情 報 シ ス テ ム 67－6
（1998．3．17）

WWWによる社内技術情報システムの開発と運用

片山 啓之

新日鉄情報通信システム(株)

最近の Internet 技術の一般化に伴い、各企業内の小グループが独自に蓄積した技術情報を自発的に WWW を利用して公開することが多くなってきた。そこで当社ではそれらの情報を収集し、分類し、再公開する「SUSIE」(SUggeStive Information Environment)を構築した。SUSIE では情報を Web ロボットプログラムを利用して収集し、それらをいくつかの分野に分類して提供する。これにより技術情報の整理と検索を容易にし、技術情報の再利用促進を図った。本論文では SUSIE の概要と開発上の検討事項、および運用における問題点とその解決策について報告する。

A Technical Information Indexing Tool Using WWW System

Hiroyuki Katayama

NIPPON STEEL Information & Communication Systems Inc.

With popularization of the latest Internet technology, the accumulation and the public presentation of spontaneous technical information using WWW system are growing in local area network at each section. Then we built "SUSIE" (SUggeStive Information Environment), a system which collects those information and classifies and re-exhibits them. In SUSIE, information is collected using Web robot program, classified and provided with some category. We meant that arrangement and reference of technical information became easier for an employee, and reuse of technical information was remarkably promoted by this system.

This paper reports the outline of a system, the devising point on development, the problem in practical use, and its solution.

1 はじめに

技術進歩の速度が加速するにつれて、個人の経験や知識のみでは技術知識の変化に追随していくことが困難になってきている。このため、製品情報や FAQ、事例など技術情報の流通をサポートするシステムの重要性が増大している。

そこで当社では「SUSIE」(SUggeStive Information Environment)と呼称するシステムを構築した。SUSIE では WWW を利用して自発的に公開されている技術情報を収集し、いくつかの分野に分類して提供する。

2 技術情報共有の現状

これまで、企業では知識の伝達は二つのルートで行われてきた。一つは組織間の職制に基づいた公的なルートであり、もう一つは私的な、個人と個人のコネクションによるルートである。以前に当社内で行ったアンケートでは、技術的な知識は必ずしも公的なルートを使用して伝播されておらず、むしろ私的なルートによって伝播されることが多いことが判明している。

公的なルートを利用すると、直接自分が知らない相手にも質問が届き、広い範囲に質問を投げかけることができる。また情報の内容の正しさについても、私的なものより信頼性が高い。しかし公的なルートは、直接面識の無い人や役職を意識しなくてはならない相手との会話が必要なため、比較的感情的なコストが高く、面倒だと感じることが多いようであった。一方、私的なルートでは自分の知っている範囲からしか情報を入手できないが、公的なルートに比較してよりきめ細かな内容を期待でき、またレスポンスが比較的速い。これが公的な問い合わせがあまり利用されない理由ではないかと推測される。

一方で、近年のインターネットの普及に伴い、企業内でもTCP/IPネットワークの構築が進んでいる。さらにWWWの普及と一般化によって、Webサーバは安価にかつ誰にでも構築が可能になった。個人や小グループによる Web サーバを用いた草の根の情報流通が、私的なルートや公的なルートに続く第3の知識の伝達手段として自然発生的に広まりつつある。

2.1 WWW を用いる利点

伝統的な社内情報システムでは、既存の社内の情報をデジタル化してネットワーク上で共有するための部署を置き、そこに一度すべての情報を集めてから公開するという手順を踏むことが多い。この方法では、担当部署によほど余力が無ければ情報流通のレスポンスが低下する。また利用する側に専用端末や専用ソフトが必要になり、情報を提供する側では規定フォーマットにあわせた情報を担当部署に提出する手順が必要となる。つまり情報の提供者と利用者にとって情報の提供や再利用の負荷が高く、情報の流通量は比較的少なかった。

WWW を利用することで、特定の部署が管理することなく、より発生源に近い部署や個人からの情報の流通が可能となる。これにより、情報の発生から公開までのレスポンスを向上させることが期待できる。また情報へのアクセス権限の管理を個々の提供者が設定できるため、情報を公開することへの感情的抵抗を比較的低くすることが期待できる。さらにブラウザで利用可能なため誰もが利用者となることができ、裾野を広げる効果が期待できる。

