E-dictionaries 交換フォーマットの国際標準化

International Standardization of Interchange Format for E-dictionaries

齋鹿　尚史†　植村　八潮‡　小町　祐史††　永田　健児‡‡

Hisashi SAIGA†　Yashio UEMURA‡　Yushi KOMACHI††　and　Kenji NAGATA‡‡

†シャープ株式会社 研究開発本部 プラットフォーム開発センター 第三開発室
† Development Dept. III, Platform Technology Center, Corporate R&D Group, SHARP Corporation
‡東京電機大学出版局　‡ Tokyo Denki University Press
††大阪工業大学 情報科学部　†† Faculty of Information Science And Technology, Osaka Institute of Technology
‡‡株式会社ディジタルアシスト　‡‡ Digital Assist, Ltd.
E-mail: † saiga.hisashi@sharp.co.jp, ‡ yashio@jim.dendai.ac.jp, †† komachi@y-adagio.com,
‡‡ nagata@d-assist.com

1. はじめに

IEC TC100/TA10 および，その日本における国内委員会である JEITA E-Book 標準化 G では，電子書籍関連の国際規格の策定を行っている．

その一環として，日本から提案され，2009 年 3 月に CD(Committee Draft; 委員会原案)[1]が発行された，E-dictionaries(電子辞書)交換フォーマットの国際標準化について述べる．

2. 電子辞書の交換フォーマットの標準化

2.1 電子辞書交換フォーマットの役割

電子辞書を含む電子書籍コンテンツが，出版社，コンテンツプロバイダを経由して，エンドユーザによって閲覧されるまでには，いくつかの段階がある．

IEC/TS 62229 [2] では，そのモデル化である Contents creation/distribution model が定義されており，後続の電子書籍フォーマット関連の標準化においてはこれを参照して，どの部分に対応するのかを示すことが行われている．

今回の標準化でも，基本的にこのモデルを踏襲しているが，電子辞書のハードウエア/ソフトウエアのメーカー(manufacturer)の役割が，通常の電子書籍における Publisher[1]と役割が重なっているという考察から，図 1 に見られるように，Publisher の入る個所に manufacturer を加えている．

今回の標準化の対象である，交換フォーマットは，図 1では Data preparer と Publisher(manufacturer)の間でのデータ交換に用いられるフォーマット（図 1の(2)）ということになる．

Contents creation/distribution model における位置は，電子書籍用の Generic format(記述フォーマット)の規格である IEC62448[3]と類似しているが，辞書コンテンツに特化した機能を多く備えている点に相違がある．

2.2 標準化検討の重要性および経緯

電子辞書の専用機の市場が近年大きく伸びており[2]，コンテンツの供給も増加している．しかし一方では，コンテンツ供給側と，それを搭載した端末を供給するメーカーとの間での交換に用いられるデータフォーマットが統一されていないという問題があった．

Author <--(1)--> Data preparer <--(2)--> Publisher(manufacturer) --(3)--> Reader

図 1 Contents creation/distribution model

[1] [2] では“organization or person that issues and distributes an e-book” と定義されており，日本語の「出版社」と同一の概念ではないことに注意．

[2] JBMIA[4]によれば，2007 年の日本メーカの出荷台数は約 300 万台，前年比 111.1%となっている．

すなわち，異なるフォーマットによる辞書コンテンツが混在しているため，出版社やコンテンツごとに異なるツールや作業を行って，電子辞書端末用のデータに変換する必要があった．これがチェックのための作業も含め，コンテンツの作成コストを上昇させる要因となっていた．

したがって，電子辞書の交換フォーマットの標準化の重要性は早くから認識されており，2005 年の JEITA E-Book 標準化 G の発足時から議論がなされ，JEPA(Japan Electronic Publishing Association; 日本電子出版協会)との意見交換も開始されていた．

このような背景の下，2007 年 10 月の TC100/TA10 のアルザス会合にて，日本側から提案を行った結果，NP（New work item Proposal; 新作業課題提案）の提出が要請された．これを受けて，E-Book 標準化 G での検討が本格化し，NP[5]が 2008 年 5 月に発行された[3].

