

オントロジー構築・アノテーション支援ツール Open Ontology Forge

川添愛 ナイジェル・コリアー

国立情報学研究所
〒101-8430 東京都千代田区一ツ橋 2-1-2
{zoeai, collier}@nii.ac.jp

要旨

本論文では、テキストや画像をオントロジーに基づいてアノテーションするツール、Open Ontology Forge (OOF) を紹介する。特に、OOFに実装されている知識モデル、OOF のユーザインターフェイスなどについて詳しく述べる。

1. はじめに

本論文では、テキストや画像をオントロジーに基づいてアノテーションするツール、Open Ontology Forge (OOF)を紹介する。テキストや画像、ビデオデータ等に対する意味内容のアノテーションは、機械にとって理解可能な意味記述を与えるという点で重要である。また、意味内容アノテーションをオントロジーと結びつけることにより、個々のアノテーションを、人々や機械が共有できる意味・概念体系の中で位置づけることが可能になる。

この「オントロジーとアノテーションとの関連付け」の重要性に対する認識は高まりつつある。例えば、生物学及び医学の情報を膨大な量の文献から自然言語処理技術を用いて情報を抽出するという研究においては、文献中で重要概念を表すテキスト（遺伝子名、タンパク質名等）を、Gene Ontology [6]等のオントロジーや FlyBase 等のデータベース内の概念と結びつけるという試みが始められている（BioCreAtIvE project [1]）。特に生物学・医学分野においては重要概念を表す用語の多義性・同義性が問題になっており（Tuason et al [13]）、テキストとオントロジー内の概念をリンクすることで、テキストが「何を意味しているか」を明示的に表現することができ、多義性・同義性の解決につながると考えられている（Morgan [11]）。また、画像に対するア

ノテーションの研究においては、画像の意味内容を記述するために、オントロジーやシソーラス等の体系が既に広く用いられている ([2] [7] [8] [10])。例えば、Hollink et. al [7]では、絵画作品の意味内容記述のために、Art & Architecture Thesaurus, WordNet 等の概念体系を利用している。

オントロジーとアノテーションの関連付けによって得られる利点として、同じ概念に対する異なる情報ソースを統合できるという点が挙げられる。すなわち、同じ概念に関する記述のうち、異なる文書に書かれたものや、異なる表現形式（テキストと画像の違いなど）で表されたものをまとめることが可能になる。本論文では、Open Ontology Forge の持つ、アノテーションをオントロジーとリンクすることにより同一指示関係 (coreference relation) を表現し、様々な情報ソースを統合するという側面について特に詳しく紹介する。

2. OOF の知識モデル

OOF に実装されている知識モデルは図 1 のように表される。ここでは、テキストや画像に対するアノテーションが、同一指示プール (coreference pool) というオブジェクトにリンクされている。同一指示プールとは、同じ概念を指し示す表現の集合であり、それぞれオントロ

ジー内のクラスに関係付けられている。同一指示プールは、テキストや画像によって指示される個々の概念 (individual) と見なすことができる。すなわち、アノテーションを介して同じ同一指示プールに関係付けられているテキストや画像どうしは、同一指示関係にあるということになる。アノテーションを同一指示プールに関係付けているのは *pool_id* というプロパティであり、これは各アノテーションが属する同一指示プールの ID を値に取るものである。

OOF の知識モデルは、他の多くのモデルとの適合性を得るため、Protégé-2000 [12] 等のオントロジー編集ツールと類似性を保っている。OOF のアノテーションは、1) 指示対象となる個体の情報、2) ソースとなる文書とのリンク 3) アノテーション作成者及び作成日時などの情報の、三種類の情報を担うプロパティを持っている。後者二つは、アノテーションの際に OOF が自動的に獲得する。これらはテキスト・画像のアノテーションに対して文脈を与えるものであり、作成されたアノテーションの信頼性を判断するのに使用される。

