Survey of RDF data on the web

このドキュメントは

Survey of RDF data on the Web

http://www.i-u.de/schools/eberhart/rdf/rdf-survey.pdf

の和訳です。

この文書には和訳上の誤りがありえます。

内容の保証はいたしかねますので、必ず正式版文書を参照して下さい。

ウェブにおける RDF データの調査

Andreas Eberhart

International University in Germany

eberhart@i-u.de

August 15, 2002

要約

Resource Description Framework(RDF)はメタデータ相互流通のためのフォーマットである。これはセマンティック Web に向けての有望な基本技術である。この論文は、Web 上に RDF データがどのくらい、どのようなものがあるかを、2001 年 12 月、2002 年 8 月に調査したものである。4 つの探索方略、クローリング、インターネットディレクトリ中の URL スキャン、特定トピックの検索、さらに以前集めた URL の検索、の結果を比較する。その結果、現在 RDF はかなり頑張って探さないと見つからないことがわかった。

1 序論

Web の初期フェーズでは、静的な情報を読んだり発行したりする便利なメディアであったが、最近はウェブアプリケーションの人気が途方もないほど成長している。 今日、オンラインで利用できないサービスはないくらいである。しかし、これらのサービスのほとんどが人間向けである。電子データ交換(EDI)コミュニティではアプリケーション統合のためにメッセージ形式の標準化にかなり成功しているが、どんな応用にも役立つような軽い標準を開発するのは不可能である。 そのため、EDI ソリューションは典型的な特定の業界にしか使えない。

セマンティック Web は、ソフトウェアエージェントが自動でコミュニケーションを行うことでウェブの可能性を最大にしようという狙いである。 中心となる考えはオントロジで異なるドメインの概念を統合しようというものである。オントロジによりエージェントがファクトやルールを伝えたりできる語彙が提供される。しかし、普遍的な API やメッセージ形式を開発するのが不可能なように、誰もが必要とする普遍的なオントロジを開発するのも不可能である。Web の精神に従えば、一連のマークアップ言語 RDF、RDFS、DAML、および

RuleML により、誰もがメタデータ、ドメインの重要概念、そしてドメインエキスパートによる規則や制約条件が書けるようになる。おそらく、そのうちにオントロジの自然淘汰がおこり、それがセマンティック Web を推進していくことになるだろう。

RDF の簡単な説明

Resource Description Framework(RDF)は、メタデータのための枠組みとしてセマンティック Web で最も基本的なマークアップ言語である。 中心となる考えはものごとを URIs として扱うことである。人間は、その人のホームページで指定することができる。 Ora Lassila について述べるのにリソース http://www.lassila.org を使うなどである。あるオフィスの机は、その会社の URL を使って http://xyz.com/inventory#K4622-ERF などである。RDF の用語では、これらはリソースと呼ばれる。リソースについてステートメントを作ることができる。ジョーがピーターズの兄弟であるということは、以下のように、Subject, Predicate, Object の 3 つ組みで表される。

Subject: http://www.mit.edu/~joe/

Predicate: http://www.cogsci.princeton.edu/~wn/concept#107127521

Object: http://www.mit.edu/~peter/

ここで「兄弟」という述語は、プリンストンの Wordnet の語彙データベースプロジェクトを示す URL であるのに注意されたい。「兄弟」という概念(同じ親を持つ男性)には、ID107127521 がついている。ジョー、ピーター、および他のリソースについて、さらにステートメントを書けるので、結局 RDF は有向ラベルつきグラフになる。リソースはグラフのノードであり、ステートメントはエッジに相当する。

上の 3 つ組みはややわかりにくい。しかし、それは「兄弟である」という関係を、広く知られた Wordnet の語彙を使うことで、多くのエージェントが正しくステートメントを解釈できるようにするためである。「ジョーがボストンに住んでいる」という次の例を見てみよう。

Subject: http://www.mit.edu/~joe/

Predicate: http://www.schema.org/rdf/livesin

Object: Boston

一つの違いは述語が別のネームスペースから来ていること。二つ目の違いは、オブジェクトは、別のリソースではなくリテラル(文字列)であることである。別のステートメントでオブジェクトとして文字列「Boston」を使っても、それは、都市であるか、ボストンというコードネームによるプロジェクトであるかは、アプリケーション次第である。さらに調べたい場合は、 RDF/RDF スキーマ(http://www.w3.org/RDF)DAML[1]、RuleML[2]、およびセマンティック Web 一般[3]のサイトを見ていただきたい。

RDF は、XML 構文にもシリアライズできる。 多くの XML ドキュメントと同じく、RDF もあるアプリケーションから別のアプリケーションに流れるバイトストリームである。ネッ

[1] http://www.daml.org
[2] http://www.dfki.de/ruleml
[3] http://www.semanticweb.org

トワーク。 また、RDF は静的ファイルにも格納することができ、HTML のファイルのヘッダに格納してもよい。

.

