言語資料の電子化に関する諸問題：特に用語法に関して *

高橋 洋成
（筑波大学）
takahashi.yona.gp@u.tsukuba.ac.jp

1 はじめに

現在まで、言語資料を電子化して研究者の間で共有しようとする試みは数多く行われている。言語資料を電子化するメリットとして挙げられるのは、第 1 に物理的にかさばる冊子体を持ち歩かずに済むという保存性、第 2 に収集したデータをその場で公開あるいは修正できるという速報性、第 3 に目的のデータを素早く見つけることができるという検索性、第 4 に大量のデータを様々な形で処理することにより新たな知見を得られる可能性である。近年のインターネット（特に World Wide Web）の普及により、保存性・速報性・検索性は世界規模のものとなり、現在は言語研究の新たな方法論が生まれつつある段階と言えよう。

その一方で、言語資料の電子化には技術的な論点のみならず、これまでの言語研究の蓄積に基づいて「何を、どのように」電子化するかを積極的に発言すべき言語学固有の問題がいくつかある。本稿は言語資料の電子化の現状と問題点を概観し、今後の方向性を模索する。

2 言語資料の電子化の事例

2.1 BNC

British National Corpus (BNC)[1] は、20 世紀後半からのイギリス英語における話し言葉と書き言葉を網羅的に収集するプロジェクトであり、2007 年に公開された版では約 1 億語が含まれている [2]。このコーパスは Text Encoding Initiative (TEI)[3] に準拠した BNC XML 形式で電子化されている。しかしながら、現版の

*本調査は平成 22〜25 年度科学研究費基盤研究 (B)「変容するエチオピア諸言語の静態と動態に関する総合的研究、ならびにデータベース構築」代表：柘植洋一（金沢大学）（課題番号 22401046）によるものである。

[1] http://www.natcorp.ox.ac.uk/

[2] 90% は新聞・雑誌・書籍・教科書・手紙等から収集された書き言葉であり、10% が会議・ラジオ番組等から収集された話し言葉である。

[3] 人文学資料を電子化する際の標準形式およびガイドラインになることを目指して策定されている。

TEI には言語分析のためのタグセットは含まれていない。そのため、BNC XML では文法のアノテーションとして「文法タグ」を独自に作成し、文法の観点から資料の検索を可能にしている[4]。

例えば、次のような例文を考える。

> Apparently we eat more chocolate than any other country.

これに BNC XML に基づく文法タグを付けると、次のような形になる。

```
<s>
  <w c5="AV0" hw="apparently" pos="ADV">Apparently </w>
  <w c5="PNP" hw="we" pos="PRON">we </w>
  <w c5="VVB" hw="eat" pos="VERB">eat </w>
  <w c5="DT0" hw="more" pos="ADJ">more </w>
  <w c5="NN1" hw="chocolate" pos="SUBST">chocolate </w>
  <w c5="CJS" hw="than" pos="CONJ">than </w>
  <w c5="DT0" hw="any" pos="ADJ">any </w>
  <w c5="AJ0" hw="other" pos="ADJ">other </w>
  <w c5="NN1" hw="country" pos="SUBST">country</w>
  <c c5="PUN">.</c>
</s>
```

ここでは、文が s 要素、語が w 要素、文字が c 要素とマークアップされている。w 要素は、見出し語形（headword）を示す hw 属性と、品詞（part of speech）を示す pos 属性を持つ。また、w 要素と c 要素は CLAWS5[5] に準拠したアノテーションを c5 属性によって埋め込むことができる[6]。上記例で使われている CLAWS5 アノテーションの意味は次の通りである。

AJ0	形容詞	AJC	形容詞（比較級）
AJS	形容詞（最上級）	AV0	副詞
AVP	副詞（小詞）	AVQ	副詞（疑問）
CJC	接続詞（等位）	CJS	接続詞（従属）
DT0	限定詞	DTQ	限定詞（疑問）

