659" end="23.921">
       し
      </v>
      <v prn="ma'su" tone="L%" begin="68.946"
                               end="69.255">
       ます
      </v>
     </su>
    </utterance>
   </audio>
   <visual>
    <character who="B">
     <attribute name="phase">
              stroke</attribute>
     <attribute name="where">
              space</attribute>
     <attribute name="location-height">
              head</attribute>
     <attribute name="location-side">
              right</attribute>
     <attribute name="handedness">
              right</attribute>
         :
    </character>
   </visual>
  </element>
     :
 </media>
</maml>
```

図 7　MAML と拡張 GDA を併用した記述例

4. 適 用 領 域

本章では，前章で策定したメタデータの適用例について述べる．

4.1　検索・要約

情報化社会の今，膨大な量のコンテンツの中からユーザが自分の目的とするデータを自分で探しだし，能動的にアクセスする作業が必要となる．そこで，多くの情報を含めたメタデータを用いることよって検索の幅が広が

```
<!ENTITY % intrasentential "su | n | v | ad |…">
<!ATTLIST %intrasentential  prn   CDATA #IMPLIED
                            tone  CDATA #IMPLIED
                            begin CDATA #IMPLIED
                            end   CDATA #IMPLIED
>
<!ELEMENT character (#PCDATA | attribute* | facs)>
<!ELEMENT attribute (#PCDATA)>
<!ATTLIST attribute name CDATA #IMPLIED>
<!ELEMENT facs (#PCDATA)>
<!ATTLIST facs expression CDATA #IMPLIED>
```

図 8　拡張を行った部分の DTD

り，効率よく意味内容に基づくインタラクティブな検索・要約が可能となる．

4.2　メディア生成

メディア生成を行うために，音声，映像などの多くの情報が必要となる．そこで，今回策定したメタデータには，各種のモーダル情報が含まれているため，このメタデータを用いることによって，多種多様な情報を含めたメディア生成を行うことが可能となる．

4.3　仮 想 現 実

仮想現実とは，コンピュータによって現実には存在しない空間を作り出し，それに接する人をあたかもその空間にいるかのような体験をさせる，というものである．仮想現実空間を作り出すためには，やはり多くの音声や映像などの情報が必要である．その点から今回策定したメタデータには多くのモーダル情報が含まれているため，仮想現実への適用が可能となる．

5.　応用事例：動画生成

本章では，前章で列挙した適用事例の中で，今回メディア生成として，動画生成を行うスクリプト言語である TVML への変換を行うことにより動画生成を行う例について述べる．

5.1　TVML への変換方法

TVML とは，一本のテレビ番組をリアルタイムに生成することができるように考えられたテキストベースの言語であり，一行に一つのイベントを記述する．MAML の `audio` タグ，`visual` タグそれぞれの下層に記述された内容を一つのイベントと考えることにより，TVML との対応を取る．以下では変換した内容の詳細を述べる．

発話情報を表す `utterance` タグの下層に記述されたアノテーションテキストは，1 文，つまり `su` タグごとにアノテーションテキストを連結させ，TVML の talk コマンドの引数内の text パラメータに記述する．

`utterance` タグの who 属性に記述された発話者を表す情報は，TVML では character イベントの casting コマンドなどを用いて人物の設定が行う必要があるため，casting コマンドの引数内の name パラメータに記述する．また，character イベントの talk などの name パラメータに記述する必要がある場合，各コマンドの name パラメータに記述する．

GDA タグの属性（sem,arg,etc）に記述された談話の種類を表す情報は，表 1 の中で TVML に反映できるものとして ir(wh) など疑問の発話の場合，TVML の talk コマンドの text パラメータの最後に”?”を付与する．これにより，音声合成を用いた時，疑問の発話文となる．

発話時間を表す GDA タグの begin，end 属性の値は，`su` タグ内の最初の begin 属性の値と最後の end 属性の値を用いることにより発話時間を求めることが可能であるが，TVML では発話時間を指定することが不可能である．そのため，TVML の talk コマンドにおける発話時間を 1 文字の発話時間を 0.2 秒として 1 文中の文字数をかけることにより求め，begin，end 属性により求めた発話時間との差を取り，その差によって talk コマンドの発話速度を変更する時に用いる rate パラメータの値を決定する．また発話間隔を求め，wait コマンドの time パラメータに記述する．

`visual` タグの下層に記述される映像情報は，その下層に記述されるタグによって TVML のコマンドを変える．背景情報を表す `background` タグで記述されている場合，set イベントの各コマンドを記述する．`character` タグの下層に動作情報を表す `attribute` タグが記述されている場合，character イベントの主に pose コマンドを記述する．pose コマンドを用いる際，あらかじめ definepose コマンドによりポーズの定義をする必要がある．表情を表す `facs` タグで記述されている場合，`facs` タグの character イベントの expression コマンドの type パラメータに記述する．