OOF の知識モデルにおいて定義されているプロパティを以下に列挙する (各プロパティに関するより詳しい説明については Collier et al [3] を

参照)。

- *ontology_id* : アノテーションをオントロジー内のクラスと関係付ける。
- *XPointer* : XPointer 値 (DeRose et al. [4]) を取り、アノテーションとソース文書に関係付ける。
- *created* : アノテーションが作成された日時を記録する。
- *modified* : アノテーションが最後に修正された日時を記録する。
- *author* : アノテーション作成者の名前を記録する。
- *comment* : アノテーション作成者が何かコメントを残したい場合に使われる。
- *sure* : アノテーション作成者が、自身のアノテーションの適切さについて自信があるか否かを記録する。
- *expression_type* : アノテーション対象である表現の種類を指定する (固有名、別名、代名詞、定名詞句、不定名詞句、画像など)。
- *text* : アノテーション対象のテキストを記録する。
- *svg* : アノテーションされた画像を SVG [5] で記録する。

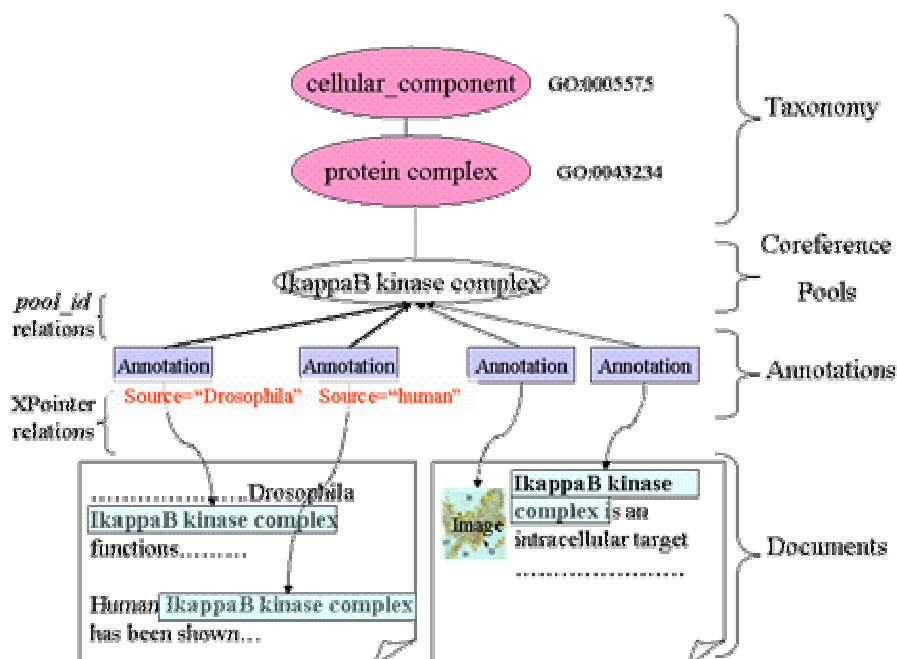


図 1. OOF の知識モデル概要

3. OOF インターフェイス

3.1 文書のインポートからアノテーションまで

OOF を用いたオントロジー構築及びアノテーションは、次に挙げる幾つかのプロセスを経て行う。

まず、Archive タブを選択し、アノテーションの対象となる文書をインポートする。OOF に読み込むことができる文書は、html 及び text である。Web 上からダウンロードすることも可能であるし、ローカルに保存されている文書と呼び出すこともできる。インポートした文書は MySQL データベースに格納される。

Class タブを選択すると、ルートクラスである ROOT_ONTOLOGY を拡張する形でオントロジーを構築することができる。ここで構築できるオントロジーは、クラス間の関係が is-a であるようなものである (part-of 関係は現バージョンでは表現できない)。OOF 上でオントロジーを構築する代わりに、すでに存在する RDF 形式のオントロジーを OOF にインポートすることも可能である。インポートの際、OOF は RDF の妥当性を検証する。オントロジーの修正の際には、部分木ごと移動する操作も行うことができる。