RDF 調査のモティベーション

RDF はセマンティック Web の基礎として事実(ファクト)を記述することができる。そのため、RDF データがどのくらい探せるかを調査することは興味深い。RDF はセマンティック Web 研究コミュニティでは、良く使われている。したがって、この調査を行うことで世の中がセマンティック Web 技術をどれだけ使い始めたかがわかる。また、どのような述語がアプリケーションでよく使われているかを調べるのも興味深い。

2 章は、どのように探索して、どのようなツールを使ったかを説明する。調査結果は 3 章に、続く章に評価とサマリを示す。

2 調査データのコレクション

いくつかの初期実験は、Karlsruhe 大学[4]で開発された RDF クローラを使って行われた。初期 URL を与えると、RDF クローラは一定の深さまで再帰的にページを収集した。見つかった RDF データはローカルファイルに格納される。この実験をやってすぐに、RDF データを Web で見つけるのは簡単でないことがわかった。単純なクローラーだと起点 URL を慎重に選ばないと、RDF データにたどり着く前に膨大なファイルを集めなければならない。そこで方針を変え、以下に解説するような方法で行った。

最初の実験は 2001 年 12 月に行った。同じソフトウェアと検索プロセスで 2002 年 8 月に再実験をおこなった。これはセマンティック Web のイニシアチブが最近ポピュラーになってきたため、その影響を見るためである。今後同様の実験を繰り返す予定である。

2.1 クローリング

1998 年の Lawrence, Giles による研究では、大手のウェブサーチエンジンでもインターネットページの 17%を集めているにすぎない。インターネットの急成長では、この数値はこれからかなり減ると言われている。今回の研究に利用できるバンド幅とコンピュータリソースでは、URL 群のごく一部を集められるにすぎない。そこで我々は、出発点として RDF コミュニティの中の有名サイトを起点として選んだ。 テーブル 1 は両方の探索実験で使った起点 URL である。2 ホップにより、計 12,507 ページを最初の実験で処理した。2 つのメジャーな RDF ファイルのコレクション、つまり、オープンディレクトリの構造/コンテンツのダンプ[5]と、

[4] http://ontobroker.semanticweb.org/rdfcrawl/
[5] http://dmoz.org/rdf.html

URL
http://www.w3.org/RDF/
http://wilbur-rdf.sourceforge.net/
http://www.daml.org/
http://www.lassila.org/
http://www-db.stanford.edu/_melnik/
http://www-db.stanford.edu/_melnik/rdf/api.html
http://www-db.stanford.edu/_stefan/
http://www-db.stanford.edu/
http://www.semanticweb.org/
http://protege.semanticweb.org/

Table 1: Popular sites within the RDF community were chosen as starting points for crawling

Wordnetの語彙データベースプロジェクト[6]のRDFバージョンは､大規模すぎるため外した。RPMソフトウェアパッケージ tool[7]に関連するサイトはソフトのディストリビューションを記述した多くのRDFファイルを含んでいるが、部分的に集めただけである。最初の実験での起点ページの選択が非常に恣意的に制限したため、２度目の実験[8]ではGoogleディレクトリのRDFカテゴリ以下のURLも含めた。www.semanticweb.orgページと同様に、これらはいろいろなRDFの関連サイトのリンクを含んでいる。2度目の実験では31,764ページが集まった。

2.2 オープンディレクトリ

Webの幅広さがいくらかはわかるため、我々はオープンディレクトリ project[9]からURLを捜した。このプロジェクトはウェブサイトをヤフーと似たようなカテゴリへまとめたものであり、URLとその説明文はRDFフォーマットで利用可能である。最初の実験では、527,408のURLが含まれていた。オープンディレクトリプロジェクト成長は早いため、8カ月で、この数は2,912,434まで増加した。これにはアダルトページ以外のすべてのカテゴリの合計である。こうして得られたURLは、通常エントリページかホームページである。URLが膨大になるため、これらのサイトはこれ以上クローリングしなかった。この方法だと明らかに単独で現れるRDFデータは集められない。しかし、以下で述べるようにHTML内のRDFは見つけることができる。

example:

```
<head>
    ...
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
             xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about=""
```

[6] http://www.semanticweb.org/library/
[7] http://rpm_nd.net/linux/RDF/
[8] Google RDFディレクトリは http://directory.google.com の>Libraries > Library and Information Science > Technical Services > Cataloguing > Metadata> Resource Description Framework.なお、Googleディレクトリはオープンディレクトリを利用している。
[9] http://dmoz.org

```
            dc:title="Ora Lassila"
            dc:description="Ora Lassila's professional home page"
        </rdf:RDF>
    </head>
```

2.3 対象を絞った探索

最初の実験では RDF データが余り見つからなかったので、対象を絞った探索をすることにした。Google ではある文字列を含む URL を検索することができる。明らかに、”RDF ”を含む URL はなんらかの RDF データが入っていそうである。Google の検索結果から URL を抜き出す簡単なパーサを作った。二番目の実験で、我々は”Java ”または”.NET client ”[10] を Google から自動で検索する Web サービスに基づくプログラムを作った。最初の実験では、このようにして 1,256URL が見つかった。