[4] http://www.natcorp.ox.ac.uk/docs/URG/posguide.html

[5] http://ucrel.lancs.ac.uk/claws5tags.html

[6] Constituent Likelihood Automatic Word-tagging System (CLAWS) は、主に英語コーパスを機械処理するためのタグ付け方法として 1980 年代に開発され、現在は CLAWS7 まで公開されている。ただし、現在の BNC に組み込まれているのは CLAWS5 であることに注意されたい。

NN0	普通名詞（単複同形）	NN1	普通名詞（単数）
NN2	普通名詞（複数）	PNI	代名詞（不定）
PNP	代名詞（人称）	PNQ	代名詞（疑問）
PRP	前置詞	PUN	句読点
VVB	語彙的動詞（基本形）	VVD	語彙的動詞（過去形）
VVG	語彙的動詞（-ing 形）	VVI	語彙的動詞（不定形）
VVN	語彙的動詞（過去分詞）	VVZ	語彙的動詞（-s 形）

さて、これらの文法タグセットは POS (Part of Speech) Tagging とも呼ばれているように、基本的に英語の品詞分類である。この枠組みでは、接辞など形態論的な内部構造に踏み込む分析や検索はできない。また、句読点という言語の表記に関するものと、品詞という言語の構造に関するものが同一に扱われていることは、文字検索と言語検索とを分けたい場合に問題になる可能性がある。このように、言語資料を電子化した際、どのような検索や分析が可能になるかは、電子化する時点でのアノテーション方法に大きく左右される。

言語資料の電子化はすでに各地で進められており、それぞれの目的に応じたアノテーションが付されている。次の段階として、電子化された言語データの相互運用性を確保する手段を考える必要があろう。具体的には次のようなことである。

- 異なる言語データを横断的に意味検索する場合の「意味の関連性」の問題。例えば、日本語の「魚」と英語 “fish” が意味的に対応しうること（あるいは逆に、意味の差異について）をデータベースに教え、適切に扱えるようにする方法はないか。
- 異なる言語データを用いて、横断的に言語学的を概念を検索する場合の「文法範疇」の問題。例えば、ある言語データで「名詞 noun」とされているものと、別の言語データで「名詞類 nominal」とされているものを、同じと見なして良いか。
- 異なる言語データを用いて、横断的に言語構造を検索する場合の「文法構造」の問題。例えば、「他動詞の直後に目的語が置かれる統語構造」のようなものを探す場合、「他動詞」「目的語」などの概念をどのように統一的に処理できるか。

以上のような問題を解決し、すでに存在する言語データの再利用性を高めていく必要性が、今後は一層高まるであろう。

2.2 LLOD

2004 年、学術情報や公共情報等をオープンに提供し、電子データを共有することを支援する非営利団体 Open Knowledge Foundation (OKFN) が発足した[7]。そこに属する部会には、政府情報を扱う部門、移動路線に関する情報を扱う部門、科学技術情報を扱う部門、経済情報を扱う部門、書誌情報を扱う部門、建築技術情報を扱う部門等がある。

そして 2010 年、OKFN の部会として 'Open Linguistics' を目指す OKFN's Working Group on Open Data in Linguistics が発足した[8]。その目標は電子的な言語学データが有機的なネットワークとして機能すること（LDL: Linked Data in Linguistics）であり、具体的な形としては Linguistic Linked Open Data Cloud (LLOD) として図 1 にまとめられている。

LLOD は大きく 3 つのデータ群から成り立っている。

- **ナレッジベース。**言語に関するメタデータや、言語分析の枠組み･用語法･アノテーションモデルといった「言語データの作り方」に関する情報。POWLA[9]、Lemon[10]、Glottolog/Langdoc[11]、GOLD、ISOcat、OLiA[12] など。
- **アノテーション付きコーパス。**ナレッジベースに基づき、アノテーションを付された言語コーパス。BNC、Project Gutenberg、The Open Library など。
- **辞書・シソーラス。**語彙の意味を分類し、機械的に意味の関連性を推測可能にするための枠組み。DBpedia[13]、YAGO、WordNet など。

本稿ではこれらの全てに触れることができないため、以下では特に、言語分析に関わる用語を表現するための枠組みである GOLD、ISOat および OLiA について紹介する。