2.2 WWW を用いる問題点

WWW を利用した社内技術情報によって、今までアクセスしにくかった情報がより公開されやすくなり、またアクセスしやすくなることが期待できる。しかし WWW 自体には、検索のための仕組みが与えられておらず、必要な情報にアクセスするための手段がわかりにくいという欠点がある。

情報の所在を確認し管理するための仕組みとして、多くのブラウザにはブックマーク機能が備わっている。またブックマークを階層的に管理することの可能なものもあり、自分自身が既に見つけたものを後で再度閲覧したい場合などには便利に利用できる。

しかしブックマークでは、自分自身がまだ確認していない Web ページの中から欲しい情報を見つけ出すことはできない。またネットワーク上のデータが増加し、一人の人間がすべてを見てまわれないほど情報が蓄積されると、この方法ではカバーしきれなくなる。

3 SUSIE の概要

WWW の欠点を補う一つの解として、社内に情報を管理し流通させる部署を設定し、情報の収集と公開、広報や閲覧権限の設定などを行うという手段がある。だが当社では支社が全国各地にあり、それぞれがかなり独立的に運営されているので、情報の中央統制と管理を行う組織やシステムを構築することが難しい。またこのような方法では WWW による自発性や分散性を損ない、従来の公的な情報流通を単にシステム化しただけということになりかねない。

今回の取り組みでは、前提として各事業部や個人の自主性を優先し、データは as is で、収集と整理の工夫のみで問題解決することを目指した。すなわち、SUSIE では情報の生成や変化を指示したり管理することは行わず、あくまで情報の流通を補助する手がかりを提供することを目標とした。

これらの要求を検討すると、インターネット上で提供されている検索サービスに非常に類似している[1]。SUSIE ではサービス対象が社内に限定され、また内容は技術情報のみに限定したため、情報を事前に分類することが望ましいと考え、ディレクトリサービス型のシステムを構築した。

SUSIE は、大きく分けて三つの部分から構成されている(図 1)。

- ●Web ロボットによる情報の収集
- ●人間による情報の分類
- ●CGI プログラムによる情報の提供

まず Web ロボットがネットワーク上にある Web サーバから情報を収集する。収集された情報は URL-DB に格納される。

次に更新された URL-DB の各データに、分野 DB から分野情報を付与する。

格納されたデータは、利用者が SUSIE にアクセスした際にメニュー作成プログラムで検索を行い、表示される。利用者は必要な情報の URL を得る。利用者は直接当該 URL を参照することで情報自体を閲覧する。

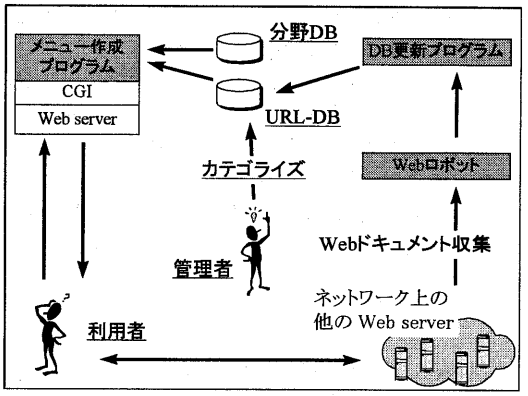


図 1 SUSIE の全体構成

なお今回は全文検索機能は利用しなかった。全文検索は適切なキーワードとその組み合わせを理解し、選択できる人にとっては、迅速に目的の情報を引き出すことができる非常に有用なツールである。しかし問題領域についての知識が少ない場合には過検索や誤検索が多く、あまり有効に利用できない。また比較的マシンパワーの必要な処理であり、ソースを公開して改造可能になっているツールも少なかった。

3.1 WWW ページの収集

情報の収集は、Web ロボットと呼ばれるプログラムを用いることにした。これによって情報収集作業を自動化できる。

Web ロボットは HTTP を使用して WWW サーバから HTML やその他のドキュメントを取得し、蓄積する。次に文中に含まれているリンク情報を分析し、さらにリンクされているドキュメントを取得する。これを繰り返すことで、ハイパーテキスト全体を取得しようとする [2]。

ハイパーテキストは人間の記憶や連想を表現しやすい、使いやすいデータ構造である。その反面、全体構造を把握したり、検索を行うことが難しい。Web ロボットはハイパーテキスト全体を一個所で収集するために考えられた仕組みである。