テーブル 2 はこの 3 つのカテゴリの URL 数をまとめたものである。カテゴリ間にはわずかにオーバラップがある。RDF コミュニティとオープンディレクトリカテゴリの両方に現れる 3 つの URL に RDF データが見つかり、RDF コミュニティにある 63URL が、Google の検索カテゴリにも現れる。 2 番目の実験において、URL に”RDF ”を含むページはあまり見つからず、数は 1,079 まで減少した。これは RDF の情報が減ったというわけではない。おそらく、Google データベース内部の変化であろう。検索結果には 241 万ページが見つかったとあっても、ブラウザと Web サービスインタフェースでは、前述の数までしかアクセスできなかった。

2.4 3 つ組の事実(ファクト)から見つかった URL

RDF のサブジェクト、述部と多くのオブジェクトは URL そのものなので、こうした URL からさらに RDF データを集めることができると考えた。まず、他カテゴリから得られたファクトを対象とし、URL としては他のカテゴリには出現しないものを選んだ。この制限は、他カテゴリとオーバーラップするものが多いだろうと考えたからである。124,374 のファクトが最初の実験で見つかり、365 の新しい URL を得ることができた。ここで、URL のアンカーとして#を含むものは、同じ URL の場所が違うものなので無視した。これらの URL のうちファクトのものは、新しい RDF が集まるかも知れないので再び収集した。

	Number of URLs scanned	
Category	Dec 2001	Aug 2002
RDF Community	12507	31764
URLs from Open Directory	527408	2912434
RDF appears in the URL	1256	1079
URLs from facts	365	6733

Table 2: URLs per category

365 の新 URL から、1,923 の新しいファクトを見つけたが、そこから先は 23 の新サイトしか見つからなかったので、収集プロセスはそこで停止した。

[10] See http://www.google.com/apis/

2 番目の実験では少し違った。139,288 のファクトが他のカテゴリの URL から見つかった。これらのファクトからの、サブジェクト、述部、およびリソースオブジェクトで 6,037 の未知 URL を指していた。それらの URL から 54,227 の新しいファクトが見つかった。この数はとても高いが、調べてみるとほとんどのファクトは、一つの URL 中の大きなファイルにあったのではなく、いくつかの URL に含まれていることがわかった。一例をあげると、http://xmlns.com では、Wordnet データベースの RDF 表現をホスティングしている。例えば、URL http://xmlns.com/wordnet/1.6/Survivor には似たような他の URL にある Wordnet リソースに関するステートメントが含まれる。それでも、それらの新しいファクトから 697 の新 URL を取り出すことができた。この時点では、ファクトと、それまで見た URL にリンクしているサイトとを区別することができなかったため、プロセスは終了した。

2.5 RDF データベースのアーキテクチャ

データをさらに分析することができるように、ファクトを RDB システムに格納することにした。図 1 はそのテーブルレイアウトである。

ファクトテーブルには、サブジェクト、述部、およびオブジェクトによる 3 つ組、さらにそれが見つかった URL を格納する。 プライマリキーにすることで二重の挿入を防いでいる。URL テーブルは、データのダブりがおきないようにさらなる一貫性制約を持っている。 最後に、URL type テーブルは、URL が前記の 3 つのカテゴリのどれに属すかを示す。URL の msgfield はネットワークエラー、XML パースエラーなど解析中のエラーを記録する。図 2 は全体のソフトウェア設計である。集められた全 URL はまず URL テーブルに挿入される。各カテゴリによってアプローチは異なる。オープンディレクトリダンプからのデータは XSLT で抜き出された。Google の検索から URL パターンで取り出した URL を集めるプログラム GetGoogleURLs は Google に対して検索を行う。パターンのエンコードは、拡張検索による膨大な結果セットにたいして 1 ページずつブラウジングして得た自動にクエリを繰り返して与えるのは、Google が禁じていることもあるので、注意したい。クローラープログラムは、複数の基点からハイパーリンクをマルチスレッドで収集する。

