問題構造アノテーション仕様書 ver.14

2015年12月8日 版

目次

問題構造アノテーション仕様書 ver.14 1

1 概要 3

2 注意事項 3

3 文字コード 3

4 文書情報 3

5 タイトル部分 5

6 問題 6

7 データ 10

7.1 テキストデータ 12

7.2 画像データ（写真、地図、概念図、グラフ等） 13

7.3 独話音声データ 14

7.4 会話音声データ 14

7.5 表データ 15

7.6 複数のデータのグループ 18

8 解答欄 19

9 選択肢 20

10 数式 24

11 テキストの装飾 24

12 空欄 28

13 クロスリファレンス 29

14 文字化け・欠落テキストの修正 30

15 レイアウト情報 30

参考文献 30

【変更点】

ver.14での変更点（青字）

\*7.3、7.4に独話音声データ、会話音声データのアノテーション方法を追加（仕様決定者：電気通信大　南泰浩先生）

ver.13での変更点（青色のマーカー）

\*refタグの対象にページ、段落、行を追加（13. クロスリファレンス）

ver.12での変更点（灰色のマーカー）

\*選択問題がある場合のexam属性の記述例を追加（4.文書情報）

ver.11での変更点（黄色のマーカー）

\*段落、および行の開始位置の記述方法を追加（15. レイアウト情報）

\*物理問題のヘッダの書き方変更

\*<pageEnd>の属性を消去

\*<pageStart>のidを文書内固有に

ver.10での変更点

\*数学問題のヘッダの書き方を変更

ver.9での変更点

\*数学問題のみ対象としていたquestion要素のanscol\_ids属性を削除し、全科目共通のanscol\_idに統一。複数回答の場合の書き方も全科目同じ形式に統一した。

\*instructionとdataの違いについて加筆

ver.8での変更点

\*kaeriタグの付与についての注意事項を追加（11.テキストの装飾）

\*ルビのアノテーション仕様を追加（11.テキストの装飾）

\*UTF-8にない文字に対する対処法を追加（11.テキストの装飾）

\*英語の空欄に関する注意事項を追加（12.空欄）

\*国語問題のページの開始位置・終了位置の記述方法を追加（15. レイアウト情報）

ver.7での変更点

\*英語問題において、blankタグ内へのansColumnタグ埋め込み禁止を追加（8.解答欄、12.空欄）

ver.6での変更点

\*okuriLタグを追加（11.テキストの装飾）

ver.5での変更点

\*@rowspanと@colspanの定義を逆にした

\*XML宣言にエンコーディングの情報を付与

# 概要

本稿は、大学入試問題を対象とした、問題構造アノテーションの仕様書である。

OCRによる認識済みの試験問題に対して、主に問題の表層的かつ直感的な構造、および参照関係等を明示的に記述することを目的としており、今後のより複雑なアノテーションの基礎となるものである。

この文書に出てくるアノテーション例はすべて、平成22年度大学入試センター試験の問題（http://www.dnc.ac.jp/modules/center\_exam/content0232.html）を利用している。

# 注意事項

* アノテーションに利用する入試問題データの管理は厳重にお願いいたします。入試問題データをアノテーション作業以外の目的に使用すること、および第三者に対して売買、貸与、刊行、配布、送信可能化することを禁止します。また、アノテーション作業および最終納品が終了し、当方から納品確認の連絡をした後は、入試問題データをすべて破棄していただけますようお願いいたします。
* アノテーションに利用するファイルの不備や欠落、作業上不明な点等ございましたら担当者までご連絡ください。
* 担当者連絡先

川添愛（国立情報学研究所　社会共有知研究センター）

zoeai@nii.ac.jp

# 文字コード

文字コードはUnicode(UTF-8)を使用する。

# 文書情報

　冒頭は、数学・物理以外の全ファイルでは次のように記述する。

<?xml version='1.0' encoding=”UTF-8”?>

<!DOCTYPE exam PUBLIC "-//TOROBO//TOROBO ANNOTATION 1.0//EN" "torobo.dtd">

<?xml-stylesheet type="text/css" href="../torobo.css"?>

数学・物理のファイルでは、次のように記述する。

<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE exam PUBLIC "-//TOROBO//TOROBO ANNOTATION 1.0//EN" "torobo-math.dtd">

<?xml-stylesheet type="text/xsl" href="final.xsl"?>

ルート要素は<exam>である。source等以下の属性によって試験問題の基本情報を記述する。

<exam>

@source　 問題の出題機関

\*代ゼミ模試は、次のようにする。

センター模試：Yoyogi Seminar Center Prep Test

東大プレ：Yoyogi Seminar Todai Prep Test

@year　 年度

@subject　 科目名

（ファイル名を参考にKokugo, Eigo, SekaishiB, ButsuriIのように入れる）

@srcTxtURL　 問題のURL（任意）

@range\_of\_options （選択問題がある場合）選択問題の範囲（選択問題のid）

@num\_of\_options （選択問題がある場合）選択する問題の数

1. 例：大学入試センター試験の場合

<?xml version='1.0'?>

<!DOCTYPE exam PUBLIC "-//TOROBO//TOROBO ANNOTATION 1.0//EN" "torobo.dtd">

<?xml-stylesheet type="text/css" href="../torobo.css"?>

<exam source="National Center For University Entrance Examination" subject="SekaishiB(makeup exam)" year="2009">

Center-2009--Supp1-SekaishiB<br/>

<title>

2009年度 追試験 世界史Ｂ<br/>

<br/>

</title>

............