今回のSUSIEでは、WebロボットとしてUCIで作成されたMOMspider [3]を用いた。今回の目的には画像情報の収集は不要であるため、GIFやjpegなどの情報を取得しないように改造を行った。またMOMspiderは同時に複数のWebサーバへアクセスすることができないため、同時に複数個のロボットを起動して収集にかかる時間を短縮している。

3.2 URLデータ

収集したURLデータは、いくつかの分野に分類して利用者に提供する。分野は互いに親子関係を持ち、全体として階層構造をなすように設計されている。これによって、一つの分野に属するURLデータの数を減らすことができ、閲覧時の使いやすさを実現している。

URLデータはタイトルとURL、最終更新日時を蓄積しており、また分野情報を付与している。取得したファイルそのものは蓄積していない。

URLデータの分類作業は現在は人手によって行っており、初回の分類には2人で約4日程度を要した。

3.3 データの表示

分野の階層構造をたどって目的の情報を検索するプログラムは、CGIを利用している。このプログラムはURL-DBと分野DBを読み込み、指定された分野および関連する分野を表示し、次に分野情報の付与されたURLを表示する(図2)。

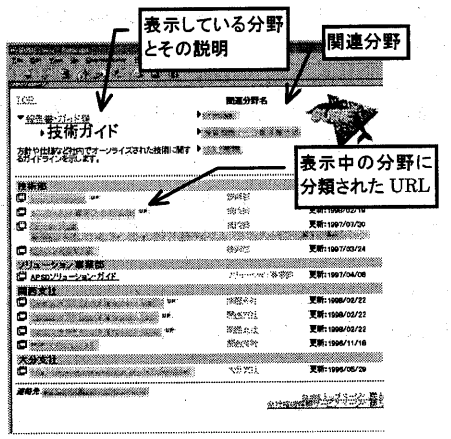


図 2 実行時画面

4 実施上の課題と解決

4.1 分野情報の作成

収集したURLの分類に使用する分野情報の作成については、以前に当社内で策定された分類や業務標準から再利用したものに拡張を行って作成したが、かなりの時間を費やした。情報の分類は図書の分類や百科事典、生物の分類などで古くから研究されてきたテーマだが、それらとは分類後の格納形式や分類の対象が大きく異なるためシンプルなtree構造になりにくく、交叉分類も多発する。さらに分類自体の変化も多く、固定的にできない。このため、図書分類等は今回の目的には不向きである。

SUSIEでは、利用者が社内に閉じていることと対象を技術的なものに限ったことから、前述の分類法でも十分理解しやすくできると判断した。しかし今後、アンケートやアクセスログを分析して分類を適宜変更していく予定である。

また分野間の関連の変更を容易にするため、分野DBのデータ形式は複数の親を持つ分野が作れるように設計し、交叉分類の発生に対応している。またURL情報に関しても、複数の分野に関連付けが行えるようにした。

4.2 URL分類作業

現在は月に一度、変更のあったページと新規ページに関してのみ分類作業を行い、2人で1日程度の作業が発生している。今後更に情報が増えた場合には、人間が当該URLを閲覧して分類を決定する方法では限界がある。これをなるべく自動化することで省力化をはかり、また更新を頻繁に行うことでup to dateな情報提供を行う方法が望まれる。このため自動分類のためのプログラムを開発し、既に動作確認を終えている。しかしキーワードの検索と重み付けや、類義語などの機能は全文検索と共に使用するのが効果的であると考え、実際には使用していない。今後テストを行って検証する予定である。

4.3 データ表示プログラム

当初、SUSIEのデータ表示部分はPerl言語で作成した。テスト時には特に問題はなかったが、データが増えるに従ってレスポンスが悪化したため、C言語で作りなおした。

レスポンスの悪化は、

- インタプリタ言語であるPerl言語を使用した
- アクセスのたびに起動される単純なCGIスクリプトとして作成したため、起動するたびにファイルの読み込みと検索を行っていた

以上の2点が原因と考えられる。

今回はプログラム言語の変更のみで対応したが、データがさらに増えた場合には、常時動作するデーモンとして実装し、CGIによるモジュールはデーモンプロセスとやり取りを行うように変更する必要があると考えている。