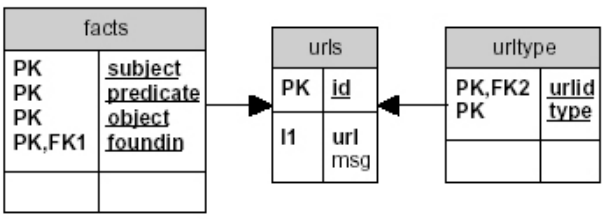

Figure 1: Design of the RDF database

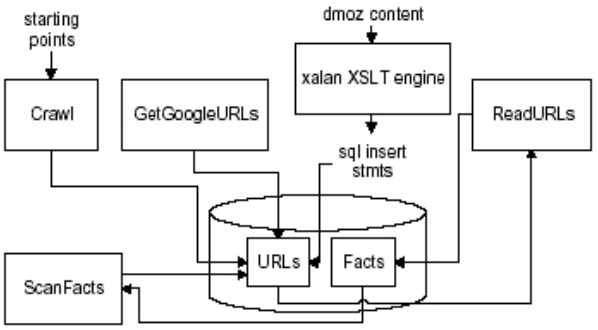

Figure 2: Overall software design

URL テーブルに格納されると、RDFLoad プログラムが URL をスキャンして RDF データを探す。ファクトを DB にアップロードするには Sergey Melnik の RDF API[11]を使用している。Java の例外で何か引っかかったものは msgfield に格納される。例えば、どのくらいのページが構文上正しくない RDF を含んでいるか、どのくらいのページがネットワークの外にあってアクセスできないかなどである。これは最も広く使われている RDF APIでなので、2001 年 1 月 19 日バージョンで、エラーがなく、3 つ組が空でない URL は、正しい RDF を含むと考えられる。 "org.xml.sax.SAXParseException"が投げられた場合、その URL には RDF は含まれないことになる。また、”java.io.EOFException: no more input ”の場合は、３つ組が空ということを表す。また、見つかった RDF データは num.rdf(num は見つかった URL の ID)というファイルに格納してさらに調べている。最後に、ScanFacts は見つかった URL を URL テーブル中のファクトテーブルに格納する。ここで、ScanFacts は ReadURLs がいくつかのファクトを挿入した後にのみ実行することに注意されたい。さらに、ReadURLs は新しく見つかった URL で見つかったファクトをロードするのに使われる。
このデータベース中心のアーキテクチャの主な利点は、検索プロセスを止めてたり再開したり自由にできることである。 データベースは、二度同じデータを挿入できないような一貫性制約を提供している。この実験は、RDF の現在のスナップショットでの利用動向を調べるため、データベースはデータを追加するだけで、削除・アップデートは行われない。 データセットおよびアプリケーションは、http://www.iu.de/schools/eberhart/rdf/からダウンロードすることができる。

3 探索結果

この章では、第 1 回目の実験での 541,536 のウェブサイト、および 2 番目の実験での 2,952,010 のウェブサイトでの検索結果について概説する。

3.1 どのくらいのページが RDF データを含んでいるか?

図 3 と図 4 は、どのくらいのページが RDF データを含んでいるかを、以下のケースに分けて示している。

[11] http://www-db.stanford.edu/~melnik/rdf/api.html

- ファイルが見つからなかったなどの一般エラー(シアン)
- ページはあったが RDF は含まれない(黄)
- 構文上、不正確な RDF データ(赤)
- 正しい RDF(青)

予想されるように、カテゴリとは強い相関がある。最初の実験では、オープンディレクトリから得た 50 万ページ以上から 16 しか RDF は見つからなかった。2 番目の実験でこの数は 290 万ページ中 180 まで増加した。セマンティック Web ポータルの近くのページでは RDF の出現率は、もう少し高いが、期待はずれであった。両方の実験[12]では、ファクトに現れる URL のおよそ 1 パーセントしか RDF を含んでいなかった。最も高く RDF を含む URL は、".rdf"で終わるページであり、最初の実験で 10%, 2 番目の実験では 17%含んでいた。他のカテゴリで見つかったファクトの RDF リンクをたどった URL についても、同じような率 RDF が含まれていた。

最初の実験では 9%、二番目の実験では 13%のページに RDF を見つけることができた。カテゴリを合わせることで、最初の実験では 541,536URL から 1,018URL の RDF を含むページが見つかり、そのうち 613 は正しい RDF, 405 が誤った RDF であった。 2 番目の実験では、2,952,010 ページから 1,479 の正しい RDF、2,940 の無効な RDF が見つかった。ただし、大量に集めても RDF ページの大半がオープンディレクトリのカテゴリに含まれていることに注意したい。

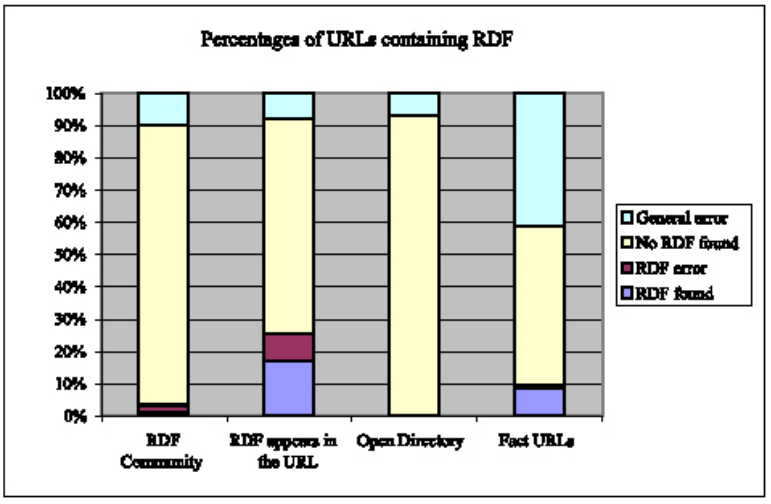

Figure 3: RDF data found per category during the _rst search Dec. 2001

[12] Originally we also counted documents on which the RDF parser used yielded an empty RDF result, i.e. a set of zero RDF statements. The exclusion of these pages explains the higher number stated in [4].

