ポータルサイト向け情報検索エンジン

阿部 真也
地方独立行政法人 東京都立産業技術研究センター
abe.shinya@iri-tokyo.jp

概要　本稿では，ポータルサイト向けの情報検索エンジンを提案する．本検索エンジンは，公設試験研究機関のポータルサイト上に設置され，各公設試の Web サイトに分散する実験設備の情報を一括検索するものである．設備情報は，Web クローラによって各 Web サイトから自動収集される．クローラが訪問する URL を限定することで，設備情報以外の情報をできる限り排除する．既存のポータルサイトに見られる横断検索と比較して，本検索エンジンの方が，検索精度の面で優れていることを示す．さらに，本検索エンジンの適用範囲について考察し，約 94%の公設試験研究機関に適用可能であることを示す．

キーワード　ポータルサイト，情報検索エンジン，公設試験研究機関

1　はじめに

公設試験研究機関いわゆる公設試は，地方自治体が設置する試験研究機関であり，地域企業への技術的コンサルティングを通して，その地域に資することを目的に設置された機関である．公設試の業務の一例として，東京都立産業技術研究センターにおける業務方法書を図 1 に示す．図 1 の第 3 条に示した業務を依頼試験と呼び，顧客から製品やその部品等の試験品を受け，その特性の測定や分析を行い，結果を証明書として発行する業務である．第 7 条に示した業務を機器利用と呼び，設備の利用を提供する業務である．依頼試験と異なり，顧客自身が設備を操作する．他に，産業技術に係る研究等の業務があるが，ここでは割愛する．

近年，地方自治体の経営資源の制約や，顧客からの多様な支援要求から，公設試が自治体の枠を越えて連携することが求められている．これを一般に広域連携と呼ぶ．中小企業庁 [1] は，今後の公設試のあるべき姿として，第一に地域企業に対する技術支援の拡充，第二に産学官連携や広域連携の推進を挙げている．広域連携のあり方には，各公設試の組織は独立で維持しつつ案件を相互に紹介するという入り口段階のものから，事実上 1 つの組織として人材や設備を共同運用するという踏み込んだものまで提案されている．全日本地域研究交流協会 [2] は，工業系公設試の定員は 30 から多くとも 50 人強であり，今後も拡大が予想される工業全体を支援対象にすることは困難であるという理由から，広域連携による技術支援分野の分担補完の必要性を指摘している．また，谷口 [3] による公設試の技術相談に関する調査では，6 割以上の公設試が技術分野の拡大に苦慮しており，他の公設試への問合せを要する案件が増えていることが明らかにされている．谷口は，これに対処する方法の 1 つとして，外部との連携による技術相談業務の拡充を挙げている．以上のように，公設試経営における直近の課題として，広域連携が求められている．

（試験に関する業務）

第 3 条　法人は，依頼に応じて，産業技術に係る試験（以下「依頼試験」という.）を実施することができる．

（中略）

（試験機器等の設備及び施設の提供に関する業務）

第 7 条　法人は，依頼に応じて試験機器等の設備及び施設を貸し付けることができる．

図 1 東京都立産業技術研究センターの業務方法書

さて，ある公設試の職員が自ら所属する公設試では対応できず，他の公設試に問合せすべきと判断した場合，その職員は問合せすべき公設試を特定しなければならない．近年は，Web サイトの利用が一般的である．この場合，各公設試の Web サイトを 1 つ 1 つ順に参照すれば，問合せすべき公設試を特定可能である．だがもし，各公設試の情報を 1 つに集約したポータルサイトがあれば，そのサイトは職員にとって有用である．そこで筆者らは，首都圏の公設試における広域連携の推進を目的としたポータルサイト [4] の開発を進めている．このポータルサイトでは，依頼試験や機器利用といった業務の連携を図るために，各公設の保有設備を一括検索するための検索エンジンを提供している．職員は，この検索エンジンを利用することで，必要な設備を保有する公設試をワンストップで特定できる．このポータルサイトは，職員だけでなく公設試を利用しようとする顧客にとっても