5 運用からの評価

現在のSUSIEは、毎月第一週の週末に社内31のWWWサーバから約5300URLを収集し、270URLに対してリンクを行い、一日に約280hitのアクセスがある。このシステムを運用する上での問題点と解決策について報告する。

5.1 Webページの品質と内容

社内のWebページを閲覧し分類していく段階で問題となったのが、公開されているデータの質である。たとえば公開されている文章に古くなった情報や事実誤認がある場合、そのままSUSIEからリンクを行うのは問題がある。また修正を要求しても、必ずしも迅速には行われない。

また類似の問題として、データがHTMLとして必要な要件を満たしていないことがある。最近よく使用されるビジュアルなHTML作成ツールで作成されたページでは、タイトルが適切につけられていないことが多い。いわゆる半角カタカナや、漢字の入ったURLが使用されていることもある。

インターネットにおけるディレクトリサービスであれば、内容に問題があると判断したページはリンクしないことで解決できると思われる。しかしSUSIEは社内システムであること、また現在は社内のWWWによる情報がさほど多くないことから、このような問題のあるページも貴重な情報源と考えてリンクを行い、その後に修正依頼を行っている。

今後十分にWeb情報が蓄積された時点で、このような問題のある情報には修正依頼を出し、修正が確認されてからリンクを行うという手順に改めることで、提供する情報の信頼性を確保する。

5.2 分類に対する評価

SUSIEを構築後約1ヶ月間、試行期間として社内各部署の約30人に対して先行公開を行い、結果をアンケート調査で確認した。

機能、構成、操作性などについてはほとんどの回答者が問題ないとしており、いくつかの操作性などに関する指摘も十分対応可能なものだった。ただ、アンケート結果の中で半数程度の人が「実際のプロジェクトに関する情報が不十分である」「プロジェクトに関する分類が大雑把すぎ、また実状を反映していない」という指摘を行っている。

今回のSUSIEは、技術的な情報に関してのみをWWWページから収集して、社内で使用されていた技術分類を利用し、開発者側で分類して提供した。一方、利用者側ではプロジェクト単位で業務を行うことが多いため、プロジェクト毎に分類された情報を求めるのではないかと推測される。また利用者がプロジェクト管理や実際のプロジェクトに関する情報を強く求めていることは、分野毎のアクセス頻度グラフからも確認できた。

分類はこのような情報システムにおいては開発者と利用者の接点となる部分である。今後はユーザにあわせて分類を調整し、またプロジェクトに関する情報の充実を図る必要がある。

5.3 絞り込み型の表示

各々の情報が増えるに従って、ある分野を選ぶと数十以上のサイトがマッチすることになる。既に一部の分野については30を超えるURLがヒットするため、分類として適切でない状態になっている部分がある。

一つの分野の URL を減らすためには、階層を増やしてサブ分野を設けることで対応している。しかし分野の階層を増やすと、必要とする分野が見つけにくくなったり、たどり着くための経路が遠くなり、必ずしも良い方法とはいえない。また階層状の分野構成は、分野に対する捉え方の違う人間にとってはむしろ理解しにくい事がある。

SUSIE では、利用者を社内に限定したこと、および技術情報のみに限定したことから、比較的分野構造が理解されやすいと考えている。しかし将来的に情報量が増加すると問題になる可能性がある。

解決策として、絞り込み型の分野検索機能を検討している。複数の分野を指定することでそれらすべてを含む情報を表示したり、関連する情報分野を見つけ出すもので、これによってより的確な検索を行えるのではないかと考えている。

5.4 Web ロボット

SUSIE では情報の収集に Web ロボットを使用したが、Web ロボットにはいくつか問題とされる点があり、これを考慮したものを使用する必要がある。しかし、現在ソースを公開しているロボットには、完全に満足できる仕様のものは存在しなかった。また市販の全文検索プロダクトでも、Web ロボットのような情報収集の仕組みまでを含むものはいまだに数が少ない[4][5]。

今後 SUSIE のために Web ロボットを開発することを検討しているが、その際に対応が必要な点について考察する。

5.4.1 robots.txt

SUSIE では Web ロボットを用いて公開されているデータをローカルにコピーし、そこからタイトルやリンク情報を取得している。しかし、たとえばアクセスがあるたびに自動的に作成されるページや、純粋に個人的な用途のページなどはコピーされたくないという要望も存在する。