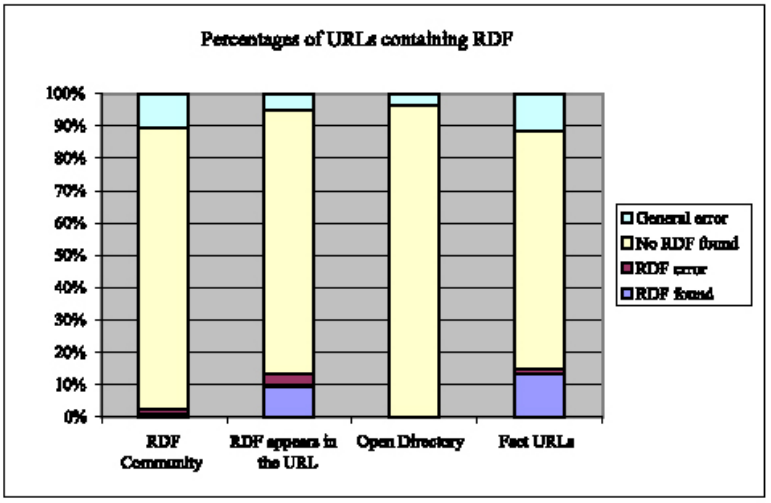

Figure 4: RDF data found per category during the second search Aug. 2002

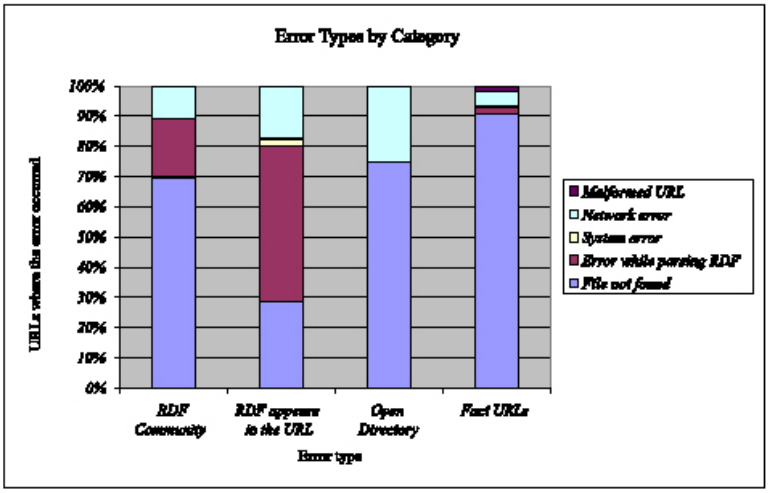

Figure 5: Error types during the_rst search Dec. 2001

3.2 エラー原因

図 5 と図 6 はエラー(図３のシアンと赤の部分)の原因を示したものである。ネットワークエラーと URL が存在しないというエラーが最も多いと思われた。最初の実験では、ReadURLs コンポーネントのシステムエラーで、対象を限定した検索から得られた URL の処理にやや影響があった。

こうしたエラーのうち 5 つは記録されていて、4 つは原因不明の例外処理、１つは大きなバイナリファイルによるメモリエラーだった。2 番目の実験では、我々は特にオープンディレクトリページの大量ページを探すのにスレッド数を増やした。これによって、メモリエラーは 1,147 と多くなった。全 290 万 URL をスキャンしたことを考えると、この数は結果に影響を与えるほど大きくはない。

興味深いのは、2,940 のエラーのうち 405 件が、古い RDF フォーマットによるエラーまたは、RDF ツールの既存バグによるものということで、今後問題になるかもしれない。いずれ

の実験でも、これらのエラーの半分以上は HTML におけるスペースエンティティ   が原因である。 残りのエラー原因としては、アトリビュートに引用符がないという XML のエラー、description が入れ子になっている RDF のエラー等であった。それぞれの実験において 14URL と 24URL で、単に<RDF>タグでくくっただけでネームスペースがないというのがあった。これは、”unresolved namespace prefix”というエラーになる。