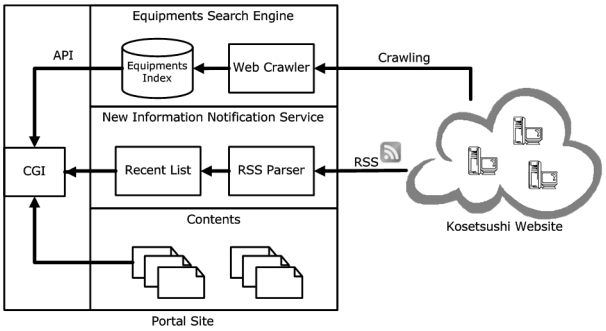

図 2 公設試ポータルサイト

有用である．したがって，本稿では，公設試の職員と顧客を対象ユーザと定義する．

以降，2 章では，筆者らが開発したポータルサイトの構成と，検索エンジンの実装について述べる．3 章では，検索エンジンの検索精度および適用範囲の評価を行う．4 章では，本検索エンジンの事業化例である，首都圏テクノナレッジ・フリーウェイの検索サービスについて述べる．5 章で本稿のまとめと今後の課題について述べる．

2 構成と実装

前章では，公設試経営における直近の課題として広域連携が求められていること，その推進には公設試ポータルサイトが有用であることを述べた．本章では，筆者らが開発したポータルサイトの構成と，その上で運用している検索エンジンの実装について述べる．

2.1 ポータルサイトの構成

ポータルサイトの全体構成を図 2 に示す．このポータルサイトは，設備検索エンジン，新着お知らせ，表示用 CGI，その他のコンテンツから成る．設備検索エンジンは，各公設試の Web サイトから設備情報を収集し，一括検索する機能である．新着お知らせは，各公設試が配信する RSS を解析し，それを一括表示する機能である．いずれの機能においても，更新情報はポータルサイトに自動で反映されるので，各公設試は自らが保有する Web サイトのみ管理すればよい．

2.2 設備検索エンジン

ここでは，設備検索エンジンの実装について述べる．設備検索エンジンは，オープンソースの全文検索システムである Hyper Estraier[5, 6] を用いて実現されている．図 2 の Equipments Index は設備情報の検索インデックスであり，その実体は Hyper Estraier のインデックスファイルである．Web Crawler は各公設試の Web サイトをクローリングし，インデックスファイルを更新するプログラムであり，Hyper Estraier に標準で付属するものを利用する．つまり，ここでいう設備検索エンジンは，設備情報そのものを管理するのではなく，設備情報が得られる外部ページへのリンクを管理する．

インデクシングが可能か否かは，Hyper Estraier が対応している文書形式に依存する．Hyper Estraier が対応している文書形式を以下に示す．

- プレーンテキスト（.txt，.text，.asc 等）
- ハイパーテキスト（.htm，.html，.xhtml，.xht 等）
- 電子メール（.eml，.mime，.mht，.mhtml 等）
- MS-Office（.doc，.xls，.ppt 等）
- PDF（.pdf）
- DocuWorks（.xdw）

プレーンテキスト，ハイパーテキスト，電子メールは，そのままの形式でインデクシング可能である．また，標準付属するフィルタによって中間ファイルに変換することで，PDF やマイクロソフト社の Microsoft Office®，富士ゼロックス社の DocuWorks™ もインデクシング可能である．つまり，各 Web サイトに掲載されている設備情報が，上記のいずれかの形式であれば，インデクシング可能である．

インデックスの作成にあたり，設備情報以外の情報が混在しないようにしなければならない．本サイトでは，クローリング対象とする URL を限定することで，これに対処している．例えば，ある公設試の Web サイトが次のような構成になっているものとする．

- foo.com/
  - jigyou/ 事業に関するコンテンツ
  - setsubi/ 設備に関するコンテンツ
  - gijutsu/ 技術に関するコンテンツ

この場合，クローリング対象とすべきは setsubi/配下であり，それ以外はクローリング対象とすべきでない．このような条件下でクローリングするための設定を以下に示す．

```
seed: 1.0| http://foo.com/setsubi/index.html
allowrx: ^http://foo\.com/setsubi/
```

1 行目はクローリング開始点の設定であり，2 行目は setsubi/の配下のみクローリング対象とするための設定である．このような設定を各公設試の Web サイトに合わせて記述し，クローラを実行することで，その時点で最新の設備情報がインデクシングされる．