3 電子的な言語分析のための方法

3.1 GOLD, ISOcat, OLiA

General Ontology for Linguistic Description (GOLD) は Farrar and Langendoen (2003) で公開され、現在も改良が進められている[14]。これは言語コーパスに電子的な文

[7] http://okfn.org/

[8] 2011 年に Open Linguistics Working Group (OWLG) となった。

[9] コーパスを RDF/OWL で表現するための枠組み。

[10] Simple Knowledge Organization System (SKOS) を語彙意味論の記述に利用するための枠組み。

[11] 言語系統や方言などを languoid を用いて表現するための枠組み。Ethnologue でも採用されている。

[12] GOLD、ISOcat、OLiA は言語学用語、特に言語分析に関わる用語を表現するための枠組み。これらは第 3 節で詳しく述べる。

[13] Wikipedia を RDF/OWL で表現し、語彙の意味を機械可読にするためのプロジェクト。

[14] http://www.linguistics-ontology.org/

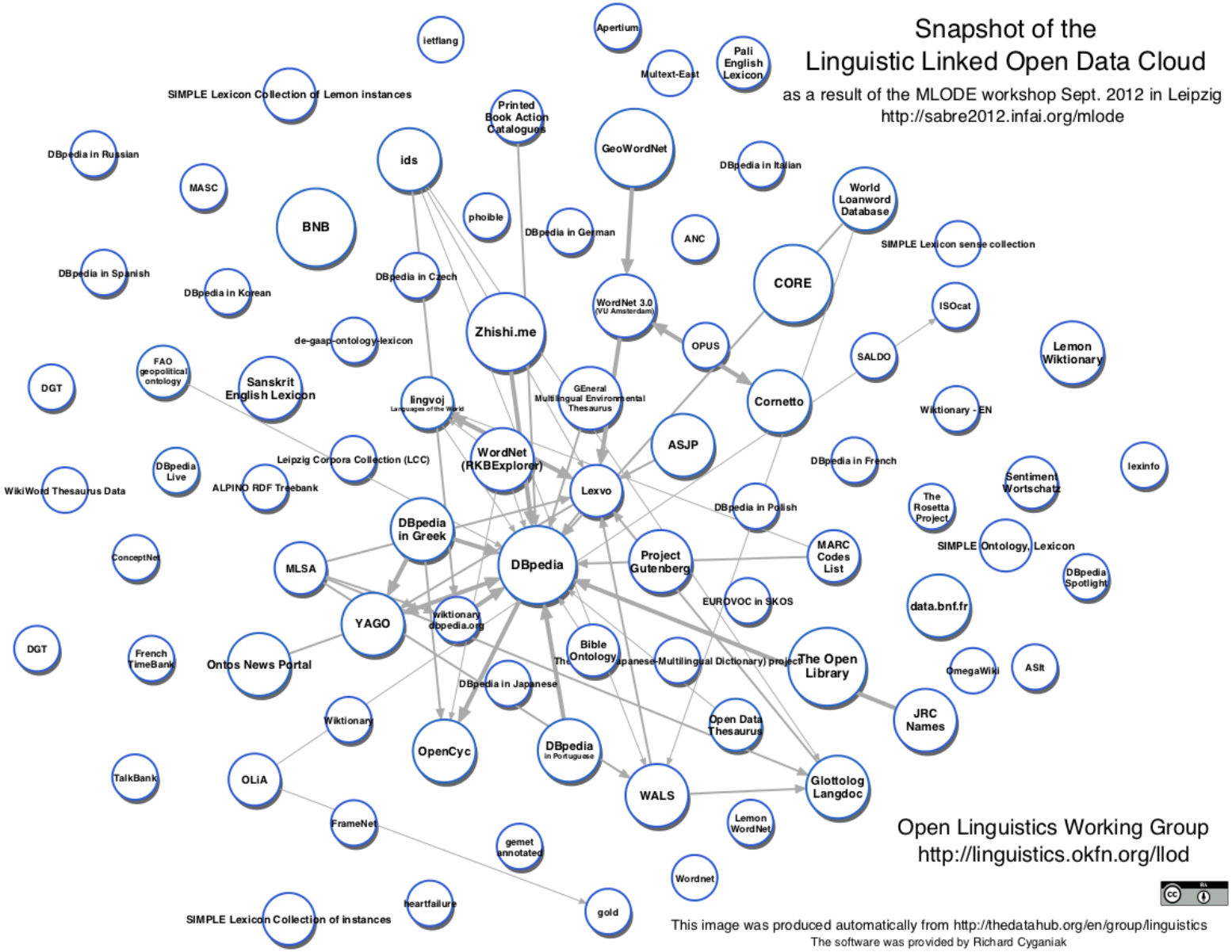
Snapshot of the
Linguistic Linked Open Data Cloud
as a result of the MLODE workshop Sept. 2012 in Leipzig
http://sabre2012.infai.org/mlode
Apertium
ietflang
Multext-East
Pali
English
Lexicon
SIMPLE Lexicon Collection of Lemon instances
Printed
Book Action
Catalogues
GeoWordNet
DBpedia in Russian
ids
DBpedia in Italian
World
Loanword
Database
MASC
BNB
phoible
DBpedia in German
ANC
SIMPLE Lexicon sense collection
CORE
DBpedia in Spanish
DBpedia in Korean
DBpedia in Czech
WordNet 3.0
(VU Amsterdam)