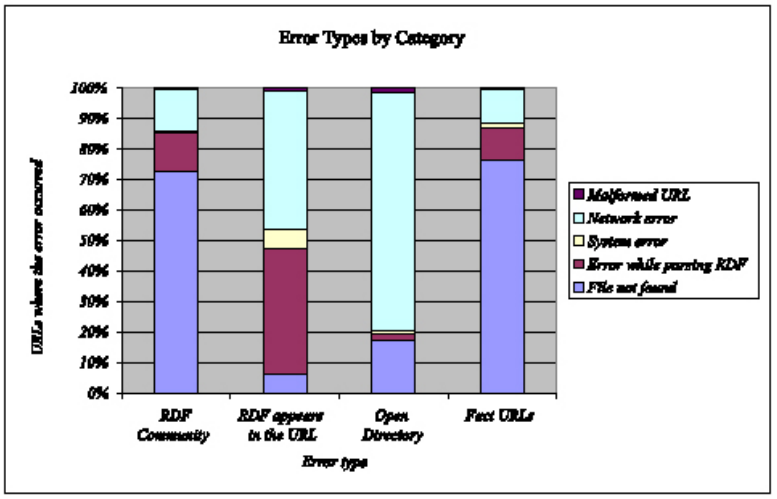

Figure 6: Error types during the second search Aug. 2002

3.3 RDF データセットのサイズ

最初の実験では、125,072 個のファクトが抽出された。そのうち 104,580 は対象限定検索から、19,696 個は RDF コミュニティカテゴリ、1,923 個は URL の形から、98 個はオープンディレクトリウェブサイトからであった。 図 7 と図 8 は、どのくらいのデータが異なった URL で見つかったかを示す。 2 番目の実験の分析では、全カテゴリで 254,783 個のファクトが見つかり、

RDF コミュニティページが 107,308 個、URL 形状から 115,495 個のファクトを含んでいた。29,168 個が対象を絞った探索、2,812 個がオープンディレクトリページからであった。最初の実験では、わずか 3 つの大きなファイルに、10,000 以上のファクトが含まれていた。すなわち、

http://www.megginson.com

http://www.ontoknowledge.org にある CIA 世界情勢、http://w.moreover.com のカテゴリ記述である。二番目の実験では、5 つに多くのファクトが含まれた。つまり、

http://opencyc.sourceforge.net の OpenCyc project

http://www.semanticweb.org/library/wordnet/wordnethyponyms-20010201.rdf にある WordNet の一部

http://orlando.drc.com/ にある２つの軍関係のオントロジ、

http://w.moreover.com

である。

全体的に見て、２つの実験の間に次のような違いがある。 まず第一に、最後のカテゴリで

は、密接に関連した 2 つのデータセット(WordNet のあるバージョンの記述で、xmlns.com から多くのファイルが moreover.com にリンク)が含まれることが多い。これによって数は大きく増加する。 他のカテゴリでは、オープンディレクトリカテゴリにおける中規模のデータセットが含まれることを除けばあまり変わっていない。

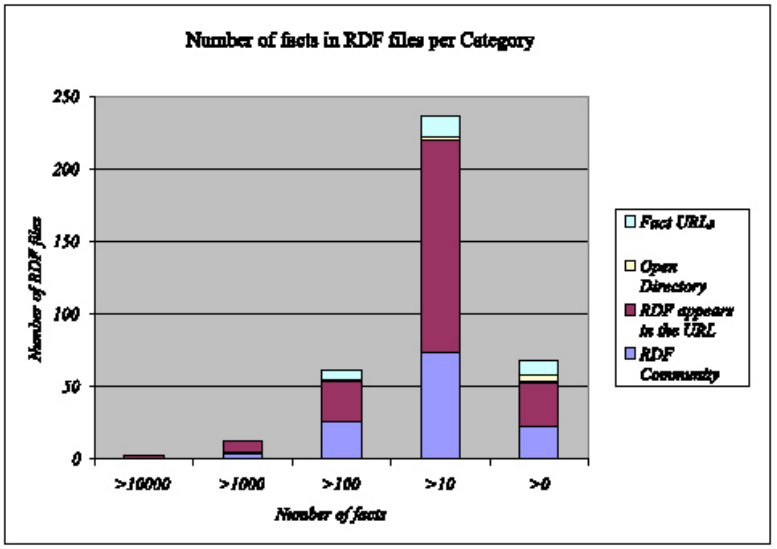

Figure 7: Distribution of the RDF data set sizes during the _rst search Dec. 2001

かなり多くのサイトが RSS (Rich Site Summary) Ver.0.9のデータを含んでいる。 RSS は Netcenter で提唱されたニュースヘッドラインを分散して配布するための軽量のシンジケートフォーマットである。RSS の例は次のようなブロックである:

```
<rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
            xmlns="http://my.netscape.com/rdf/simple/0.9/">
<rss>
        <channel>
                <title>BBspot</title>
                <link>http://www.bbspot.com</link>
                <description>Your Spot for Tech Humor</description>
        </channel>
...
```

3.4 典型的なネームスペースの利用

RDF データの見つかる確率と典型的なサイズがわかったので、さらにファクトを調べてみよう。正しくデータを解釈するには、エージェントが 3 つ組みにおける述語を正しく理解または解釈できなければならない。 ここで最も顕著に使われたのは、ダブリン Core メタデータの語彙である。テーブル 3 とテーブル 4 は、集まったファクトの中でどのくらい特定のネームスペースの prefix が現れたかをまとめたものである。最初の実験では、Ontoknowledge のケーススタディや、David Megginson の空港の例に関するネームスペースが多く現れているが、特定のサイトでしか使われていない。そこで、URL の異なりでなくホストの異なりで数えることにした。これはシステムではできないので、人手でチェックを行った。Wordnet やオープンディレクトリの述語は現れなかった。