3 評価

前章では，筆者らが開発したポータルサイトの構成と，設備検索エンジンの実装について述べた．本章では，設備検索エンジンの検索精度と適用範囲を評価する．

表 1 実験環境・条件

対象公設試	埼玉県産業技術総合センター 千葉県産業支援技術研究所 東京都立産業技術研究センター 神奈川県産業技術センター
キーワード	造形 万能試験機 耐候性 雷サージ Ge 半導体 三次元測定 3D スキャナ 分光 エミッション 赤外線
正解判定	目視
検索エンジンの設定	図 3 のとおり

3.1 検索精度

ここでは，提案した設備検索エンジンの検索精度について評価する．まず，検索精度の指標となる適合率と再現率について述べる．適合率は，正確性の指標であり，検索ヒット数（検索結果として得られた文書の総数）に対する正解ヒット数（検索結果として得られ，かつ検索要求を満たす文書の総数）の比で表される．検索ヒット数を N，正解ヒット数を R とすると，適合率は R/N となる．一方，再現率は，網羅性の指標であり，正解文書数（検索要求を満たす文書の総数）に対する正解ヒット数の比で表される．正解文書数を C，正解ヒット数を R とすると，再現率は R/C となる．

実験環境や条件を表 1 に示す．評価対象の公設試は首都圏の 4 公設試である．評価に用いるキーワードは，東京都立産業技術研究センターの Web サイトにおける検索ワードランキングから実験設備に関するワードを抽出し，その上位 10 ワードとする．正解文書か否かの判定は，具体的な設備情報，たとえば機器分類，メーカ，型番，仕様等が文書に含まれているかどうかを目視によって行う．Web クローラの設定は図 3 のとおりである．

この条件下でクローリングと検索を実行したときの適合率を表 2 に示す．表中の横断検索の列は，既存のポータルサイトによく見られる横断検索による結果であり，提案手法の列は，提案した設備検索エンジンにおける結果である．それぞれの方法で検索したときの検索結果として得られた全文書数 N，検索結果として得られた正解文書数 R，適合率 R/N を示している．いずれのキーワードにおいても，設備検索エンジンの方が適合率が高いことが分かる．

```
# Saitama
seed: 1.0 | http://www.saitec.pref.saitama.lg.jp/kaihou/kikilist.html
seed: 1.0 | http://www.saitec.pref.saitama.lg.jp/kaihou/denjiha_yoyaku.html
allowrx: ^http://www.saitec.pref.saitama.lg.jp/kaihou/kikilist.html

# Chiba
seed: 1.0 | http://www.pref.chiba.lg.jp/sanken/kikisetsubi/index.html
allowrx: ^http://www.pref.chiba.lg.jp/sanken/kikisetsubi/

# Tokyo
seed: 1.0 | http://www.iri-tokyo.jp/setsubi/index.html
allowrx: ^http://www.iri-tokyo.jp/setsubi/

# Kanagawa
seed: 1.0 | http://www.kanagawa-iri.go.jp/equipment.html
allowrx: ^http://www.kanagawa-iri.go.jp/equipment/
```

図 3 設備検索エンジンの設定

次に再現率の比較を行う．まず，再現率を具体的な実数値として求めることは難しい．なぜならば，検索システム用テストコレクションなど特殊な場合を除き，C を把握することが困難だからである．ただし，再現率の大小比較は可能である．検索対象とする文書集合（この場合は対象公設試の Web サイト上にある全ての文書）が等しいならば C も等しいので，R の大小によって比較が可能である．表 2 から，いずれのキーワードでも両者の R は等しく，したがってまた再現率も等しいことが分かる．