</exam>

1. 例：代ゼミ模擬試験の場合

<?xml version='1.0'?>

<!DOCTYPE exam PUBLIC "-//TOROBO//TOROBO ANNOTATION 1.0//EN" "torobo.dtd">

<?xml-stylesheet type="text/css" href="../torobo.css"?>

<exam source="Yoyogi Seminar Center Prep Test" subject="SekaishiB" year="2012">

YozemiCenter-2012--1-SekaishiB<br/>

<title>

2012年度　代々木ゼミナール第1回全国センター模試　世界史B<br/>

<br/>

</title>

............

</exam>

1. 例：数学問題で、選択問題がある場合

<?xml version='1.0' encoding="UTF-8"?>

<!DOCTYPE exam PUBLIC "-//TOROBO//TOROBO ANNOTATION 1.0//EN" "torobo-math.dtd">

<?xml-stylesheet type="text/xsl" href="final.xsl"?>

<exam source="National Center For University Entrance Examination" subject="SuugakuIIB(main exam)" year="2005" range\_of\_options="Q5 Q6 Q7 Q8" num\_of\_options="2">

Center-2005--Main-SuugakuIIB<br/>

<title>

2005年度 本試験 数学Ⅱ・Ｂ<br/>

</title>

............

</exam>

# タイトル部分

タイトル部分のアノテーションに使う要素は以下の通りである。

<title>　 タイトル

<info>　 解答番号の範囲や配点などの情報、「全問必答」などの指示。

（タイトルの直後だけでなく、問題の内部にも出現する。）

1. 例

<title>世界史A</title>

<info>解答番号（1~33）</info>

# 問題

問題は多くの場合、大きな問題がより小さな問いを含む入れ子構造になっている（例(4)を参照）。ここでは、大小問わずすべての問題に対して<question>要素を使ってタグ付をする。それ以上小さい問題を含まない問題を明示的に示すために、minimal属性を用意している。入れ子になった問題のアノテーション例については、(4)を参照。

<question> 問題の範囲を表す

@id 識別番号（文書内で固有。「Q」の後に半角数字を付けたものを入れること）

@minimal yes/noを取る。それ以上小問を含まない問題の場合のみyes。

@answer\_style 回答方式による問題の分類。（「問題分類アノテーションスキーマ」を参照）

@answer\_type 答えのタイプによる分類。minimal属性がyesの場合のみ指定。（別紙「問題分類アノテーションスキーマ」を参照）

@knowledge\_type 回答の際に必要となる知識による分類。minimal属性がyesの場合のみ指定。（別紙「問題分類アノテーションスキーマ」を参照）

@anscol 数学以外の問題では、minimal属性がyesの場合に、対応する解答欄（<ansColumn>要素）のidを入れる。数学では、大問番号（半角）の後に、対応する空欄の記号（全角カタカナ）を続けたもの（例：1ア、4エ）を入れる。全科目において、問題に対応する解答欄が複数ある場合は、「A1 A2」「1ア 1イ」のように、スペースを挟んで値を複数記入する。ただし、「順不同」の指定がある場合は「A1|A2」「1ア|1イ」のように縦棒（半角）で区切る。また、正しい組み合わせで正解を入れなければ点数が付かない解答欄の組は、「A1,A2」「1ア,1イ」のようにカンマで区切る。

例：「大問１　3+7=[アイ]、1+2=[ウ] である。」のような問題のanscol属性は　anscol="1ア,1イ 1ウ"　となります。

問題の内部には主に以下のものが含まれる。

* 問題のラベル

<label> 問題に振られているラベル（「第一問」「A」「問1」など）

* 指示（「以下の文章を読み、○○について答えよ」「○○は何か」のような文を含む）

<instruction>　 解答の指示や質問を記述した箇所

* データ（文章や図、表）

<data>　 文章や図、表などを含む、解答者が解答を出すにあたって参考にするデータ全般。

@id 識別番号（文書内で固有）

@type text、image、table, complex等の値をとる。

詳しくは、第7節「データ」を参照。

* 解答欄

<ansColumn> 解答欄

@id 識別番号（文書内で固有）

詳しくは、第8節「解答欄」を参照。

* 選択肢

<choices> 選択肢のまとまり。

@anscol 対応する回答欄の識別番号（複数指定可）

<choice> 個々の選択肢。

@ra 選択肢が正解かどうかを記述する。yes/noの値をとる。

<cNum> 選択肢の番号。

詳しくは、第9節「選択肢」を参照。

* 数式

<formula>

詳しくは、第10節「数式」を参照。

1. 例：問題が入れ子になっている場合

**<question id="Q1" minimal=”no”>**

<label>第1問</label>　<instruction>次の問い（A・B）に答えよ。（配点　14）</instruction>

**<question id="Q2" minimal="no">**

<label>A</label>　<instruction>次の問い（問1～4）において、下線部の発音がほかの三つの場合と異なるものを、それぞれ1～4のうちから一つずつ選べ。</instruction>