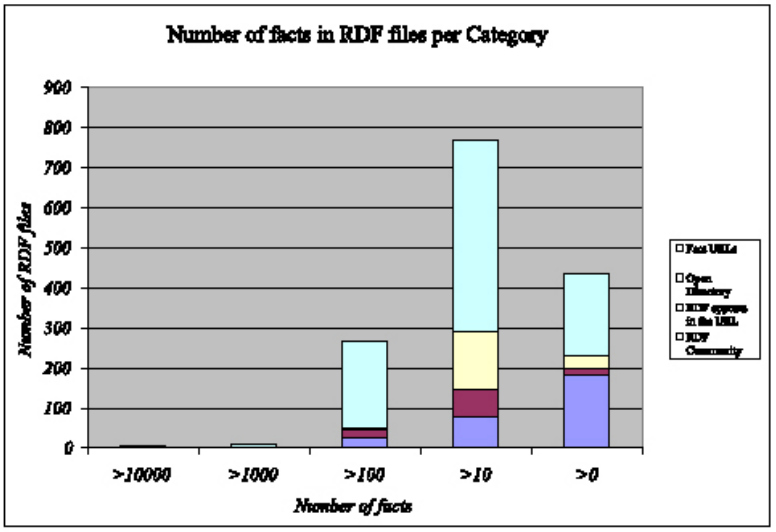

Figure 8: Distribution of the RDF data set sizes during the second search Aug. 2002

2 番目の実験でも同じような状況だった。多くのファクトには使われても、限られた数の文書にしか現われないネームスペースはあった。それ以外で、さらに W3C のものでもないものは、やはりダブリン Coreが一番多く、新しく現れたものとしては Adobe ネームスペースがあった。これらのページは Adobe XMP (eXtensible Metadata Platform)に基づくものである[1]。 XMP は、RDF で記述され、アプリケーションファイルに埋め込むメタデータとして設計されている。メジャーな IT 企業が RDF を使うということはセマンティック Web コミュニティにとって大きな励みとなっている。

Predicate namespace prefix	in # of URLs	in # of facts
http://www.ontoknowledge.org/oil/case-studies	1	23259
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	326	21011
http://www.w3.org/1999/02/22-rdf-syntax-ns#	326	17298
http://www.megginson.com/exp/ns/airports#	2	13589
http://alchemy.openjava.org/ocs/directory#	1	7014
http://www.w3.org/2000/01/rdf-schema#	62	6182
http://purl.org/	123	5198
http://interdataworking.com/vocabulary/	27	4698
http://www.trustix.net/schema/rdf/spi-0.0.1#	2	3012
http://my.netscape.com/rdf/simple/	93	2446
Other http://www.w3.org	331	2212
http://www.daml.org	27	2032
http://www.rpm.org	7	1716
http://metainfo.hauN.org	1	1351
http://home.netscape.com/	1	801
Other	164	13253

Table 3: Predicate namespace pre_xes used by the RDF data found during therst search Dec. 2001

エージェントが RDF ファクトを理解するには、述部だけでなく、よく参照されるオブジェクトも重要である。例えば、オープンディレクトリカテゴリにおけるウェブサイトに関するメタデータが良い例である。これによってオープンディレクトリを知っているどんなエージェントでも、例えば、サイトの内容を引き出すことができるだろう。 テーブル 5 はこの調査結果である。どちらの実験でも、オブジェクトのおよそ 57%がリテラル(文字列)であり、その多くが”en”,“text/plain” という文字列であった。RDF Type の述語の多くに見られるように、オブジェクトの多くは RDFS クラスである。多くの異なったサイトから頻繁にリンクされるクラス以外のオブジェクトは見つけられなかった。Wordnet やオープンディレクトリ以外で、良く参照されるリポジトリもなかった。

3.5 2 つの実験の比較

評価の前に、2001 年 12 月、2002 年 8 月の二回の実験の比較の傾向を分析して見よう。
全体的に見て、最後のカテゴリによって見つかった URL の数の違い(他のファクトから参照されるサイトが多い)を除けば大きな変化はない。これは RDF ファクトが比較的小さく閉じていることを示唆している。しかし、もっと細かく調べてみると、これらの大部分は数少ないソースから得られている。最初の実験では、15,214 個の異なったホストからの URL が見つかった。二番目の実験では、同じ比率にすると、異なりホスト数は 269 である。

Predicate namespace prefix	in # of URLs	in # of facts
http://www.cogsci.princeton.edu/	1	78445
http://www.w3.org/2000/01/rdf-schema	693	57132
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	1205	37926
http://orlando.drc.com/	19	27773
Other http://www.w3.org	435	11454
http://alchemy.openjava.org/	2	9793
http://purl.org/	463	9411
http://interdataworking.com/	16	5247
http://www.daml.org/	53	4490
http://ilrt.org/	9	2124
http://opencyc.sourceforge.net/	1	1630
http://ns.adobe.com/	152	1589
http://my.netscape.com/	34	902
http://www.rpm.org/	3	734
http://www.ontoknowledge.org/	2	645
http://dublincore.org/	82	544
http://www.omg.org/	3	523
http://www.semanticweb.org/	41	466
http://annotation.semanticweb.org/	5	375
http://xmlns.com/	48	351
http://example.org/	95	121
http://www.nesstar.org/	6	106
Other	129	3002