ISOcat
Zhishi.me
de-gaap-ontology-lexicon
OPUS
SALDO
Lemon
Wiktionary
FAO
geopolitical
ontology
DGT
Sanskrit
English Lexicon
lingvoj
Languages of the World
GEneral
Multilingual Environmental
Thesaurus
Cornetto
Wiktionary - EN
ASJP
WordNet
(RKBExplorer)
DBpedia in French
lexinfo
DBpedia
Live
Leipzig Corpora Collection (LCC)
Lexvo
Sentiment
Wortschatz
WikiWord Thesaurus Data
ALPINO RDF Treebank
The
Rosetta
Project
DBpedia in Polish
DBpedia
in Greek
SIMPLE Ontology, Lexicon
ConceptNet
MLSA
DBpedia
Project
Gutenberg
MARC
Codes
List
DBpedia
Spotlight
EUROVOC in SKOS
data.bnf.fr
YAGO
wiktionary
dbpedia.org
DGT
French
TimeBank
Ontos News Portal
Bible
Ontology
The Open
Library
OmegaWiki
ASIt
DBpedia in Japanese
Wiktionary
Open Data
Thesaurus
JRC
Names
OpenCyc
DBpedia
in Portuguese
TalkBank
OLiA
WALS
Glottolog
Langdoc
FrameNet
Lemon
WordNet
gemet
annotated
Wordnet
heartfailure
gold
SIMPLE Lexicon Collection of instances
Open Linguistics Working Group
http://linguistics.okfn.org/llod
This image was produced automatically from http://thedatahub.org/en/group/linguistics
The software was provided by Richard Cyganiak