**<question anscol="A1" answer\_style="multipleChoice" answer\_type="orthography" id="Q3" knowledge\_type="PRN" minimal="yes">**

<label>問1</label> <ansColumn id=”A1”>1</ansColumn>

<choices anscol=”A1”>

<choice ansnum="1" ><cNum>1</cNum> b<uText id=”u01”>oo</uText>t</choice>

<choice ansnum="2" ><cNum>2</cNum> g<uText id=”u01”>oose</uText></choice>

<choice ansnum="3" ><cNum>3</cNum> pr<uText id=”u01”>oof</uText></choice>

<choice ansnum="4" ><cNum>4</cNum> w<uText id=”u01”>ool</uText></choice>

</choices>

**</question>**

**<question anscol="A2" answer\_style="multipleChoice" answer\_type="orthography" id="Q4" knowledge\_type="PRN" minimal="yes">**

<label>問2</label><ansColumn id=”A2”>2</ansColumn>

<choices anscol=”A2”>

<choice ansnum="1" ><cNum>1</cNum> bread<uText id=”u01”>th</uText> </choice>

<choice ansnum="2" ><cNum>2</cNum> fai<uText id=”u01”>th</uText></choice>

<choice ansnum="3" ><cNum>3</cNum> leng<uText id=”u01”>th</uText></choice>

<choice ansnum="4" ><cNum>4</cNum> smoo<uText id=”u01”>th</uText> </choice>

</choices>

**</question>**

...........

**</question>**

...........

**</question>**

何をもって一つの独立した問題とみなすかについては、判断が難しい場合もある。ここでは「ラベル（番号、アルファベットその他の記号）」を持つものは「問題」と見なす。例えば、(4)の例で赤字で示したQ3、Q4の問題のように、明示的な指示や質問を含まず、解答欄と選択肢のみ含むものや、あるいはそれらに加えて解答選択のキーとしての語句を含むもの（後の(7)を参照）であっても、「問2」のようなラベルが付いていれば一つの問題と見なし、<question>を用いてタグ付けする。これに対して、(5)のように、解答欄と選択肢があっても、他の部分から独立したことを示すようなラベルがない場合は、独立した問題とはみなさない。

1. 独立した問題とはみなされない部分（赤字）

<question anscol="A29 A30 A31" answer\_style="multipleChoice" answer\_type="sentence" id="Q34" knowledge\_type="DIS\_W" minimal="yes">

<label>B</label> <instruction>次の英文は、地域の公開討論会で行われた高齢者問題についての議論の一部である。<ref target="B35">29</ref>～<ref target="B37">31</ref>に入れるのに最も適当なものを、それぞれ下の1～4のうちから一つずつ選べ。</instruction>

<data id="d11" type="text">

Chair: Today's discussion is about how younger people should communicate with the elderly. We invited three panelists:Dr.McDonald, a medical doctor, and Mr. Johnson and Ms. West, who are researchers specializing in gerontology, or the study of the elderly. Mr. Johnson, can you start our discussion?

Mr. Johnson:Certainly. I often notice people communicate in a particular way with the elderly. People exaggerate their-ways of speaking, for instance, by using a loud voice. Sometimes people speak as if ....（中略）.....

Chair: That's an interesting observation. You're saying that <blank id="B35"><ansColumn id=”A29”>29<ansColumn></blank>. What do you think, Dr. McDonald?

<choices anscol=”A29”>

<choice ansnum="1" ><cNum>1</cNum> doctors and nurses often use too many technical

expressions</choice>

<choice ansnum="2" ><cNum>2</cNum> many elderly people are comfortable when they speak to doctors</choice>

<choice ansnum="3" ><cNum>3</cNum> medical professionals also speak differently to the elderly</choice>

<choice ansnum="4" ><cNum>4</cNum> some people can't communicate without speaking like

children</choice>

</choices>

（中略）

Chair: OK. Dr. McDonald, with his professional experience, thinks that <blank id="b36" label="30"><answer\_column id=”A30”>30</answer\_column></blank>. What are your thoughts on this point, Ms. West?

<choices anscol=”A30”>

<choice ansnum="1" ><cNum>1</cNum> age has nothing to do with speech style</choice>

<choice ansnum="2" ><cNum>2</cNum> more intelligent words should be used</choice>

<choice ansnum="3" ><cNum>3</cNum> special attention should be paid to the elderly</choice>

<choice ansnum="4" ><cNum>4</cNum> the elderly should speak more slowly and clearly</choice>

</choices>

（中略）

</data>

</question>

# データ

解答者が解答を出すにあたって参考にする情報には、文章や図、表、およびそれらに対するラベル、説明文などが含まれる。ここではそのような情報全般を、<data>要素を利用してタグ付けする。

<data>　 文章や図、表などを含む、解答者が解答を出すにあたって参考にするデータ全般。

@id 識別番号（文書内で固有。「D」の後に半角数字をつけたものを入れる）

@type text, image、speech\_monolog, speech\_conversation, table, complex等の値をとる。

<data>に共通して含まれる要素には、以下のものがある。

<label> データにつけられたラベル（「表1」「図1」「A」「ア」など）

<caption> （主に表や図などに対する）説明文

テキストと図、テキストと表、図と表など、異なる種類のデータが隣接して現れている場合は、基本的にそれぞれ独立に<data>要素としてタグ付けする。

　<data>と<instruction>の区別については、基本的に、「～せよ」のような明示的な指示や「～はどれか」のような明示的な質問が現れている部分を<instruction>としてタグ付けする。それ以外の、解答の参考になるデータは<data>でタグ付けする。指示および質問の部分と、解答の参考になるデータの部分が一つの部分に混在する場合、段落が分かれて間が一行空いている、あるいは固有のラベルで区別できる場合のみ、別個にアノテーションする。区別がしづらい場合は、全体を<instruction>でアノテーションする。