以上より，適合率は高く，再現率は両者等しいゆえ，設備検索エンジンの方が検索精度が良いといえる．

3.2 適用範囲

ここでは，設備検索エンジンが全国の公設試の何割に適用可能であるかを評価する．調査対象とする公設試は都道府県が設置した工業系公設試とし，1 県あたり 1

表 2　適合率

	横断検索			提案手法		
キーワード	N	R	R/N	N	R	R/N
造形	358	5	0.01	11	5	0.45
万能試験機	416	11	0.03	19	11	0.58
耐候性	162	8	0.05	9	8	0.89
雷サージ	83	7	0.08	10	7	0.70
Ge 半導体	44	1	0.02	2	1	0.50
三次元測定	477	15	0.03	26	15	0.58
3D スキャナ	11	1	0.09	2	1	0.50
分光	201	8	0.04	8	8	1.00
エミッション	130	7	0.05	8	7	0.88
赤外線	411	13	0.03	17	13	0.76

機関を無作為に抽出する．以下に，対象公設試の一覧を示す．

- 北海道・東北（7 機関）
  北海道立総合研究機構，青森県産業技術センター，岩手県工業技術センター，秋田県産業技術センター，山形県工業技術センター，宮城県産業技術総合センター，福島県ハイテクプラザ．
- 関東（7 機関）
  茨城県工業技術センター，栃木県産業技術センター，群馬県立産業技術センター，埼玉県産業技術総合センター，千葉県産業支援技術研究所，東京都立産業技術研究センター，神奈川県産業技術センター．
- 中部（9 機関）
  新潟県工業技術総合研究所，富山県工業技術センター，石川県工業試験場，福井県工業技術センター，山梨県工業技術センター，長野県工業技術総合センター，静岡県工業技術研究所，愛知県産業技術研究所，岐阜県産業技術センター．
- 近畿（7 機関）
  三重県工業研究所，滋賀県工業技術総合センター，京都府中小企業技術センター，大阪府立産業技術総合研究所，兵庫県立工業技術センター，奈良県工業技術センター，和歌山県工業技術センター．
- 中国（5 機関）
  鳥取県産業技術センター，島根県産業技術センター，岡山県工業技術センター，広島県立総合技術研究所，山口県産業技術センター．
- 四国（4 機関）
  徳島県立工業技術センター，香川県産業技術センター，愛媛県産業技術研究所，高知県工業技術センター．
- 九州・沖縄（8 機関）
  福岡県工業技術センター，佐賀県工業技術センター，長崎県工業技術センター，熊本県産業技術センター，大分県産業科学技術センター，宮崎県工業技術センター，鹿児島県工業技術センター，沖縄県工業技術センター．

2 章で述べたように，プレーンテキスト，ハイパーテキスト，電子メール，MS-Office，PDF，DocuWorks のいずれかの文書形式で，設備情報が Web 上に公開されていれば，設備検索エンジンを適用可能である．対象公設試の Web サイトを調査した結果，全 47 機関中 44 機関は設備情報を上記のいずれかの形式で Web 上に公開していることが分かった．また，該当サイトを実際にクローリングし，正常にインデクシングされることを確認した．

以上より，設備検索エンジンは，約 94%の工業系公設試に対して適用可能である．

4　事業化

前章では，設備検索エンジンの検索精度と適用範囲を評価した．本章では，本検索エンジンの事業化例である首都圏テクノナレッジ・フリーウェイの検索サービスについて述べる．

4.1　首都圏テクノナレッジ・フリーウェイ

首都圏テクノナレッジ・フリーウェイ（Metropolitan Techno Knowledge Freeway: TKF）[7] は，埼玉県産業技術総合センター，千葉県産業支援技術研究所，東京都立産業技術研究センター，神奈川県産業技術センター，横浜市工業技術支援センターの 5 つの公設試からなる連携体である．

TKF の主要事業の 1 つに，TKF Web サイト [8] の運営がある．TKF Web サイトは，顧客への情報提供の一元化を目的にしたサイトである．TKF Web サイトの主な機能を次に示す．

- 設備検索
- お知らせ機能
- 相談フォーム

設備検索は，各公設試の設備情報を一括検索するサービスである．図 4 に検索フォームのブラウザイメージを示す．このサービスは，本稿で提案した設備検索エンジンを用いて実現されている．ユーザは，この検索サービスを利用することで，必要な設備を保有する公設試をワン

