# GODDAG 再考

師 茂樹*

2005 年 3 月 3 日

## 1 はじめに

SGML や XML などによる電子テキストのマークアップにおいては、ツリー構造の限界がしばしば指摘されてきた。東洋学におけるマークアップの利用においても、文書の論理構造とレイアウト構造の共存問題や、掛詞のようにそもそもツリー構造では表現できない構造などがしばしば指摘されてきている。例えば、『大阪市史』に収録される史料を電子化し SGML によってマークアップを試みた柴山守氏は、「SGML 化に際しての種々の問題点」として次の 4 点を挙げている [15]。

1. SGML 化の目的、すなわち印刷出力、検索、文書交換などの内容によって SGML 化の設計が異なる。これは、言い換えれば、ある目的のための SGML 文書を別の目的には利用できない。
2. SGML 化の目的によっては、原テキストの文書構造と異なり、SGML 化が不可能な場合がある。すなわち、歴史テキストでは明確な構造化が不可能な場合が存在する。
3. SGML 化の目的によっては、原テキストに手を加え、変形しなければならない。すなわち、タグを取り除いても元のテキストを復元することができない場合がある。
4. 属性値の付与などで参照関係の設定が不可能である場合が存在する。

一方、SGML の CONCUR を始めとして、これまでに種々の試みがなされてきた。その多くは、SGML や XML など既存の標準マークアップ言語の枠内での試みであるが（一例をあげれば、[4, 8, 9, 10, 12] など）、未だ決定的な解決策は見出されていない[*1]。そのため、標準マークアップ言語以外のマークアップ言語を提案する者もいる。

本発表でとりあげる GODDAG[11] は、後者のアプローチによる提案のひとつである。以下、GODDAG を題材にその意義と問題点について考えることで、よりよいテキスト処理技術について考える一助としたい。

* 花園大学文学部専任講師; s-moro@hanazono.ac.jp

*1 先行研究の調査に際しては、XML ユーザーメーリングリストに対する村田真氏の投稿（http://www2.xml.gr.jp/log.html?MLID=xmlusers&TID=9060&F=10&L=10&R=1#9065）を参考にした。なお、この投稿において村田氏は、自身の研究を含めたこれら過去の取り組みを「死屍累々」と評している。

# 2　GODDAG の意義

## 2.1　GODDAG について

C. M. Sperberg-McQueen 氏と Claus Huitfeldt 氏による GODDAG[*2] は、Huitfeldt 氏が提案する MECS[5] の表記法に準じ、次のようなタグのオーヴァーラップを認めるマークアップを可能とする。

```
<s/<a/ John <b/ likes /a> Mary /b>/s>
```

ここでは、要素 a と要素 b が互い違いになっており、「likes」の部分にオーヴァーラップしている。

Sperberg-McQueen 氏らは、このようなタグのオーヴァーラップの構造を図 1 のように表現する。GODDAG の具名 “General Ordered-Descendant Directed Acyclic Graph” からもわかるように、従来のツリー構造ではなく、親から子へという方向を持つ有向非巡回グラフとして表現しうると考えられている。グラフは “GODDAG” という略称にするために無理矢理考え出されたものかもしれないが、将来の拡張として、

- graphs with disordered nodes
- graphs representing multiple orderings of data

があげられているので、名前だけではないのだろう。最近になって、グラフ理論によるテキスト処理の可能性について安岡孝一氏 [17, 18] によって提起されているが、マークアップによるテキスト処理の分野において限定的ながらグラフ理論を導入していたことは、もっと注目されてよいように思う。

しかしながら、[11] はごく基本的な仕様について述べるのみで、タグ名に使ってよい文字の規定や属性の書き方など、実用に必要な情報についてはまったく知ることができない（そもそも思考実験的なものなのかもしれないが)。特にオーヴァーラップを表現する上で、

```
<a/ John <a/ likes /a> Mary /a>
```

が入れ子なのかオーヴァーラップなのかを属性などで区別する仕組みが存在しない点は、これまでの議論で何度も指摘されてきたことであることから考えても不十分である。今回、研究にあたって、Perl の正規表現によるごく簡単なパーサを作成したが、タグ名用の文字の問題などについては XML に準じた[*3]。