図 1: LLOD の概念図

法タグを付与するにあたり、前もって言語学的用語を網羅しておき、用語同士の意味関係を記述するものである。現在、その成果は危機言語のデータを電子化する Electronic Metastructure for Endangered Language Data (EMELD)[15] や、IEEE[16] で言語学的データを通信するための規格 Standard Upper Merged Ontology (SUMO)[17] などに反映されている。

一方で、自然言語処理（NLP）や機械可読辞書等の分野で使用されている電子的アノテーションも様々であり、ISO にはそうした言語処理に関する技術をまとめる技術委員会 TC37 が設置されている。2004 年、ISO/TC37 の下で、言語学的用語法を網羅的に収集する分科会 SC4 が発足した。TC37/SC4 は数々の先行プロジェクトの成果をもとに、収集された電子的用語法を ISOcat というレポジトリに公開している [18]。GOLD も ISOcat に取り入れられている。

さらに、Chiarcos (2008) から徐々に進められてきた OLiA (Ontologies of Linguistic Annotation)[19] プロジェクトも、GOLD と ISOcat との連携を密にしながら、記述言語学や自然言語処理における用語法を独自にまとめている。

現在まで、これらに収録された言語学用語は膨大な数にのぼり、本稿で一つ一つを検討することはできない。そこで、最も基本的と思われる GOLD のアノテーションモデルに的を絞り、その構成と使い方を紹介する。

3.2 GOLD のアノテーションモデル

GOLD に収録されている言語学用語の数は膨大であるが、大まかな構造を示すため、現行のバージョンにおける主な範疇・分類法を以下に抜粋する。各範疇に属している電子的用語の正確な分類位置、およびそれがどのような言語で確認されるものかについては、GOLD の Web サイトで確認されたい [20]。

- 言語学的特性
  - 音声学的特性
    調音特性、音響特性
  - 形態意味論的特性
    テンス特性、アスペクト特性、ムード特性
  - 形態統語論的特性

[15] http://emeld.org/
[16] The Institute of Electrical and Electronics Engineers, Inc.
[17] http://suo.ieee.org/SUO/SUMO/
[18] http://www.isocat.org/
[19] http://nachhalt.sfb632.uni-potsdam.de/owl/
[20] http://www.linguistics-ontology.org/gold.html

ケース特性、モダリティ特性、ヴォイス特性、フォース特性、エヴィデンシャリティ特性、エヴァリュアティブ特性、ジェンダー特性、ナンバー特性、サイズ特性、人称特性、ポラリティ特性

  - POS 特性
    述語子（動詞類、形容詞類、副詞類）、機能子（接続詞類、置詞類）、限定子（冠詞）、名詞（普通名詞、固有名詞、動名詞）、代形類（代名詞、指示詞、代形容詞、代副詞）、分類子、不変化詞、量化子（数詞類）、否定子、分詞
  - 音韻論的特性、意味論的特性、談話論的特性
    未定

- 言語学的単位
  - 形式的単位
    字素、分節（子音、母音）、超分節的、音節、音調素、脚韻、モーラ、音素
  - 意味的単位
    語彙概念
  - 文法的単位
    * 形態素
      拘束形態素（派生形態素・屈折形態素・過程的・重複的）、接語、複合語、語根、語幹、接辞
    * 統語的単位
      統語構造（句・節）
    * 統語的語
  - 談話的単位
    未定

- 言語学的体系
  - 人間言語種別
    確認済種、記述済種、絶滅種、絶滅危惧種、第二言語種、音声言語種、口語変異種、手話言語種、文字言語種

- 言語学的分類

  - 系統的分類
    語族、語群、方言、孤立的
  - 地理的分類、政治的分類、使用層分類
  - 文字、正書体系、音韻論的体系
- 言語学的表現
  - 文字言語表現、手話言語表現、音声言語表現

さて、GOLD は RDF/OWL[21] を用いて定義されている。したがって、あるデータベースで独自の言語分析の枠組みと用語法を用いていたとしても、それを GOLD にマッピングすることで、データベースを横断的に検索可能になりうる。例えば、ある聖書ヘブライ語コーパスのデータベースで PrefixConjugation（接頭辞活用）、SuffixConjugation（接尾辞活用）という形態論的特徴を扱っていたとしよう。これらの用語について、以下では「PrefixConjugation は GOLD における PastTense と NonPastTense もしくは PerfectiveAspect を包括するもの」、また「SuffixConjugation は GOLD における VerbalAdjective もしくは PerfectiveAspect である」と定義している。

```
<owl:Class rdf:about="http://example.org/terms1/PrefixConjugation">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <owl:Class
          rdf:about="http://purl.org/linguistics/gold/PastTense"/>
        <owl:Class
          rdf:about="http://purl.org/linguistics/gold/NonPastTense"/>
        <owl:Class
          rdf:about="http://purl.org/linguistics/gold/ImperfectiveAspect"/>
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>
<owl:Class rdf:about="http://example.org/terms1/SuffixConjugation">
  <owl:equivalentClass>
    <owl:Class>
      <owl:intersectionOf rdf:parseType="Collection">
        <owl:Class
          rdf:about="http://purl.org/linguistics/gold/VerbalAdjective"/>
        <owl:Class
          rdf:about="http://purl.org/linguistics/gold/PerfectiveAspect"/>
```