図 4 TKF Web サイトの検索フォーム

ストップで特定できる．その他に，各公設試からの新着情報を一括表示するお知らせ機能，問合せ窓口を一元化した相談フォームがある．

4.2 事業化に至る経緯

ここでは，本稿で提案した設備検索エンジンが事業化に至った経緯について述べる．ここで述べることは，あくまで TKF における経験に基づくものであり，必ずしも全国の公設試にあてはまるものではない．

初期の TKF Web サイトでは，手動によるデータ保守が行われていた．具体的には，TKF Web サイトの管理者が，各公設試から設備情報を保存した外部記録媒体を受け，その情報をもとに保守するという運用である．運用当初から挙げられていた課題は，保守作業が負担になることに加え，更新の遅延が発生することである．ここでいう更新の遅延とは，各公設試において設備の導入または廃止が行われてから，その情報が TKF Web サイトに反映されるまでの期間をいう．この遅延は，すでに設備を廃止したにもかかわらず，設備情報を掲載している期間が生じることを意味しており，顧客トラブルに発展しかねない．

そこで，遅延を最小限にするために，データ保守をサイト管理者に一任するのではなく，図 5 に示す保守用フォームを設け，各公設試がそれぞれ保守を行う運用に改めた．その結果，遅延は改善されたが，新たな問題が指摘された．それは，将来的に外部ページの編集が禁止される公設試の存在である．TKF Web サイトは，東京都立産業技術研究センターが保有するサーバ上にあるので，それ以外の公設試から見れば外部ページであり，禁止対象となる．この制限は，自治体における情報漏洩防止策の一つとして計画されたもので，自治体の設置機関である公設試もそれに従わざるを得ない．ゆえに，各公設試は自身の Web ページのみ編集が許可されるという

図 5 旧サイトの保守用フォーム

条件下で，運用を再検討せざるをえなくなった．

その代案として，設備情報更新用の XML フォーマットを定め，これを利用して更新する方法を検討した．フォーマットの例を図 6 に示す．各公設試がフォーマットにのっとって設備情報をまとめ，それを自身の Web サイトに追加する．TKF Web サイトは，このファイルを解析し，それをデータベースに反映させる．こうすることで，外部ページを編集することなく，データ保守が可能となる．ところが，この方法であっても対応できない公設試が存在した．それは，自治体が保有するコンテンツ管理システムを用いて Web サイトを運営している公設試である．自治体が運用するコンテンツ管理システムでは，公設試の裁量で編集可能なページは限られるので，図 6 のようなファイルを追加できない．

ここまでで課せられた制約条件を以下に示す．

条件 1 公設試の裁量で編集可能な Web ページは，自身の Web サイト上にあるページのみである．

条件 2 公設試の裁量でファイルを追加することは，自身の Web サイトであっても不可能である．

これらの打開策として，Web クローラを利用して，設備情報が掲載された Web ページへのリンクを管理する方法を検討した．これが，本稿で提案した設備検索エンジンの原型である．この方法は，外部ページの編集を要しないので，条件 1 を満たす．また，各公設試の Web サイトにファイルを追加する必要もないので，条件 2 も満たす．さらに，各公設試は，自身の Web ページのみ編集すれば，TKF Web サイトにも自動反映されるので，

```
<?xml version="1.0" encoding="UTF-8"?>
<equipments pref="tokyo">
  <equipment>
    <name> 100kV 電子ビーム描画装置 </name>
    <spec> 加速電圧:100kV </spec>
    <spec> 最小線幅:8nm </spec>
    <spec> フィールド継ぎ精度:40nm </spec>
    <model> ELS-7000Ac </model>
    <year> 2011 </year>
    <remarks 事前講習必須 </remarks>
  </equipment>
  <equipment>
    <name> Ge 半導体検出器 </name>
    <spec> 相対効率:kV36 </spec>
    <spec> 分解能:1.90keV </spec>
    <model> DSA1000 </model>
    <year> 1995 </year>
    <remarks />
  </equipment>
</equipments>
```

図 6 設備情報更新用 XML フォーマット

旧サイトに比べて保守作業量も少ない．

その後，先の 2 案と本案をまとめ，TKF の最高意志決定機関である首都圏公設試連携推進会議に提案したところ，本案の採用が認められ事業化に至った．現在，TKF Web サイトおよび設備検索エンジンは，顧客と職員を対象とした首都圏公設試の総合カタログとして，継続運営されている．