1. 例：データを含む問題

<question id="Q1" minimal=”no”>

<label>第1問</label>　<instruction>歴史上の支配者・指導者による近代化の試みについて述べた次の文章A～Cを読み、下の問い（問1～11）に答えよ。（配点　33）</instruction>

**<data id="D1" type="text">**

**<label>A</label>** オスマン帝国の支配機構は、<uText id="U1">先行するイスラーム諸王朝</uText>の様々な特徴を継承しつつ形成され、独特の発展を遂げた後、18世紀後半以降、西洋化の道を歩む。セリム3世（在位1789～1807年）は、<uText id="U2">ヨーロッパ列強の圧力</uText>が強まるなかで、帝国の体系的な近代化を初めて実行した君主だった。彼は、西洋式軍隊を創設するとともに、西洋の軍事技術の導入を推進し、（中略）また、主要な<uText id="U3">ヨーロッパ諸国</uText>の首都に常駐大使館を開設して、（中略）彼の試みは、守旧派の反乱によって挫折したが、やがて<uText id="U4">④ 19世紀</uText>の大半を通じて進められる全般的な西洋化改革に継承される。

**</data>**

<question id="Q2" ....（中略）....... minimal="yes">

<label>問1<label><instruction><ref target="U1">下線部</ref>に関連して、イスラーム諸王朝の支配者について述べた文として正しいものを、次の1～4のうちから、一つ選べ。</instruction>

<ansColumn id=”A1”>1</ansColumn>

<choices anscol=”A1”>

<choice ansnum="1"><cNum>1</cNum> ファーティマ朝の君主は、後ウマイヤ朝を滅ぼしてカリフ位を奪った。</choice>

<choice ansnum="2"><cNum>2</cNum> セルジューク朝の君主は、ウマイヤ朝カリフからスルタンの称号を与えられた。</choice>

......（中略）

</choices>

</question>

...... （中略）

</question>

語彙問題などで、解答選択のキーとして選択肢の直前に現れる語句なども<data>要素でタグ付けする。

1. 例：選択肢の直前に現れる語句

<question id="Q7" minimal ="no">

<label>B</label>　<instruction>次の問い（問1～3）において、与えられた語と第一アクセント（第一強勢）の位置が同じ語を、それぞれ1～4のうちから一つずつ選べ。</instruction>

<question id="Q8" ....（中略）....... minimal="yes">

<label>問1</label>

**<data id="D1" type="text">damage</data>**

<answer\_column id=”A5”>5</answer\_column>

<choices anscol=”A5”>

<choice ansnum="1"><cNum>1</cNum> convince</choice>

<choice ansnum="2”><cNum>2</cNum> effort</choice>

<choice ansnum="3"><cNum>3</cNum> prefer</choice>

<choice ansnum="4"><cNum>4</cNum> throughout</choice>

</choices>

</question>

......（中略）

</question>

## テキストデータ

　解答の参考になるデータのうち、文章はtype値としてtextを取る（参考：例(6)(7)）。そのような<data>要素は以下の要素を含む場合がある。

<quote> 文章中に現れる引用。

<note> 注。

@id 注の識別番号（文書内で固有）

<source> 文章の出典情報。

1. テキストデータのアノテーション例

**<data id="D1" type="text">**

　<ref target="n01">注1</ref>フロイトによれば、人間の自己愛は過去に三度ほど大きな痛手をこうむったことがあるという。...（中略）...まずは<ref target="n02">（注2）</ref>「ヴェニスの商人」について語らなければならない。...（中略）...

だが、<uText id="U1"><label>A</label>経済学という学問は、まさに、このヴェニスの商人を抹殺することから出発した</uText>。

**<quote>**年々の労働こそ、いずれの国においても、年々の生活のために消費されるあらゆる必需品と有用な物質を本源的に供給する基金であり、...（中略）...したものである。**</quote>**

『国富論』の冒頭にあるこのアダム・スミスの言葉...（中略）...一国の富の増大のためには外国貿易からの利潤を貨幣のかたちで<uText id="U2"><label>(ア)</label>チク</uText>セキしなければならないとする、重商主義者に対する挑戦状にほかならない...（以下略）

**<source>**（岩井克人「資本主義と『人間』」による）<**/source>**

（注）**<note id="n01">**1　フロイト――オーストリアの精神医学者（1858～1939）。**</note>**

**<note id="n02">**2　「ヴェニスの商人」――シェークスピアの戯曲「ヴェニスの商人」を...（以下略）**</note>**

**</data>**

## 画像データ（写真、地図、概念図、グラフ等）

解答の参考になるデータのうち、写真、地図、概念図、グラフ等の画像からなるものは、type値としてimageを取る。そのような<data>要素は以下の要素を含む。

<img/> 画像

@src 画像のファイル名（拡張子必須）

@comment コメント（任意）

1. 画像データのアノテーション例。左は元の画像（地理B、p.158より）。



**<data id="d01" type="image”>**

**<img src=”meatconsum.jpg"/>**

**<caption>**統計年次は2001～2003年のいずれか。FAOの資料により作成。**</caption>**

<label>図3</label>

**</data>**

## 独話音声データ

リスニングデータは，、いくつかの問題のリスニングデータが含まれているので，各問題ごとの分割する必要がある．type値としてspeech\_monologを取る。そのような<data>要素は以下の要素を含む。