その他，MAML には多くの情報が含まれているが，TVML への変換を行っていない情報が多くある．例えば GDA タグの prn，tone 属性に関してである．TVML は，音声合成エンジンを用いて発話を行っているため，ローマ字による単語情報は必要としない．また，形態素ごとに発話の高さの調整を行なうことが不可能であるため，GDA のタグに付与した prn 属性と tone 属性の情報は反映できないと考えられるからである．

5.2　変換の実行例と TVML 画面出力例

図 7 の拡張された GDA を用いた MAML の記述例の TVML 形式への変換の例を示す．図 7 はマルチモーダル対話に関する動画のメタデータである [9]．その動画の画面の例を図 9 に示す．そして図 7 を前節で説明した方法に

図 9　マルチモーダル対話に関する動画

```
skipscript(switch=on)
      :
character: casting(name="A")
character:openmodel(modelname=A
 ,filename=Datafiles/Character/Mina/Mina.bm)
character: bindmodel(name=A, modelname=A)
character: position(name=A,
 x=0.35,y=0.05,z=2.3,d=0.0,posture=sitting)
character:setvoice(name=A, voicetype=j_woman)
character: look(name=A, what=camera)
character: gesture(name=A, degree=-10.0)
character: casting(name="B")
      :
character: gesture(name=B, degree=-10.0)
      :
character: definepose(name=B, pose=R, joint
 =RightUpperArm, rotx=20, roty=70, rotz=0)
character: definepose(name=B, pose=R, joint
 =RightLowerArm,rotx=-170,roty=-90,rotz=180)
character: definepose(name=B, pose=R, joint
 =RightHand, rotx=0, roty=0, rotz=0)
      :
character: talk(name=B, text="解答します"
        d           , rate=0.0, volume=0.0)
character: pose(name=B, pose=R)
wait(time=1.004)
character: talk_wait(name=B)
      :
```

図 10　TVML への変換例

より変換した TVML を図 10 に，その TVML を TVML プレイヤを用いて出力した画面の例を図 11 に示す．

図 9 の二人の女性は，図 7 において人物 A と B で表され，その人物 A と B は図 10 において，character イベントの casting コマンド等を用いてることにより二人の人物の設定が行われる．発話を表す `utterance` タグの下層に記述された「解答します」という発話は，character イベントの talk コマンドを用いて記述される．begin, end 属性により求めた発話時間と TVML の talk コマンドにおける発話時間との差がほぼなかったため，発話速度を表す rate パラメータは 0.0 と記述される．`character` タグの下層の動作を表す `attribute` タグに記述された「右手を挙げる」という動作は，character イベントの definepose コマンドを用いて右腕の各関節の角度を指定した後，character イベントの pose コマンドを用いて記述されることによってその動作が行われる．

図 11　TVML 出力画面の例

今回の TVML への変換は発話テキスト，時間，談話，動作の情報に限定されるが，これは TVML の仕様により，策定したメタデータ仕様の他の情報が反映できないためである．メタデータに記述された全ての情報を踏まえることができれば，より情報量の豊富なメディアコンテンツの生成が行えるようになる．

6. 実　　験

本章では，生成した動画の聴衆実験とその結果について述べる．

6.1　実験の概要

聴衆実験の 2 つの動画に関して比較を行った．(1) 発話テキストと背景や人物のみを TVML に変換し，TVML プレイヤを用いて再生した CG 動画．(2) (1) の他に時間，映像などの情報を含めて TVML に変換し，TVML プレイヤを用いて再生した CG 動画．

(1) と (2) に用いたメタデータはマルチモーダル対話に関する動画（図 9）のメタデータである．この動画は二人の対話者が人物写真の問題シートを見ながら，一人がある人の顔の特徴を説明し，もう一人がその人物を当てる「顔課題」という内容である．この動画では，顔の特徴を説明するときに，モーダル情報による意志疎通が行われている．

(1) の発話テキストとは，`su` タグの下層の各々の GDA タグの下層に記述されたアノテーションテキスト部の情報のみのことである．次に (2) は，発話テキスト以外に TVML への変換を行うことが可能である時間，談話，動

作の情報を含めて TVML への変換を行う．図 10 がこのメタデータの記述例である．

実験の流れは元の動画, (1) , (2) の順に動画を視聴し, (1) と (2) の動画においてモーダル情報が再現されているか 7 段階評価を行う．7 段階の評価項目は以下のようにした．

1. 再現できていない
2. ほとんど再現できていない
3. あまり再現できていない
4. どちらともいえない
5. 多少再現できている
6. ほとんど再現できている
7. 再現できている

この 7 つの項目からそれぞれの TVML 動画について 1 つ選ぶ．また，この評価の他に動画に関する意見などを記述してもらった．

6.2 実験結果

12 名に対して前節で説明した元動画, (1) , (2) の動画の評価結果の平均点を表 2 に示す.