Property タブでは、前節で列挙したアノテーションにデフォルトで付いているプロパティ以外の、新しいプロパティを作成することができる。

Individual タブでは、オントロジーのクラスに属する個体を定義することができる。これは、同一指示関係にあるアノテーションの所属する同一指示プールとして働く。次節でも述べるが、同一指示プールはここで作成する他に、アノテーションの段階でも新規に作成することができる。

Event タブは、述語一項関係のアノテーションに使用する述語を定義するためのものである。

3.2 テキスト・画像のアノテーション

テキスト及び画像のアノテーションは、Annotation タブ (図 2 を参照) で行う。ここでは画面の左側にオントロジーとそれに属する個体が表示され、中央にアノテーションの対象とな

る文書が、右側に各アノテーションのプロパティが表示される。テキストのアノテーションは、アノテーションしたい部分を選択し、オントロジー内のクラスか個体 (同一指示プール) にドラッグすることによって行う。テキストをクラスにドラッグすればそのクラスの下に新規の同一指示プールが作成される。テキストをオントロジーにドラッグする際には、OOF が同じ文書内にある全く同じ文字列をすべて探しだし、それらを同時にアノテーションするオプションを提示する。この機能は、大きな文書で数多く出現するテキストをアノテーションする場合などに有効である。アノテーションされたテキストには第 2 節で述べたデフォルトのプロパティが自動的に付与される。

画像のアノテーションは、画像を選択して図 3 (左) のような専用のウィンドウを表示し、ここでアノテーションしたい部分を SVG の基本図形を用いて選択後、テキストの場合と同様にオントロジーにドラッグするという方法で行われる。図 3 (右) に示されるように、同じ同一指示プールに画像とテキストを関係づけることによって、これらの間の同一指示関係をアノテーションすることができる。

OOF では、アーカイブ化した複数の文書をアノテーション画面に呼び出し、同じオントロジー及び同一指示プールを用いてアノテーションすることが可能である。すなわち、異なる文書間にまたがる同一指示関係もアノテーションすることができる。

3.3. アノテーションサポート機能

OOF にはアノテーションのプロセスを簡単にするための機能がいくつかある。たとえば、アノテーションとクラスの関係を見やすくするために、クラス毎に色を定義してアノテーションを色分けすることができる。また、特定のクラスに対するアノテーションにはホットキーを定義することもできる。

アノテーションをクラス毎に行ったり、表示したりする場合には、アノテーションレイヤーを使用すると便利である。アノテーションレイヤーを用いると、選択されたクラスに属するアノテーションのみが表示され、またそのクラス

に対してのみテキスト・画像を新規にドラッグしてアノテーションすることができる。

その他の機能としては、html 文書でテキストの中に埋め込まれた画像を扱えるというものがある。オンラインで閲覧できる生物学等の論文には、しばしばギリシア文字や数学記号などが画像データとして埋め込まれている。OOF は