Web ロボットによってデータを収集されることを拒否したいという要望に対して、解決策として規定されているのが、Robots Exclusion Protocol[6] によって規定された robots.txt である。

robots.txt には、アクセスされたくない URL とアクセスされたくないロボット名を記述する。Web ロボットは最初に robots.txt にアクセスし、許可された URL にのみアクセスする。これが Robots Exclusion Protocol であるが、そのような仕様を無視したプログラムは簡単に作成できるし、事実これを考慮していないプログラムは多い。

Robots Exclusion Protocol はいまだに RFC には規定されておらず、あくまで提案に過ぎない段階だが、なるべくこれに従った Web ロボットを使用することが望ましいだろう。

```
User-agent: webcrawler
Disallow:

User-agent: lycra
Disallow: /

User-agent: *
Disallow: /tmp
Disallow: /logs
```

robots.txt の例

またこれとは別に、特に明確な理由がないのにあらゆるアクセスを拒否するような robots.txt を記述してある場合もある。SUSIE を運用する際にも、「ロボットを拒否する方法を試したい」という理由だけで、すべての Web ロボットを拒否するような robots.txt が作成されていたことがあった。同じ社内なので管理者に依頼して当該ファイルを削除したが、この機能が濫用されると Web ロボットが役に立たなくなる。今後も広報などの手段で Web サーバ管理ポリシーについて啓蒙を行う必要がある。

5.4.2 収集するファイルの指定

URL 情報を集める際に、Web ロボットに対してリンクされているすべてのファイルを収集するように指定すると、画像ファイルや実行可能バイナリなどといったものまで収集してしまう。このようなファイルまで収集していては、ネットワーク負荷や時間的な負荷、そしてローカルのディスクの負荷が高くなりすぎるので拡張子などで判別し、HTML ファイルやテキストファイルのみを収集するように指定できるロボットが望ましい。

5.4.3 FRAMEへの対応

HTML の FRAME 機能は使い方によってはWeb サイトの見通しをよくし、閲覧性向上に高い効果がある。しかし Web ロボットの多くはこれに対応しておらず、FRAME を利用したページから先のリンクをたどることができなくなる。今後は FRAME を使用したサイトの増加が予想され、対応の重要性が増してくるため、今後のロボットには必要な機能だと考えている。

5.4.4 HTTP プロトコル

Web ロボットは Web サーバに HTTP でアクセスし、情報を収集する。このため、正しく HTTP に対応しているかどうかは Web ロボット選択の上で重要である[7][8][9]。

- HTTP の Referer:ヘッダはクライアントからサーバへのリクエストの際に使用され、そのリクエストがどの URL からのリンクをたどって行われているかを意味する。このヘッダを正しく付与することで、サーバ管理者は自サイトへのリンクを検知することができる。
- From:ヘッダを正しく設定することで、Web ロボットの使用者のメールアドレスをサーバ管理者に伝えることができる。
- HTTP には、「If-Modified-Since ヘッダ」が規定されている。これはクライアントからサーバへのリクエストの際に使用され、「リクエスト対象の URL が指定した日付以降に更新されている」場合にのみデータを取得するというヘッダである。これを利用することにより、クライアント側に蓄積されたデータをキャッシュとして利用し、トラフィックを減らすことができる。

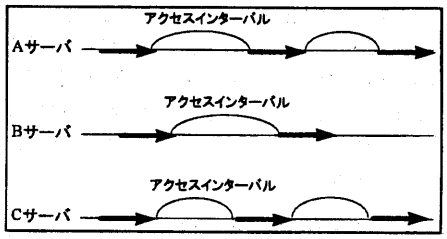
これらのヘッダを Web ロボットが正しく設定し、Web サーバへのリクエストに際して使用することで、アクセスされる Web サーバ管理者に便宜を図ることができる。礼儀正しい(ethical)振る舞い[10]をする Web ロボットを使用することで、ターゲットとなるサーバに問題を起こさないように注意し、円滑な運用を心がけるべきである。