<spc /> 独話音声

@src 音声のファイル名（拡張子　flac）

音声ファイルフォーマットは以下の通りである．

サンプリングレート：１６ｋHz

チャンネル：モノ

ファイルフォーマット：flac

<caption> 音声の開始時間と終了時間をsec 単位で記述する．

1. 独話音声データのアノテーション例。

<data id="d01" type="speech\_monolog”>

<spc src=”mono1.flac"/>

<caption> start:0 end:1.0</caption>

<label>d01</label>

</data>

## 会話音声データ

リスニングデータは，、いくつかの問題のリスニングデータが含まれているので，各問題ごとの分割する必要がある．type値としてspeech\_conversationを取る。そのような<data>要素は以下の要素を含む。

<spc/> 会話音声

@src 音声のファイル名（拡張子　flac）

音声ファイルフォーマットは以下の通りである．

サンプリングレート：１６ｋHz

チャンネル：モノ

ファイルフォーマット：flac

<caption> 個別の会話音声の場合，話者番号，音声の開始時間と終了時間をsec 単位で記述する．

　　　　　　全体の会話音声の場合，話者数，音声の開始時間と終了時間をsec 単位で記述する．

会話音声の場合，将来のことも考えて，切り出し前のデータにもアクセスできるように，このファイルも参照できるようにしておく．このため，アノテーションは階層的になる．

1. 対話音声データのアノテーション例。

<data id="D1" type="speech\_conversation”>

<spc src=”conversation\_all1.flac"/>

<data id=”D1\_1” type=”speech\_conversation”>

<spc src=”conversation1\_1.flac"/>

<caption>話者:1　start:0 end:1.0 </caption>

<label> D1\_1 </label></data>

<data id=”D1\_2” type=”speech\_conversation”>

<spc src=” conversation1\_2.flac"/>

<caption>話者:2 start:0 end:1.0 </caption>

<label> D1\_2</label></data>

<data id=”D1\_3” type=”speech\_conversation”>

<spc src=” conversation1\_3.flac"/>

<caption>話者:1 start:0 end:1.5 </caption>

<label> D1\_3 </label></data>

<caption>話者数:2 start:0 end: 3.5 </caption>

<label> D1 </label>

</data>

問題ファイル中には，問題の説明，英語の問題指示，独話が含まれているので，この対話の部分だけを切り出す．conversation1\_all.flacには対話全体を，conversation1\_❓.flacには切り出した対話を保存する．データは対応する<instruction>の後や，<question anscol=>の後に挿入する．

## 表データ

解答の参考になるデータのうち、表からなるものはtype値としてtableを取る。そのような<data>要素に含まれる要素は以下の通りである。ほぼhtmlにおける表の表現に準拠している。（<tbl>はhtmlの<table>, <row>はhtmlの<tr>、<cell>はhtmlの<td>に相当）

<tbl> 表。<row>、<cell>を含む。

<row> 表中の各々の行。<tbl>に含まれる。

<cell> 表中の各々のセル。<row>に含まれる。行内の<cell>の表示順は、左から右である。

（ただし、国語の場合は、右から左とする。）

@colspan　複数のセルが横方向に連結されている場合。結合されているセルの数を記入

@rowspan　 複数のセルが縦方向に連結されている場合。結合されているセルの数を記入

1. 表データのアノテーション例。上は元の表（地理B, p.145より）



<data id=”D1” type=”table”>

<label>表1</label>

<tbl>

<row>

<cell/>

<cell >新庄市</cell>

<cell >金山町</cell>

<cell >舟形町</cell>

<cell >真室川町</cell>

<cell >最上町</cell>

<cell >大蔵村</cell>

<cell >鮭川村</cell>

<cell >戸沢村</cell>

</row>

<row>

<cell >15歳未満人口割合（%）</cell>

<cell >15.0 </cell>

<cell >14.1</cell>

<cell >11.6</cell>

<cell >12.4</cell>

<cell >13.3</cell>

<cell >13.3</cell>

<cel >13.4</cell>

<cell >13.1</cell>

</row>

........