アノテーションの際に、選択されたテキスト内にそのような画像が埋め込まれていればそれを探知し、ユーザに提示して適切なテキストに置き換えることができる。

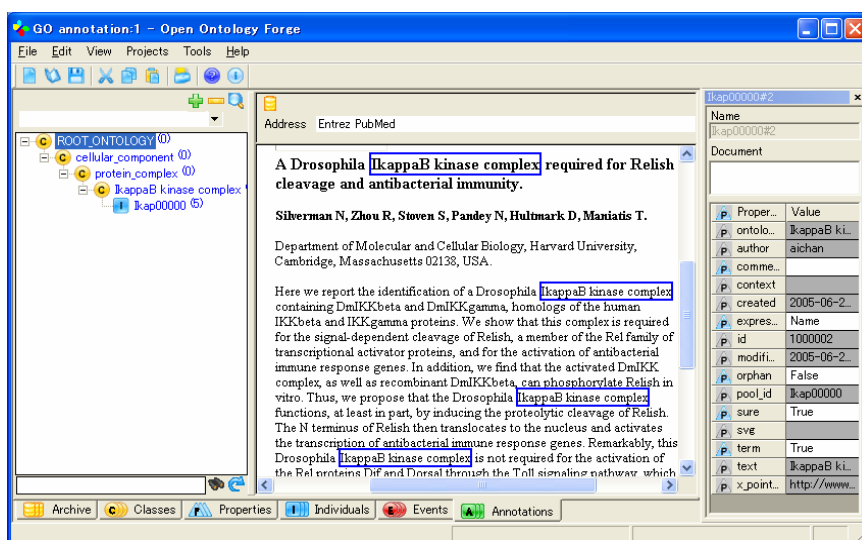


図 2. OOF アノテーション画面のスクリーンショット

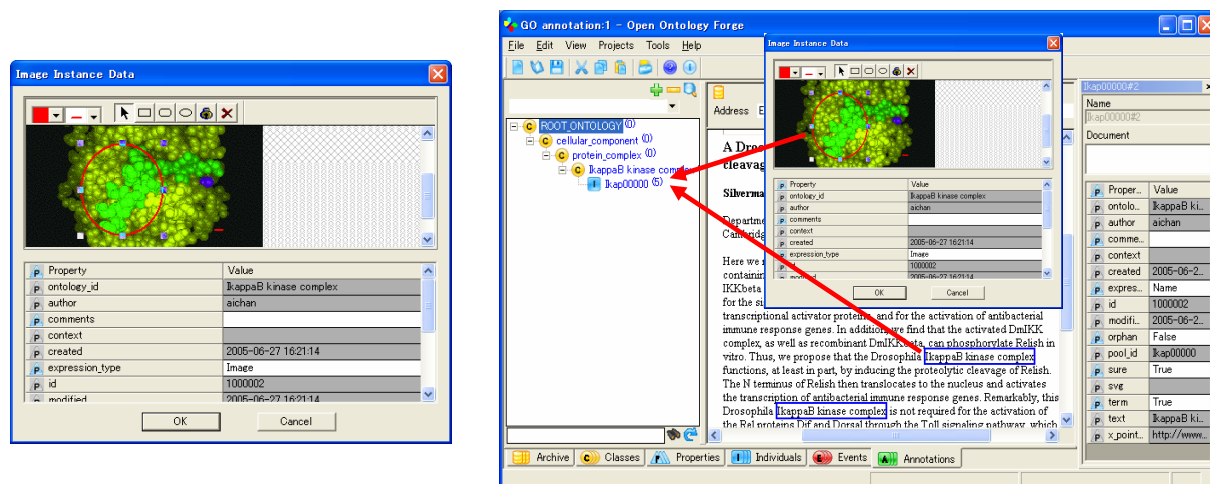


図 3. (左) 画像編集用ウィンドウ (右) テキスト-画像間の同一指示関係アノテーションの一例.

3.4. オントロジー及びアノテーションの出力

OOF 上で作成したオントロジー及びアノテーションは、RDF [9]形式かテキスト内に埋め込まれた XML の形式で出力することが可能である。

RDF 形式の出力にはオントロジーについての記述と、アノテーションされたすべての表現に関する情報の記述が含まれている。XML 形式の場合は、テキストに対するアノテーションのみが含まれる。RDF 形式で出力した画像アノテーション

ョンの例、及び XML 形式で出力の例を以下に示す。

RDF 出力の例：

```
<annotation>
  <item>
    <properties>
      <author>zoeai</author>
      <created>2005-07-10
14:06:37</created>
      <expression_type>Image</expression_t
ype>
      <id>1000000</id>
      <modified>2005-07-10
14:06:37</modified>
      <ontology_id>protein-complex</ontolo
gy_id>
      <pool_id>C 000000</pool_id>
      <sure>True</sure>
      <term>True</term>
      <text><svg
xmlns="http://www.w3.org/2000/svg"
version="1.1"><rect x="1" y="1"
width="1198" height="398" /><image
x="200" y="200" width="100px"
height="100px"
xlink:href="myimage.png"/></svg></te
xt>
<XPointer>http://www.pnas.org/cgi/co
ntent/full/100/10/5823(//img\[1\])</X
Pointer>
    </properties>
  </item>
</annotation>
```