[*2] デンマーク語で “goddag” は、英語の “good day” に相当する挨拶（発音は「ゴデー」に近い)。

[*3] 作成にあたっては、[3, 6, 13] を参考にした。

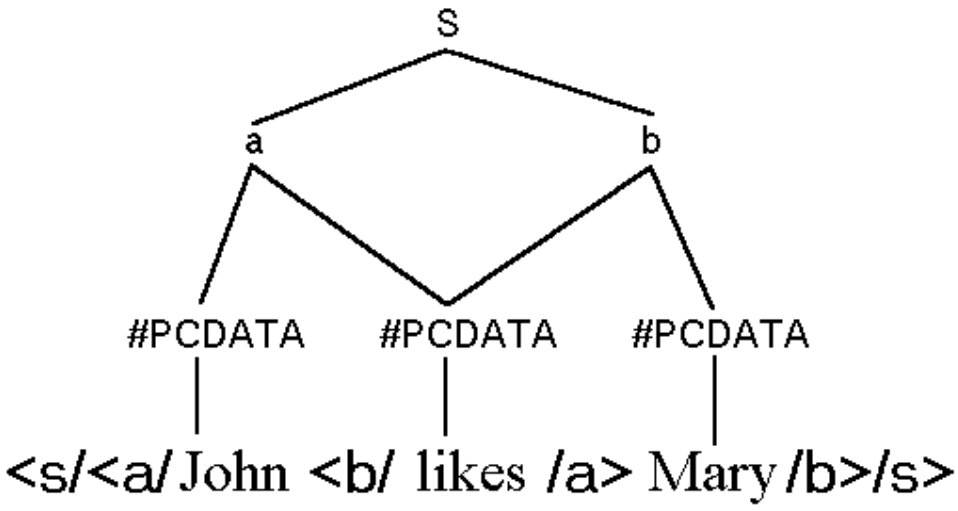


図 1　タグのオーヴァーラップ（[11] より引用）

## 2.2　オーヴァーラップの意義

ツリー構造の限界性とオーヴァーラップの意義については、所々に述べられているので再説はしないが、文章構造に限らず、様々な分野で論じられていることを付言しておきたい。柄谷行人氏 [14] は、建築家 Christopher Alexander 氏の論考 “A city is not a tree” [2] に言及して次のように述べる。

> アレグザンダーは、自然都市はセミ・ラティスの組織をもっており、人工都市はツリーの組織を持っているという。われわれが都市を人工的に組織するとき、それをツリーとして組織してしまう。ツリーとセミ・ラティスは、多くの小さなシステムがどのようにして大きな複雑なシステムを形成するかについて考える方法であり、もっと一般的にいえば、それらは集合の順序的構造に対する名称である。(pp. 52-53)。

ここで言う「セミ・ラティス」とは、GODDAG と同じオーヴァーラップ構造のことである。柄谷氏は、これに、ヴァレリーのテクスト論を重ねる。

> 人間によって作られたものの特徴は、その形態の構造が素材の構造より単純だという点にあると、ヴァレリーはいう。たとえば、ある文学作品の「構造」が把握されるとき、それはつねにテクストより単純である。いかなる「構造」も、何らかの意図・目的なしには考えられない。あるテクストの「隠された構造」をみるとき、すでに隠された意味、あるいは、制作者が想定されている。ところがテクストは人間によって作られたものでありながら、それよ

りも複雑で過剰な「構造」をもつ。なぜなら、それは「自然言語」という素材によっているからである。制作者がどう統御しようと、言語そのものが別の意味をもってしまうことを避けられない。その意味で、テクストは「自然が作ったもの」なのである。(pp. 48-49)

アレグザンダーによる、人工都市と自然都市の差異に関する数学的考察は、それが形式的であるがゆえに豊かである。たとえば、それは組織機構にかんしてもあてはまる。軍隊や官僚機構はツリーであり、上位組織を介さない横断的交通は許されない。(中略) この意味で興味深いのは、ヴァレリーが「人工的なもの」の構造についてのべるとき、軍隊を例にとっていたことだ。いいかえれば、ヴァレリーは、「人工的なもの」の特性がツリーであることを省察していたのである。(pp. 57-59)

この議論をふまえれば、XML に代表されるようなテクストのツリー構造が、ヴァレリーやアレグザンダーの言う「人工的なもの」である、と言えるだろう。言うまでもなく、柄谷氏が「自然な」テキスト=セミ・ラティス構造と考えているわけではないが、ツリー構造の限界性が、都市や組織など、他の人工物に敷衍されて論じられているのは興味深い。

## 3 GODDAG の問題一端

GODDAG の問題点として、同じ範囲のオーヴァーラップが表現できない点を指摘しうる。

例えば、論理構造とレイアウト構造を同居させたい場合、ページ (page) と文 (s) がずれている箇所では、下のようにオーヴァーラップさせることができる。

```
<page/...<s/我が輩は猫/page><page/である。/s>.../page>
```

しかしながら、たまたまページと文との範囲が一致した場合、例えば、

```
<page/<s/我が輩は猫である。/s>/page>
```

と書くと、s は page の子となってしまう。苦し紛れに、

```
<page/<s/我が輩は猫である。/page>/s>
```

と書くこともできるが、この場合 GODDAG で規定される “spurious overlap” となってしまう。その際、GODDAG の規定では、「オーヴァーラップのない形に書き換えることができる (the document can be rewritten without overlap) 」とあるが、親子関係に書き替わってしまうのは構造の変化を意味するため、問題が多いのではないかと思われる[*4]。