</tbl>

</data>

世界史に出てくる年表も、表データとしてタグ付けする。

1. 表データのアノテーション例。左下は元の年表（世界史B, p.44より）

<data id=”D7” type=”table”>

<tbl>

<row>

<cell >1917年</cell>

<cell >ロシア十月（十一月）革命</cell>

</row>

<row>

<cell colspan=”2”>

<blank id=”b01”><label>a</label>

</cell>

</row>

<row>

<cell >1947年</cell>

<cell >コミンフォルム結成</cell>

</row>

<row>

<cell colspan=”2”>

<blank id=”b02”><label>b</label>

</cell>

</row>

........（以下略）

</tbl>

</data>

## 複数のデータのグループ

　複数のデータ（それぞれラベルや説明文を持つ）のグループに、一つのラベルや説明文などが付いている場合は、<data>要素はtype値としてcomplexを取る。

1. 複数の画像のグループに一つのラベルが付いている例。左は元の画像（地理B、p.138より）



**<data id="D1" type="complex”>**

<data id=”D2” type=”image”>

<img src=”vegetation1.jpg"/>

<label>カ</label></data>

<data id=”D3” type=”image”>

<img src=”vegetation2.jpg"/>

<label>キ</label></data>

<data id=”D4” type=”image”>

<img src=”vegetation3.jpg"/>

<label>ク</label></data>

**<label>写真1</label>**

**</data>**

# 解答欄

解答欄は、<ansColumn>でタグ付けする。

<ansColumn> 解答欄

@id 識別番号（文書内で固有）。

抜けを防ぐため、「A」の後に解答欄の番号（半角数字）を続けたものを

入れる。

（例：解答欄の番号が23の場合：A23）

数学問題の囲み文字は、原則としてすべて、解答欄ではなく空欄とみなし、<blank>でアノテーションする（12節を参照）。

一つの問題に複数の解答欄が現れている場合は、より選択肢の近くに現れる解答欄に対して<ansColumn>タグを付与する。

穴埋め問題では、空欄の下に解答欄が埋め込まれる場合があるが、英語問題では、<blank>内への<ansColumn>タグの埋め込みを禁止する。

# 選択肢

選択肢のアノテーションに使う要素は以下の通りである。

<choices> 選択肢のまとまり。

@anscol 対応する解答欄の識別番号。複数ある場合は、「A1 A2」のように、スペース

を挟んで複数記入する。

数学の場合は、大問番号（半角）の後に、対応する空欄の記号（全角カタカ

ナ）を続けたものを入れる。

（例：空欄が大問1のアの場合：「1ア」と入れる。複数の場合は、「2サ 2

シ」のようにスペースを挟んで記入）

@comment コメント（任意）

<choice> 個々の選択肢。

@comment コメント（任意）

@ansnum 選択肢の番号を半角数字で記述する。

<cNum> 選択肢の番号（記号の場合も含む）。

1. 数学の例。anscol属性の値と対応する空欄を赤字で示す。

<question id="Q1" minimal="no"><label>【１】</label>

（...中略...）

<instruction>

次の<ref target="B3">ケ</ref>&#65374;<ref target="B5">サ</ref>に当てはまるものを，下の①～④のうちから一つずつ選べ。ただし，同じものを繰り返し選んでもよい。（...中略...）

自然数nに関する条件p，q，rを次のように定める。

p：nは5で割ると1余る数である

q：は10で割ると1余る数である

r：nは奇数である

s：nは2より大きい素数である

また，条件rの否定を<formula />，条件sの否定を<formula />で表す。このとき 「pかつr」はqであるための<blank digits="1" id="B3">ケ</blank>。

<formula />は<formula />であるための<blank digits="1" id="B4">コ</blank>。

「pかつs」は「qかつs」であるための<blank digits="1" id="B5">サ</blank>。 </instruction>

<choices anscol="1ケ 1コ 1サ">

<choice ansnum="1">

<cNum>①</cNum> 必要十分条件である

</choice>

<choice ansnum="2">

<cNum>②</cNum> 必要条件であるが，十分条件でない

</choice>

<choice ansnum="3">

<cNum>③</cNum> 十分条件であるが，必要条件でない

</choice>

<choice ansnum="4">

<cNum>④</cNum> 必要条件でも十分条件でもない

</choice>

<br/>

</choices>

まれに、選択肢がテキスト内に埋め込まれている場合がある。

1. テキストに埋め込まれた選択肢。上は元のテキスト（地理B、p.146）



<choices anscol=”A31”>新田平岡集落の北を通る金山川に沿って<choice ansnum="1"><cNum>1</cNum><uText id=”U4”>氾濫原</uText></choice>がみられる。共栄集落は<choice ansnum="2"><cNum>2</cNum><uText id=”U5”>河岸段丘</uText></choice>にあり、その集落の<choice ansnum="3"><cNum>3</cNum><uText id=”U6”>約600m</uText></choice>南には丘陵地もみられる。共栄集落と新田平岡集落との標高差は<choice ansnum="4"><cNum>4</cNum><uText id=”U7”>約30m</uText></choice>あるため、共栄集落は新田平岡集落に比べて水利条件は良くなかった。</choices>

　各選択肢が表に埋め込まれている場合もみられる。その場合は、以下の(16)のようにタグ付けする。

1. 表の中に埋め込まれた選択肢。左下は元の表。



<choices anscol=”A3>

<tbl>

<row>

<cell/>

<cell >

<ref target=”b05”>ア</ref>

</cell>

<cell >

<ref target=”b07”>イ</ref>

</cell>

</row>

<row>

<choice ansnum="1">

<cell ><cNum>1</cNum></cell>

<cell >1.30</cell>

<cell >0.50</cell>

</choice>

</row>

<row>

<choice ansnum="2">

<cell ><cNum>2</cNum></cell>

<cell >1.30</cell>

<cell >1.0</cell>

</choice>

</row>

<choice ansnum="3">

<cell ><cNum>3</cNum></cell>

<cell >1.40</cell>

<cell >0.50</cell>

</choice>

</row>

........（以下略）

</tbl>

</choices>

　　選択肢が画像データに埋め込まれている場合は、<img>要素を利用して、画像データの情報を<choices>要素の下に記述する。画像の内部の個々の選択肢については現時点ではタグ付けをしない。