[21] セマンティック・ウェブの基盤技術で、Web 上あるいは非 Web におけるリソースの関係を記述し、機械的な推論でリソース間のネットワークを形成するための枠組みである。

```
      </owl:intersectionOf>
    </owl:Class>
  </owl:equivalentClass>
</owl:Class>
```

一方、別のデータベースではそれらを Imperfect、Perfect と名付けていたとする。この場合も同じようにして GOLD の語彙にマッピングする。こうすることで図 2 のように、異なる言語分析の枠組みで作られた 2 つの聖書ヘブライ語データベースを、互いに横断しながら利用することが可能になる。

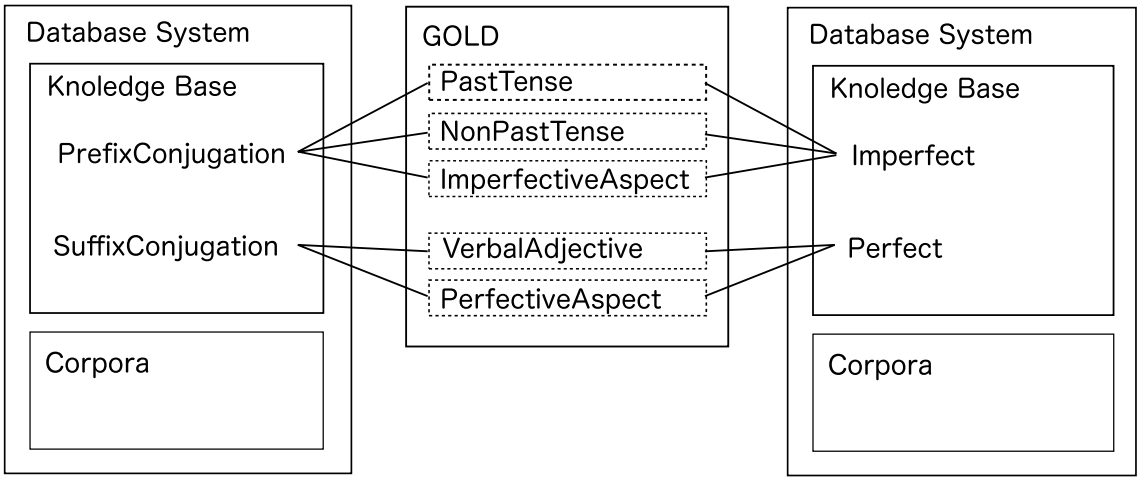

図 2: 各データベースの用語法を GOLD へマッピング

4 おわりに

言語資料の電子化は、様々な目的や切り口、方法論のもとに行われている。そうした既存のプロジェクトと連携しつつ、横断的な処理を可能にするための枠組み（あるいはガイドライン）の構築が現在の急務と言えよう。本稿では OWLG が進めている LLOD プロジェクト、およびその基盤技術の中でも言語分析に関わる OLiA、ISOcat そして GOLD を中心に、その目的と使い方を簡単に紹介した。これらの電子的言語研究が今後どのような方向に進んでいくべきか、エチオピア諸語研究、セム諸語研究、アフロ・アジア諸語研究、そして日本語研究など、多くの言語研究の立場からの積極的な提案が求められている。

【参照文献】

Chiarcos, C. (2008) “An Ontology of Linguistic Annotations,” *LDV Forum* 23/1, 1-16.

Chiarcos, C. et al. (2012a) *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata,* Heidelberg: Springer.

Chiarcos, C. et al. (2012b) “Towards a Linguistic Linked Open Data cloud: The Open Linguistic Working Group,” *Traitement automatiques des langues* 52, 245- 275.

Farrar, S and D. T. Langendoen (2003) “A Linguistic Ontology for the Semantic Web,” *GLOT International* 7/3, 97-100.