表 2　実験結果 (平均点)

動画	平均点
(1) 発話と背景と人物情報を変換した TVML	3.0
(2) (1) の他に動画と時間情報を変換した TVML	5.0

7. 考　　察

実験結果において (1) の TVML 動画より (2) の TVML 動画の方が良い結果が出ており，また意見として「動作は話の文脈理解の助けになる」,「読点による句の区切りは不可欠」という意見などがあった．これは今回のメタデータの仕様として動作のモーダル情報や時間情報を追加したためである．このことから, (1) の TVML 動画のように発話だけの情報を含めた動画を生成するより, (2) の TVML 動画のように時間や動作を考慮した方が有効であるということが言える．

一方,「TVML を用いた動画だけを見ても内容を理解しづらい」という意見があった．これは実際元動画では，画面が 2 分割され二人の対話者が離れたところに座っているのに，TVML では隣同士に座ってしまっているためであった．それは今回用いたメタデータでは画面の 2 分割など細かい設定を記述していないため，TVML を用いた動画生成においても隣同士に座っている設定の動画になってしまった．TVML の仕様において画面の 2 分割などの細かい設定が不可能であり，マルチモーダル記述のメタデータ仕様においてもそのような記述の想定をしていない．しかし今回用いたマルチモーダル対話に関する動画などのメタデータにおいて，対話者の位置関係などは重要なことであると思われる．そこで，映像に関するメタデータの記述の仕様を詳細な記述が可能なように考える必要があると思われる．

また, (2) の動画においてもメタデータに含まれていた韻律の情報を今回の動画生成において加味していないため，イントネーションなど不自然な点が多く，内容理解のためには韻律の情報を含めることが重要であると思われる．

8. ま と め

本稿では GDA の拡張を行うことにより，音声に関するモーダル情報を記述し，MAML を用いることにより，映像に関するモーダル情報を記述した。そして拡張した GDA を MAML に適用することにより，一つの記述仕様で複数のモーダル情報が記述可能となった．これにより，複数のモーダル情報の相互の対応付けが困難であるという既存の問題は解決された．また，複数のモーダル情報を記述したメタデータを用いることにより，モーダル情報を加味したメディアコンテンツを生成することが可能となった．今後は，今回策定したメタデータ記述仕様に含まれなかったモーダル情報の融合に関しての検討を行っていく必要があると思われる。

謝　　辞

GDA の提唱者であり，また今回のタグ仕様の策定にあたり多くの助言をいただきました，産業技術総合研究所の橋田浩一さんに深く感謝いたします。

文　　献

[1] MPEG-7 Japan: http://www.itscj.or.jp/mpeg7/
[2] 橋田浩一: GDA 意味的修飾に基づく多用途の知的コンテンツ, 人工知能学会論文誌, Vol.13, No.4, pp.528–535, 1999.
[3] 伊藤一成, 斎藤博昭: メディアデータに対するアノテーション記述言語 (MAML) の策定とその応用, 情報処理学会研究報告, FI70–4, pp.19–26, 2003.
[4] TV program Making Language(TVML): http://www.strl.nhk.or.jp/TVML/
[5] J.J.Venditti:Japanese ToBI Labelling Guidelines, Ohio-State University, Columbus, U.S.A, 1995.
[6] James Allen and Mark Core: Draft of DAMSL, University of Rpchester, 1997.
[7] Michael Kipp: ANVIL A Generic Annotation Tool for Multimodal Dialogue, Eurospeech, 2001.
[8] Extensible Markup Language(XML): http://www.w3.org/XML/
[9] JEITA 電子情報技術産業協会対話コンテンツ技術専門委員会 (橋田浩一，斎藤博昭，伊藤一成他) “ JEITA マルチモーダル対話コーパス“ , 第 16 回人工知能学会全国大会 (JSAI2002) ポスターセッション，2002.