NP[5]では，フォーマットの具体的な定義は行わず[4]，その内容は2.1，2.3および2.4で述べるような一般論となっている．

2.3 要求される条件(Requirements)

電子辞書交換フォーマットが持つべき機能について，NP[5]では，以下を挙げている[5].

a) キーワードの記述，順序の記述およびキーワードとエントリ（各見出し語の定義）との間のリンク (Keywords and their order, Link data)
b) エントリの記述(テキスト，画像，マルチメディア機能を含む) (Entry Data)
c) 書誌（著者名，題名など），その他の情報（凡例など）(Bibliographical data, etc.)
d) いろいろな言語で記述されたコンテンツの記述

特に，a)b)c)の関係を模式的に表しているのが図 2 である．

d)については，辞書が多くの場合複数の言語からなり（たとえば独和辞典であればドイツ語，日本語というように），しかもその対象言語としては多様なものがあり得ることから，自然な要求と考えられる．

2.4 スコープに含まれない項目について

[1]の第 1 章”Scope”では，取り扱わない問題として，下記を列挙している[6].

e) 閲覧装置のためのデータフォーマット
f) 印刷にのみ関係する要素
g) 物理的デバイスに関連するレンダリングの問題
h) DRM などのセキュリティの問題

このような問題は，いずれもそれ自体では重要であるが，電子辞書交換フォーマットの標準化で取り扱うにはふさわしくないと考えられるものである．

e)が今回の標準化の対象でないのは，2.1ですでに説明した通りである．

3. 規格原案

3.1 規格原案作成の方針

NP[5]発行後は，作業は具体的なフォーマット（以下「標準化フォーマット」と呼ぶ）を記載した規格原案の作成に移った．(NP[5]は，2008 年 9 月に各国の投票を通過し，採択された[7]．)

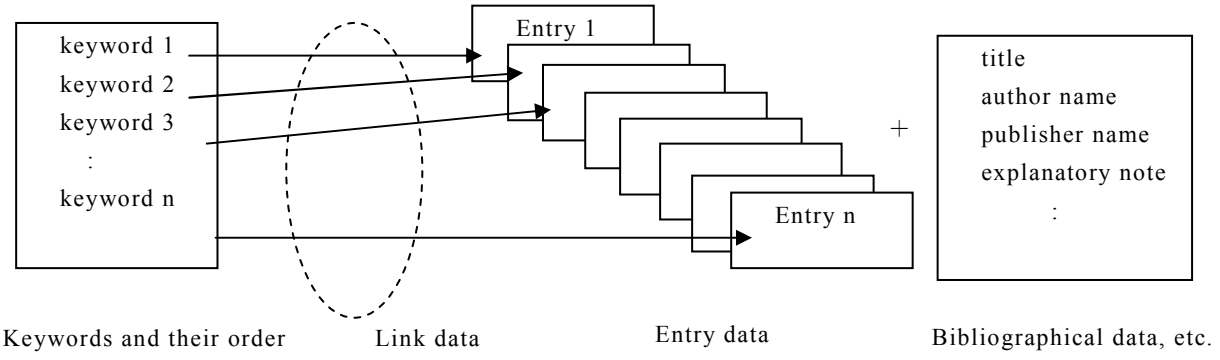

図 2 各要素の関係模式図

[3] 著者の一人（齋鹿）がプロジェクトリーダとなっている．
[4] 「XML で表現される．詳細は今後決定.」と記載．
[5] a)からd)の記号はこの稿のために付けた．
[6] e)からh)の記号はこの稿のために付けた．
[7] これにより，正式に IEC TC100/TA10 のプロジェクトとなる．