この問題は、GODDAG を含む従来の方法において、タグがひとつのタグ名 (=セマンティクス) しか持ち得ないことが背景にある。従来の方法において、ひとつのタグに複数のセマンティクスを持たせる方法として属性をつかうことがあげられるが、属性の場合、スコープがタグに限られ、かつ構造化することができないなどの問題が指摘されており [16]、解決策にはならない。

[*4] この問題は、マイルストーンなどの他の方法でも起こりうるのではないかと思われる。

## 4 終わりに

以上、GODDAG の意義と問題点について簡単に見てきた。オーヴァーラップを含むツリーを超えた構造をテキストとして扱う際、グラフ構造を用いることがひとつの解決策になることが期待され、GODDAG はその意味で画期的であったが、既存のタグ表記法に依存している部分が大きく、不十分な記述しかできないことがわかった。

既存のテキスト処理の文脈の延長線上でグラフ表現について考えてみると、XML などで記述する方法[*5]や Emacsen の文字プロパティなどが考えられるが、どのような方法が適切なのか、今後の研究課題としたい。

## 参考文献


[1] XML Topic Maps (XTM) 1.0: TopicMaps.Org Specification. Online: `http://www.topicmaps.org/xtm/1.0/`, Aug 2001.

[2] Christopher Alexander. A city is not a tree. *Architectural Forum*, Vol. 122, No. 1-2, 1965.

[3] Robert D. Cameron. REX: XML Shallow Parsing with Regular Expressions. Online: `http://www.cs.sfu.ca/~cameron/REX.html`, Nov 1998.

[4] Steven DeRose. Markup overlap: A review and a horse. Online: `http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01/EML%2004DeRose01.html`, 2004.

[5] Claus Huitfeldt. MECS - A Multi-Element Code System. Online: `http://xml.coverpages.org/MECS-summary2001.html`. forthcoming in Working Papers from the Wittgenstein Archives at the University of Bergen, No 3 (ISBN 82-91071-02-0, ISSN 0803-3137).

[6] Paul Kulchenko. XML::Parser::Lite - Lightweight regexp-based XML parser. Online: `http://search.cpan.org/dist/SOAP-Lite/lib/XML/Parser/Lite.pm`, 2000.

[7] Olivier Liechti, Mark J. Sifer, and Tadao Ichikawa. Structured graph format: XML metadata for describing Web site structure. Online: `http://www.oasis-open.org/cover/sgfWWW7.html`, (No date).

[8] Makoto Murata. File format for documents containing both logical structures and layout structures. *Electronic Publishing*, Vol. 8, No. 4, Dec 1995.

[9] Dot Porter. Overlapping Markup SIG Minutes 23 October, 2004. Online: `http://www.tei-c.org/Activities/SIG/Overlap/olm02.html`, Nov 2004.


[*5] XML でグラフ構造を扱う方法としては、[7] や [1] など。ただしいずれもテキスト処理とは関係がない。

[10] Dot Porter. Report of TEI Overlapping Markup SIG meeting in Baltimore. Online: http://xml.coverpages.org/TEI-Overlap200410.html, Nov 2004.

[11] C. M. Sperberg-McQueen and Claus Huitfeldt. GODDAG: A data structure for overlapping hierarchies. In Anne Bruggemann-Klein and Ethan Munson, editors, *Proceedings of PODDP'00 and DDEP'00*. New York: Springer, 2001. Presented at ACH-ALLC '99. Online: http://helmer.aksis.uib.no/claus/goddag.html.

[12] Jeni Tennison and Wendell Piez. The Layered Markup and Annotation Language (LMNL). Online: http://xml.coverpages.org/LMNL-Abstract.html. Extended abstract of the presentation at Extreme Markup 2002.

[13] wepmaster@donzoko.net. XML::Parser::Lite の改良. Online: http://www.donzoko.net/cgi/xmlparser/, (No date).

[14] 柄谷行人. 隠喩としての建築, 定本柄谷行人集, 第 2 巻. 岩波書店, Jan 2004.

[15] 柴山守. 大坂町触の SGML/XML 化と全文検索. 京都大学大型計算機センター第 62 回研究セミナー報告, 1999.

[16] 豊島正之. XML の骨抜き利用法—アジア・アフリカ言語文化研究所データベースの例. Online: http://www.classics.jp/Contents/Assets/notice/xmt-main.pdf, Oct 2001.

[17] 安岡孝一. テキスト検索は文字列検索でも木検索でもない. Online: http://kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/2004%-11-02/yasuoka2004-07-02.ppt, Jul 2004.

[18] 安岡孝一. 紙テープの呪縛. Online: http://kura.hanazono.ac.jp/paper/20040609yasuoka.pdf, Jun 2004.