1. 選択肢が画像データに埋め込まれている例



<choices anscol=”A3”>

<img src=”filename.png"/>

</choices>

# 数式

　アノテーション対象のテキストファイルでは、数式、数値、座標、および変数の一部の情報が落ちている。それらに対しては、<formula>要素を使ってタグ付けする。

<formula> 数式が欠落した部分

@comment　コメント

　また、情報が欠落していない場合でも、変数と思われる表や数値は<formula>でタグ付けする。（第11節、第12節の例を参照）

# テキストの装飾

　上付き文字・下付き文字、漢文の送り仮名・返り点などに対しては、以下のタグを使用する。

<sup>　上付き文字

<sub>　下付き文字

<kaeri>　漢文の返り点：送り仮名との順序は、「返り点」→「送り仮名」とする。

<okuri>　漢文の右側に出現する送り仮名

<okuriL> 漢文の左側に出現する送り仮名

語と語をつなぐハイフンは厳密には返り点ではないが、ここでは返り点として扱う。但し、他の返り点と同時に出てきている場合は，ハイフンだけ異なる<kaeri>タグで囲む。  
他のタグ(lTextとか)との重なりが出てくる場合<okuri>，<kaeri>タグが一番内側に来るようにする。

　国語問題のルビについては、以下のタグを使用する。

<ruby> ルビ対象とルビの組

<rb> ルビを振る対象

<rt> ルビ

<ruby>タグは原則として文字単位で振る。<ruby>タグ内では<rb>と<rt>の数は一致しなければならない。

例:

<ruby><rb>寧</rb><rt>ねい</rt></ruby>

<ruby><rb>歳</rb><rt>さい</rt></ruby>  
  
<ruby>タグ内で返り点が出現することを許す。返り点は<rb>タグ内でのみ振り、<rt>には付与しない。  
ハイフン"-"で文字が結合されている場合はそれを一つのまとまりとしてルビを振る。返り点などは<rb>内に付与する。  
  
例:<ruby><rb>諳<kaeri>-</kaeri><kaeri>二</kaeri>知</rb><rt>あんち</rt></ruby>

　下線や傍線によって装飾されたテキストには<uText>をつける。また、下線や傍線はないが、ラベルの付いたテキストは<lText>でタグ付けする。

　UTF-8で表示できない文字があった場合は、<unk>タグを使ってその情報を記述する。対応する他の文字があればそれに置き換えて、置き換えた文字を<unk>タグで囲む。読みが分かっている場合は、カタカナでその読みを記述し、<unk>タグで囲む。

<unk> UTF-8で表示できない文字に対応する他の文字、あるいはカタカナによる読み

@imgsrc 漢字の画像がある場合、そのファイルのパス（任意）

@comment コメント（任意）

例："山﨑"の"﨑"がない場合  
山<unk comment="大の代わりに立">崎</unk>  
読みが分かっている場合はカタカナで読みを書く(<unk>サキ</unk>)  
  
　下線付きテキストやラベル付きテキストには、以下のタグを利用する。

<uText>　下線（傍線）付きテキスト

@id 識別番号（文書内で固有。「U」の後に半角数字をつけたものを入れる）

<lText>　ラベル付きテキスト（下線（傍線）なし。「A」や「ア」などの記号が付いた短いテキスト）

@id 識別番号（文書内で固有。「L」の後に半角数字をつけたものを入れる）

<label>　テキストにつけられたラベル

1. 傍線付きテキスト、ラベル付きテキストへのアノテーション例（国語、p.26）



<data id="D11" type="text">

　暁方に<uText id=”U40”><label>a</label>なる</uText>ままに、おびただしう吹きまさりたる風の紛れに、いと疾う内裏へ参り給ひぬ。「今宵は中宮の御宿直なりけるが、下りさせ給ひけるままに、上は藤壷にわたらせ給ふ」と聞こゆれば、....（中略）....かの雪の朝の御面影<uText id=”U41”><label>b</label>なる</uText>ものから、なほけしき異にて気高う、...（中略）...世の常に思ひ寄るべき御年のほどならねど、<uText id="u42"><label>（ア）</label>ただまぼり奉らまほしきに</uText>、「あはれ、...（中略）...げにと思したるさまにて、<uText id="U43"><label>（イ）</label>まめだち給へる御まみのわたり</uText>、見る我もうち笑まれて、...（中略）...口惜しうて歩み過ぎ給ふ。

　<lText id=”L01”><label>A</label> 宮城野にまだうら若き女郎花移して見ばやおのが垣根に</lText>

....（以下略）

</data>

上のように、下線付きテキストおよびラベル付きテキストにつけられたラベル（「ア」「a」「A」など）は、<uText>, <lText>の範囲に含める。

数学においては、ラベル付きテキスト<lText>は、主に数式に①や②などの番号が付き、後から参照されるような要素へのタグ付けにも使う。以下に例を示す。

元データ（PDF or WORD）

*a*，*b*を実数とし，*x*の二つの2次関数

 ①

 ②

のグラフをそれぞれ，とする。

テキストデータ（数式情報欠落）

　a，bを実数とし，xの二つの2次関数

①

②

のグラフをそれぞれ，とする。

アノテーション済みデータ

<formula>a</formula>，<formula>b</formula>を実数とし，<formula>x</formula>の二つの2次関数<br />