今回の規格原案作成にあたっては，2.2で述べたような経緯を踏まえて，JEITA E-Book 標準化 G と JEPA とで連携しつつ進めることとした．

JEPA との協議の結果，日本において，辞書の電子化の分野で比較的普及している，株式会社ディジタルアシストの LeXML のフォーマットをベースとすることとした[8]．(LeXML の仕様については，ディジタルアシスト[6]参照．)

一方，LeXML に含まれていないか，強化が必要な機能（書誌，全体データ構造，マルチメディア機能，テキスト機能など）については，IEC62448[3] Annex B[9]をベースに仕様を策定することで，出版社にとって使いやすい標準化フォーマットの策定を目指すとともに，従来の電子書籍フォーマットとの親和性にも配慮している．

また，今回の標準化では，単に LeXML に IEC62448[3] Annex B をマージするのではなく，規格が長く実用に耐えるよう，後述するような拡張・改良を行っている．

このようにして定義された標準化フォーマットは，後に具体例を示すように，キーワードやその順序の記述，エントリとのリンクを記述できることにより，2.3で述べた要求される条件のa)を充たしている．

また，マルチメディア機能を含むエントリの記述をカバーすることで，要求される条件のb)を充たしている．

さらに，前述のように，書誌データやテキストデータについても機能を充実させたことで，要求される条件のc)も充たしている．

要求条件のd)に対する対応については後述する．

3.2 拡張と改良

3.1で触れた改良と拡張について，ここでは全てを紹介できないが，以下に例を挙げる．

1) Text-to-speech への対応

text-to-speech に送られる文字列を指定する<tts>タグを新設している．

2) 環境によって異なる記述を有効とする

異なる環境向けのソースファイルを共有しつつ，異なる環境では異なる表示としたいという要求を満足するため，環境によって異なる記述を有効とするための<select>タグを策定した．その記述例を図 3に示す．

図 3の記述例は，PC ではルビを表示するが，それ以外の環境（デフォルト）では，漢字の後のかっこで読みを示すということを表現している．

```
<select>
<select_item default="yes">
夏目漱石（なつめそうせき）
</select_item>
<select_item type="PC">
<ruby>
<rb>夏目漱石</rb>
<rt>なつめそうせき</rt>
</ruby>
</select_item>
</select>
```

図 3 記述例

3) 言語コードの指定

コンテンツで用いられる言語[10]を部分ごとに指定することができる（3.3の記述例も参照）．また，その際に用いるコード体系自体を指定することが可能である．（デフォルトは ISO 639-3[7]とする．）

IEC62448[3] Annex B の方針を受け継ぎ，文字セットや禁則文字などは，各言語ごとのローカライゼーションにゆだね（規格では規定しない），自由に指定[11]できるようにしている．

これにより，2.3で述べた要求される条件のd)を充たしている．

4) フォントファミリの指定

フォントファミリとして，表 1にあるものを使用可能としている．

フォントを指定するタグ（複数ある）の family 属性の値として指定される．

表 1 使用可能なフォントファミリ

値	意味
monospace	固定幅フォント
san-serif	ゴシック体など
serif	明朝体など
cursive	筆記体
fantasy	装飾系

なお，標準化フォーマットでは，2.4でも述べた通り，実際に上記の指定がどのようにレンダリングされるかまでは規定していない．

[8] 国際標準化のため，日本国外に通用しにくいタグ名など，仕様の見直しは行っている．

[9] シャープ株式会社による電子ドキュメント閲覧技術である，XMDF の記述フォーマットがベースとなっている．

[10] 文字コードセットの指定とは別である．

[11] エンコーディングについては，UTF-8 または UTF-16 を[1]では推奨している．

3.3 記述例

参考までに図 4に，辞書コンテンツの記述例を挙げる．図 4で使われているのは基本的な要素・属性のみである[12].