Table 4: Predicate namespace pre_xes used by the RDF data found during the second search Aug. 2002

RDF Object	in number of Facts Dec. 2001	in number of facts Aug. 2002
Other literals	58949	237163
Other resources	44562	175110
http://www.w3.org/1999/ 02/22-rdf-syntax-ns#	7646	7947
Numbers	8115	9667
en	2414	3278
hourly	2361	3265
text/plain	1002	1410

Table 5: Overview over RDF Objects

多くのページを集めたが、その多くがオープンディレクトリだったため、RDF の数も増えたが、総合的な RDF ページの比率は下がっている。ただし、違いはそれほど顕著ではないため、これから即結論とすることは早計である。

4 評価

この調査の結果、RDF はまだ大きなユーザコミュニティには使われていないことがわかった。探索範囲はそれほど広くないため、もしかすると RDF の大きな島を逃しているかもしれない恐れはある。しかし、RDF データは Web であまねく使われているわけではないことはわかった。それは、Web サービスにおいても同様と言えよう。技術的には Web サービスと RDF のインタフェースの組み合わせで利用できる何百万ものデータソースがあるはずである。
Web サービスの方がより一般に受け入れられているのは疑いもないだろう。しかし、Google ウェブサービス API やマイクロソフトの Map Point service13 のようないくつかの有名なサービスを除いて、現在 UDDI レジストリに入っているサービスの大半はプロトタイプレベルのものである。
今後、Web サービスやセマンティック Web で、自動処理が可能になり、状況は抜本的に変わるというのがビジネス上の見通しである。現在の Web は主として広告によって無料の情報提供がされている。しかし、今後は人によるアクセスではなく機械によるアクセスになれば、この状況は変わる必要がある。小額決済や一括支払いのような課金のバリエーションも必要となるだろうが、現状この方向について明確なトレンドや標準化はよくわかっていない。いずれにしても、Web の自動化が進めば、RDF はより重要なデータフォーマットとなるだろう。

RDF を HTML ページのメタデータフォーマットとして使うのは、すでに HTML には META タグがあるためあまり意味はない。RDF の強さ(すなわち、その拡張性と標準的な語彙とグローバルな識別子)を生かすこともできない。オープンディレクトリにおける一般のウェブサイトの検索結果が非常に貧弱なのを見ると、この考えはさらに確からしくなる。さらに、今回見つかったファクトにおいては、相互接続性は非常に低い(すなわち、ほとんどのオブジェクトが、リテラルだったり RDF スキーマ)。ダブリン Core と Adobe XMP ネームスペース以外では、非 W3C ボキャブラリはほとんど使われていない。ただし、Adobe の RDF のサポートは非常に有望な兆候である。

にもかかわらず、我々は、RDF には多くの可能性があると考える。例えば NEC CiteSeer リサーチインデクス[3]を見ると、Web 上のメタデータの必要性さらに対象を絞った検索の必要性がよくわかる。CiteSeer では、論文の引用関係を抽出している。被引用数は論文の質を測る指針である。
もし RDF メタデータを出版物につければ、こうしたシステムの実装はもっと容易になるだろう。さらに、様々な RDF アプリケーションも実現できるかもしれない。

我々は、セマンティック Web が成功するには、研究コミュニティが、より大きいユーザコミュニティがびっくりするようなアプリケーションを作ることが必要と考える。それこそが、現状の鶏と卵問題: アプリがなければデータもマークアップしないし、大量のマークアップデータがないとアプリが成功しない、というのを打開する可能性を秘めている。

エラーの原因

限定した探索では多くの RDF データが見つからないというのは別として、ページをスキャンしても RDF が見つからないというエラーがある。いくつかのケースを調べたところ、フォーマットが間違っていたり、ネームスペースが定義されていなかったりということがわかった。我々が実験で使用した RDF API ではこうした問題は、エラーとなってしまう。また、XML または XHTML ファイルを RDF API にかけると空データセットというエラーが返る。他の不具合は、ランダムに調べたところでは見つからなかった。

References

[1] Adobe Inc. A managers introduction to adobe extensible metadata platform, the Adobe XML metadata framework.
http://www.adobe.com/products/xmp/pdfs/whitepaper.pdf.

[2] M. Bergman. The deep web: Surfacing hidden value, 2001.

[3] K. Bollacker, S. Lawrence, and C. L. Giles. CiteSeer: An autonomous web agent for automatic retrieval and identi_cation of interesting publications.
In K. P. Sycara and M. Wooldridge, editors, Proceedings of the Second In 1998. ACM Press.

[4] A. Eberhart. Survey of RDF data on the web. In Proc. of the 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002), July 2002.

[5] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science,