XML 出力の例：

```
In the adult, <NAME author="aichan"
created="2005-07-10 11:23:16"
expression_type="Name" id="1000001"
modified="2005-07-10
11:23:16" ontology_id="protein"
pool_id="C 000000" sure="True"
term="True" text="FHL2">FHL2</NAME>
is expressed in the myocardium of the
heart and in the epithelial cells of
```

```
the prostate, where <NAME
author="aichan" created="2005-07-10
11:23:53"
expression_type="Others" id="100021
3" modified="2005-07-10 11:23:53"
ontology_id="protein" pool_id="C
000000" sure="True" term="False"
text="it">it</NAME> colocalizes with
the AR in the nucleus.
```

4. 結語

本論文では、テキストや画像をオントロジーに基づいてアノテーションするツール、Open Ontology Forge (OOF)の知識モデルとインターフェイスについて述べた。本論文で紹介したバージョンは現在テスト中であり、2005 年秋に配布する予定である。以前のバージョンの OOF は、<http://research.nii.ac.jp/~collier/resources/OOF/index.htm>よりダウンロード可能である。

参考文献

1. BioCreAtIvE Project.
<http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html>
2. Clark, K. (2004) Photostuff: Semantic Image Annotations.
<http://monkeyfist.com/kendall/pstuff.pdf>
3. Collier, N., Takeuchi, K., Kawazoe, A., Mullen, T., and Wattarujeekrit, T. (2003) A Framework for Integrating Deep and Shallow Semantic Structures in Text Mining. In *Proceedings of the Seventh International Conference on Knowledge-based Intelligent Information and Engineering Systems (KES'2003)*, pp.824-834. Oxford, UK.
4. DeRose, S., Maler, E., and Daniel, R. (2001) XML Pointer Language (XPointer) Version 1.0. W3C candidate recommendation, 11th September
5. Ferraiolo, J., Fujisawa, J., Jackson, D. (2003) Scalable Vector Graphics (SVG) 1.1 Specification, *W3C Recommendation* 14 January 2003.
<http://www.w3.org/TR/SVG11/>

6. Gene Ontology Consortium.
<http://www.geneontology.org/>
7. Hollink, L., Schreiber, A. Th., Wielemaker, J.,
Wielinga, B. (2003) Semantic Annotation of Image
Collections. In *Proceedings of the KCAP'03
Workshop on Knowledge Capture and Semantic
Annotation*, Florida.
8. Lafon, Y., and Bos, B. (2002) Describing and
retrieving photos using RDF and HTTP. W3C
Note 19 April 2002
<http://www.w3.org/TR/photo-rdf/>
9. Lassila, O., and Swick, R. (1999) Resource
Description Framework (RDF) Model and Syntax
Specification. Recommendation, W3C, February
1999.
10. Miller, L., and Poulter, M. (2003) Easy Image
Annotation for the Semantic Web. *ILRT Technical
Report 1065*.
11. Morgan, Alex. (2005) Linking Text Mentions to
Biological Identifiers. Presented at e-Biology
Initiative: Towards New Frontiers of Biology held
on March 11th, 2005, the University of Tokyo.
12. Noy, N. F., Sintek, M., Decker, S., Crubezy, M.,
Ferguson, R. W., and Musen, M. A. (2001)
Creating semantic web contents with
Protégé-2000. In *IEEE Intelligent Systems*,
16(2):60–71.
13. Tuason, Olivia., Lifeng Chen, Hongfang Liu,
Judith A. Blake, Carol Friedman (2004) Biological
Nomenclatures: A Source of Lexical Knowledge
and Ambiguity. *Pacific Symposium on
Biocomputing 2004*: 238-249