<lText id="L3"><formula /><label>①</label></lText> <lText id="L4"><formula /><label>②</label></lText> のグラフをそれぞれ<formula />，<formula />とする。

ただし、①や②のような番号がアノテーション対象データで欠落している場合もある。そのような場合は、以下のようにアノテーションする。

元データ（PDF or WORD）

*n*を整数とし，*x*の連立不等式



を考える。

　①の左辺は



と因数分解される。

テキストデータ（数式情報とラベル①、②が欠落）

nを整数とし，xの連立不等式

を考える。

　①の左辺は

と因数分解される。

アノテーション済みデータ（数式情報を<formula />で、ラベル①、②は直接復元）

<formula>n</formula>を整数とし，<formula>x</formula>の連立不等式<br /><lText id="L1"><formula />①</lText>

<lText id="L2"><formula />②</lText> <br />

を考える。<br />

<ref target="L1">①</ref>の左辺は<br />

<formula /><br />

と因数分解される。<br />

# 空欄

空欄のアノテーションには以下の要素を用いる。

<blank>　空欄

@id 識別番号（文書内で固有。「B」の後に半角数字をつけたものを入れる）

@digits 空欄に指定された桁数（\*理数系の問題で指定がある場合のみ）

<label>　空欄につけられたラベル

同じラベルを持つ空欄が複数現れており、同一のものを指示していると考えられる場合は、最初に出てきた空欄を<blank>でアノテーションし、残りの空欄は<ref>を付けtarget属性に最初の<blank>のidの値を指定する。

　穴埋め問題では、空欄の下に解答欄が埋め込まれる場合があるが、英語問題では、<blank>内への<ansColumn>タグの埋め込みを禁止する。

理数系の問題については、数式内に現れる空欄はアノテーションしない。

元データ（PDF or WORD）

を*x*軸方向に，*y*軸方向にも同じくだけ平行移動しても頂点は上にある。ただし，は0でない数とする。

テキストデータ（数式情報、空欄のラベル欠落）

をx軸方向に，y軸方向にも同じくだけ平行移動しても頂点は上にある。ただし，は0でない数とする。

アノテーション済みデータ（数式情報を<formula />で、空欄ラベル「二」は直接復元）

<formula />を<formula>x</formula>軸方向に<blank digits="1" id="B2">ニ</blank>，<formula>y</formula>軸方向にも同じく<ref target="B2">ニ</ref>だけ平行移動しても頂点は<formula />上にある。ただし，<ref target="B2">ニ</ref>は<formula>0</formula>でない数とする。<br />

# クロスリファレンス

　クロスリファレンスには以下の要素を用いる。

<ref>　参照要素。

@target 参照先の要素のidの値をとる。複数指定可。

@comment コメント（任意）

参照先の要素が画像内に埋め込まれているなどの理由でidを指定できない場合は、target属性は空にし、comment属性に「画像内」などのコメントを入れる。

<ref>は以下の要素への参照を含む。

* データ <data>
* 下線付きテキスト <uText>
* ラベル付きテキスト <lText>
* 注釈 <note>
* 空欄 <blank>
* ページ <pageStart> （国語問題のみ。pn属性の値を記述）
* 形式段落 <paragraph>（国語・英語問題のみ。prn属性の値を記述）
* 行 <lineStart> （国語・英語問題のみ。ln属性の値を記述）

# 文字化け・欠落テキストの修正

アノテーション対象ファイルに文字化けがある場合は、テキストを直接修正する。アノテーション対象ファイルに欠落したテキストがある場合も、直接テキストを追加する。数式の欠落部分はテキストを追加した後、<formula>でタグ付けする。ただし、複雑な数式は空のまま<formula/>で置き換える。図に埋め込まれたキャプション、画像になっている表の中のテキスト、画像になっている漢文も復元し、元ファイルに追加する。

# レイアウト情報

<br/> 元テキストの改行位置を示す

　以下、国語問題のみ、ページの開始位置と終了位置を示すタグを入れる（国語の問題で、「～ページの記述について」という指定のある問題が散見されるため。）

<pageStart/> ページの開始位置

@pn ページの番号（文書内で固有。「P」の後に半角数字をつけたものを入れる）

<pageEnd/> ページの終了位置

　また、国語あるいは英語の長文問題のうち、形式段落および行番号が明記されている場合に限り、以下のタグを使って段落および行の開始位置をタグ付けする。（行の場合、5行おきに番号が振られている場合もあるが、その場合は全行を<lineStart/>でタグ付けする）

<paragraph> 形式段落の範囲

@prn 段落の番号（文書内で固有。「PR」の後に半角数字を付けたものを入れる）

<lineStart/> 行の開始位置

@ln 行番号（文書内で固有。「LN」の後に半角数字を付けたものを入れる）

# 参考文献

The TEI Consortium (2010) TEI P5: Guidelines for Electronic Text Encoding and Interchange.