5.4.5 アクセスインターバル

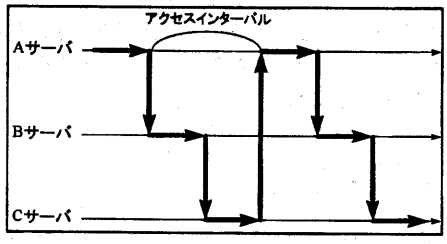
不用意な Web ロボットによくある問題が、ラピッドファイヤ[7]と呼ばれる現象である。これは短時間に繰り返し同一サーバに対してリクエストを行う行動を指す。これはサーバの過負荷を引き起こすため、一つのサーバに対する連続したリクエストはインターバルを置いたり、同時に複数のリクエストを行わないなどの配慮が必要である。しかし間隔を長く取ると、多数の情報を集めたい場合に非常に時間がかかってしまい、実用的ではない。

同一サーバに対するアクセスインターバルの指定ができるロボットでは、

- 各スレッドは一つのサーバにのみアクセスし、同時に複数のスレッドが走る。各スレッドは設定したアクセスインターバルを取る。ただし、このような場合にはいくつものスレッドが同時に実行されるとマシンの負荷が高くなりすぎる恐れがあり、同時に走るスレッド数の上限を指定できるような実装が必要になる。



- 単一スレッドで実行するが、あるサーバへのアクセスインターバルを取っている間に他のサーバに対するアクセスを行う。



などの手法を取り、全体として処理の速度を上げる必要がある。

今回 SUSIE で使用した MOMspider は、同時に一つのサーバにのみアクセスするタイプであったため、平行して複数のプロセスを実行するように運用している。

5.5 今後の目標

SUSIE は一般公開も果たし、運用も落ち着きつつある。しかし現在の状態で十分に社員の要求を満たしているとは言えず、今後もさまざまな機能拡張を行っていかねばならない。

要求の多い機能の一つが、全文検索機能である。最近、実績のあるツールのソースが公開されつつあり [11]、これを利用して全文検索サービスを行う予定である。ただ適切な WWW ロボットがないことを問題と考えており、今後は独自に開発することを検討している。

次に考慮しているのが、不定形かつ小規模の情報の収集である。現在のように WWW で公開されている HTML ファイルを収集するような運用形態では、収集する情報の粒度は比較的大きくなる。このため、より日常的な粒度の小さな技術情報、具体的にはテキストのみで数行で記述できるような情報はカバーしきれない。このような情報を収集するためには、掲示板的なサービスが有効ではないかと考えられる。

また社内の一部の部署では情報提供に関する報奨制度が既に開始されている。これは情報を見たい人間がその都度社員番号と名前を入力して投票を行い、評価の高かった情報に関しては報奨金を支給する制度である。実施部署では順調に成果を上げており、今後はこの制度の対象範囲を拡大し、社員の情報共有に対する意識を改善することが望ましい。

6 おわりに

SUSIE では、比較的少ない労力で社内の WWW ページに対する有益なインデックスを作成できるようになった。WWW は個人の自発性と草の根的な運用の効果を全体で受益できるところに特徴があるが、SUSIE によって、ボトムアップという WWW の利点を生かしたまま広い範囲での情報の流通促進を実現できた。今後は全文検索機能の追加や自動分類機能などを追加して、更に効果的なシステムを構築していく。

7 参考文献

[1] 原田昌紀:「サーチエンジン徹底活用術」, オーム社, 1997

[2] The Web Robots Pages
<URL:http://info.webcrawler.com/mak/projects/robots/robots.html>

[3] MOMspider: Multi-owner Maintenance Spider
<URL:http://www.ics.uci.edu/pub/websoft/MOMspider/>

[4] 日本語全文検索エンジンソフトウェアのリスト
<URL:http://www.kusastro.kyoto-u.ac.jp/~baba/wais/other-system.html>

[5] The Web Robots Database
<URL:http://info.webcrawler.com/mak/projects/robots/active.html>

[6] The Robots Exclusion Protocol
<URL:http://info.webcrawler.com/mak/projects/robots/norobots-rfc.txt>

[7] The Web Robots FAQ
<URL:http://info.webcrawler.com/mak/projects/robots/faq.html>

[8] 私の探知したロボット君
<URL:http://shidahara1.earth.s.kobe-u.ac.jp/~takawata/robots.html>

[9] User-Agent についてのたわごと
<URL:http://www.dais.is.tohoku.ac.jp/logs/agentgripes.html>

[10] Ethical Web Agents
<URL:http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Agents/eichmann.ethical/eichmann.html>

[11] 日本語全文検索エンジン「Freya」
<URL:http://kichijiro.c.u-tokyo.ac.jp/odin/freya/>