5 おわりに

本稿では，公設設試験研究機関のポータルサイト上に設置した，設備検索エンジンについて述べた．設備検索エンジンは，各公設試の Web サイトに分散する実験設備の情報を一括検索する機能である．設備情報は，Web クローラによって各 Web サイトから自動収集される．クローラが訪問する URL を限定することで，設備情報以外の情報をできる限り排除する．既存のポータルサイトに見られる横断検索と比較して，設備検索エンジンの方が検索精度の面で優れていることを示した．また，設備検索エンジンの適用範囲について考察し，約 94%の工業系公設試に適用可能であることを示した．

今後の課題として，広域首都圏輸出製品技術支援センター（Metropolitan Technical Support Network for Export Products: MTEP）[9] への対応が挙げられる．MTEP は，広域関東圏の公設試が一体となって，国際規格に関する技術的コンサルティングを行う連携体である．これへの対応として，設備情報を収集する際に，規格に準拠した試験が可能か否かの情報を合わせて収集し，それを付加した形で提供する等の案が考えられる．

謝辞

本稿の執筆にあたって，産業技術総合研究所名誉リサーチャー小島俊雄氏には，様々なご指導をいただきました．また，千葉県産業支援技術研究所城之内一茂氏，神奈川県産業技術センター中谷吉久氏には，連携機関の立場からご助言をいただきました．重ねて御礼申し上げます．

参考文献

[1] 中小企業庁：公設試経営の基本戦略 ～中小企業の技術的支援における公設試のあり方に関する研究会中間報告～，中小企業庁（オンライン），http://www.chusho.meti.go.jp/keiei/gijut/2005/download/051220kousetushi_senryaku_houkokusho.pdf，2011 年 10 月参照．

[2] 全日本地域研究交流協会 : 地域の産学官連携への公設試の効果的な取組みに関する調査研究 —地域イノベーションの加速を目指して—，全日本地域研究交流協会（オンライン），http://www.jarec.or.jp/pdf/cyosa/18-chiikino.pdf，2011 年 11 月参照．

[3] 谷口邦彦 : 公設試験研究機関の役割 : 技術集積機関と中堅・中小企業との橋渡し，研究技術計画，Vol.15, No.3, pp.162–167, 2003.

[4] 阿部真也，北原枢，五十嵐美穂子，山田一徳，近藤幹也，吉野学，片岡正俊 : 設備データベースと新着情報機能を有する公設試広域連携 Web サイトの開発，情報処理学会論文誌：データベース，Vol.6, No.4, pp.58–69, 2013.

[5] 平林幹雄，江渡浩一郎 : N.M-gram: ハッシュ値付き N-gram 索引による全文検索の一手法，情報処理学会論文誌：データベース，Vol.48, No.7, pp.29–37, 2007.

[6] Mikio H. : 全文検索システム Hyper Estraier, FAL Labs（オンライン），http://fallabs.com/hyperestraier/，2014 年 7 月参照．

[7] 片岡正俊：首都圏の連携体「テクノナレッジ・フリーウェイ」，産学官連携ジャーナル，Vol.7, No.10, pp.6–7, 2011.

[8] 埼玉県産業技術総合センター，千葉県産業支援技術研究所，地方独立行政法人東京都立産業技術研究センター，神奈川県産業技術センター，横浜市工業技術支援センター : 首都圏テクノナレッジ・フリーウェイ，地方独立行政法人東京都立産業技術研究センター（オンライン），http://tkm.iri-tokyo.jp/，2014 年 10 月参照．

[9] 茨城県工業技術センター，栃木県産業技術センター，群馬県立産業技術センター，埼玉県産業技術総合センター，千葉県産業支援技術研究所，地方独立行政法人東京都立産業技術研究センター，神奈川県産業技術センター，新潟県工業技術総合研究所，山梨県工業技術センター，長野県工業技術総合センター，静岡県工業技術研究所 : 広域首都圏輸出製品技術支援センター，地方独立行政法人東京都立産業技術研究センター（オンライン），http://www.iri-tokyo.jp/mtep/，2014 年 10 月参照．