<dict_body>タグ，<dict_data>タグは全体構造のために導入されたタグである．<dict_default_attribute>タグはフォントや背景色などデフォルトの設定値を示す[13].

見出し語ごとに情報をまとめているのが<dic-item>タグであり，見出し語ごとに１つずつ存在する．

<headword>タグが各単語ごとに表示される文字列（見出し語，発音記号など）を，<key>タグが検索のために入力される文字列を示す．これが，キーワードとエントリとのリンクを指定していることになる．

図 4の例では，"standard", "stay" の 2 単語が並んでいる．"standard"について「標準.」と「知名度が高く，良く歌われる歌曲.」の 2 つの語義が<meaning>タグによって記述されている．

table_id 属性は，検索テーブルに各見出し語を登録する際，登録先の検索テーブル[14]を識別するために導入されているものである．

なお，検索テーブル上でのキーワードの順序についても，XML 上での記述順とは別に，指定することもできる[15].

<psp>タグは品詞を，<meaning>タグが語義を，<example>タグが用例をあらわす．図 4の例では，"standard" の英・日の例文のため，タグの lang_code 属性によって言語コードを切り替えている[16].

3.4 今後の作業

次のステップは，CDV(Committee Draft for Vote; 投票用委員会原案)に向けた仕様のブラッシュアップである．その一環として，出版社など関係者に対する説明を行い，さらなるニーズの汲みあげを図る予定である．

その後，CDV に対する各国投票を通過すれば，国際規格となる．

4. まとめ

電子辞書交換フォーマットの国際標準化について，その内容を経緯も含めて紹介した．

```
<dict_data>
<dict_default_attribute>
…
</dict_default_attribute>
<dict_body>
<dic-item>
<head>
<headword table_id=“ST0001”>
standard
</headword>
<key>standard</key>
<headword type=“pronunciation”>
stændəd
</headword>
</head>
<psp>名詞</psp>
<meaning level="1" no="1">
標準.
</meaning>
<example>
<span lang="eng" lang_code ="ISO 639-3">
He set a standard.
</span>
<span lang="jpn" lang_code="ISO 639-3>
彼がお手本となった.
</span>
</example>
<meaning level="1" no="2">
知名度が高く，良く歌われる歌曲.
</meaning>
</dic-item>
<dic-item>
<head>
<headword table_id=“ST0001”>
stay
</headword>
        :
</dic-item>
        :
</dict_body>
</dict_data>
```

図 4 記述例

[12] 以下の説明は，図 4に出てくる全ての要素・属性の説明を意図したものではない.
[13] ここでは詳細は省略する.
[14] 検索テーブルは必須ではない.
[15] ここでは詳細は省略する.
[16] 言語の指定は，この場合必須ではない.

今回の標準化においては，出版社側(JEPA)との連携による仕様策定が大きな特徴となっている．

今後の標準化の進捗については審議状況に大きく依存するが，このような推進方法によって，国際規格を使う立場からのニーズを効率よく汲みあげることができ，実効性の高い規格となることが期待される．

5. 謝辞

今回の標準化の過程で議論，協力頂いている，IEC TC100/TA10，JEITA E-Book 標準化 G のメンバーの皆様，株式会社ディジタルアシスト，シャープ株式会社の皆様に感謝致します．

文　献

[1] 100/1532/CD, 2009

[2] IEC/TS 62229 Multimedia systems and equipment - Multimedia e-publishing and e-book - Conceptual model for multimedia e-publishing, 2006

[3] IEC62448 Edition 2.0 Multimedia systems and equipment - Multimedia e-publishing and e-books - Generic format for e-publishing, 2009

[4] http://www.jbmia.or.jp/MoBS/market/jisyo-data090226.pdf,
ビジネス機械・情報システム産業協会(JBMIA), 2009

[5] 100/1400/NP, 2008

[6] http://www.d-assist.com/lexml200.pdf,
ディジタルアシスト, 2007

[7] ISO 639-3, 2007