# 全文検索を使用した，学術研究情報データベースの構築

SI 事業部 大阪センター 開発部
鈴 木 一 也

## 1. はじめに

産学連携が推進される昨今，保有する研究資源の情報を整備することは，大学にとって重要な課題である。したがって，多くの大学が，研究者のもつ学術研究情報を蓄積し，広く情報を発信することに取り組んでいる。

本稿では，こうした取り組みのひとつである，立命館大学様の「立命館大学研究者学術研究情報データベース」（以下，本システムと記載）を通じて，全文検索による学術研究情報データベースの構築例を，紹介する。

## 2. 目　　的

本システムは，フリーワード検索や条件の設定と選択によって，『ユーザが求める学術研究情報をもつ，研究者を探し出す』機能の提供を，主な目的とする。

また，本システムの構築にあたり，主目的の他に，以下の二次的なシステム目標を設ける。

・情報の一元管理を行う

以前は，管理や公開の目的に応じて，該当情報を担当する部署が，独自にデータを管理することがあった。しかし，同一ソースの情報を複数箇所で保存することは，無駄が多いだけでなく，情報間で不整合が生じる危険性をもっている。

このため，学術研究情報は全て，本システムで管理する。

・研究者自身によって，情報を管理する

学術研究情報は個別性・専門性の高い情報である。それゆえ，専門知識をもたないオペレータによるメンテナンスには，限界が生じることになる。

本システムは，原則として，情報を研究者自身が管理する。

なお，メンテナンスされた情報は，即座に本システムに反映する。

・情報を公開する

蓄積した情報は，広く公開されなければならない。したがって，特定のクライアントを必要とするシステムは，不適切である。

本システムは，Web インターフェイスを装備することによって，インターネットを通じて，情報を広く公開する。

・利用目的に応じた，情報を提供する

条件に一致した研究者を，単純に示すだけでは，利便性の低いシステムといえる。

本システムは，目的に応じて，情報を整理して表示する。

また，他システムとのデータ変換や，データダウンロード機能によって，情報の流用性を高める。

## 3. 機　　能

本システムは，以下の基本機能を提供する。こ

れによって，主目的の実現はもちろん，さらに活用範囲を拡大することを目指す。

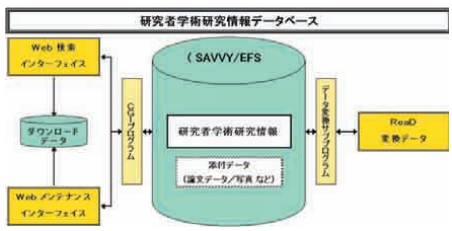
・専用の Web インターフェイスによって，Web ブラウザから，検索を行う機能。（一般公開機能）
・専用の Web インターフェイスによって，Web ブラウザから，研究者自身がデータのメンテナンスを行う機能。（不正使用防止のセキュリティ機構を含む）
・組織別一覧など，予め定められた複数のカテゴライズに従って，動的に研究者の一覧を生成する機能。
・複数の予備項目をもち，研究者毎に表示設定を行うことができる機能。
・管理者によって，表示対象外の管理項目を追加することができる機能。
・全ての管理項目を，全文検索対象として利用することができる機能。
・英文情報（和文情報と連動）の表示及び検索を行うことができる機能。
・「ReaD (Directory Database of Research and Development Activites) 研究開発支援総合ディレクトリ」[1] など，外部機関システムとのデータ連携を行うことができる機能。
・オンラインシラバスシステムと連携することができる機能。

## 4. 構　　成

本章では，本システムの構成と，そこに至った経緯について考察する。

### 4.1 構成の概要

本システムは，三つのパートで構成される。




SAVVY/EFS という，全文検索エンジンを使った文書管理システムを，データベースの核とする。これに，Ｗｅｂインターフェイスやデータ変換のためのサブプログラムを加えて，システムを構成する。

### 42 SAVVY/EFS の概要

SAVVY/EFS[2] は，全文検索エンジン[3] を使用した文書管理ソフトウエアである。

文書管理には，キャビネット方式と呼ばれる，階層構造を採用している。オフィスで使われるキャビネットを仮想的に実現したもので，【キャビネット → ドロワ（引き出し）→ フォルダ → ドキュメント】という階層をもつ。

SAVVY/EFS の特長は，柔軟な検索機能である。管理文書の全文を，自由な言葉で高速に検索する，高速全文検索機能。パターン認識技術による曖昧検索と，従来の全文検索エンジンが苦手とした完全一致検索を，同時に実行する，ハイブリッド検索機能。ドキュメント毎に設定された，

[1] 国内の国立研究機関，独立行政法人，特殊法人，地方公設試，国公私立大学の附属研究施設，公益法人等に関する研究機関情報，研究者情報，研究課題情報，研究資源情報等を中心とした研究情報を提供し，研究開発活動を情報面から支援することを目的に，科学技術振興事業団（JST）が運営するサービス。

[2] SAVVY/EFS は，ジップインフォブリッジ株式会社が発売するソフトウェアである。

[3] 同社は，SAVVY/TRS という超高速全文検索エンジンを発売している。

数値や日付などの属性を検索する，属性項目検索機能。これら基本的な検索機能に加えて，階層名につけたラベルを使用して検索範囲を限定することも可能となっている。

### 43 構成の考察

何故全文検索なのか，何故 SAVVY/EFS なのか。

ユーザが本システムを利用するのは，例えば，『全文検索に関する知識を持った，研究者を探したい』といった場合であろう。つまり，全文検索というキーワードを含む，学術研究情報を探しだす作業である。それは，専門分野や研究課題として発見できるかもしれない。あるいは，自己紹介の中に登場するかもしれない。何処にあるか特定できないが，何処かに含んでいる情報を，探す必要がある。

非常に単純な構成とするならば，研究者一人の情報を一つの文書ファイルに雑多に詰め込み，全研究者の文書ファイルを，全文検索エンジンを使って一気に検索するといった手法を考えることができる。

これで，『ユーザが求める，学術研究情報をもつ研究者を，探し出す』という目的は達成できる。ユーザの指定した条件を，全文書から検索して，文書を特定すれば，すなわち研究者を特定できたことになるのだから。

しかし，柔軟な情報提供[4]と情報管理を行うためには，不十分である。

先の例が，『全文検索に関する知識を持った，理工学部に所属する研究者を探したい』という条件に変化すると，この手法では対応できなくなる。理工学部というキーワードを追加して検索しても，それは任意の場所に含まれる可能性があるため，不正確である。この場合は，所属学部という項目が理工学部でなければいけない。

つまり，情報を整理し，項目毎に管理する必要があるということだ。項目管理を行えば，飛躍的に柔軟な検索が可能になる。

項目を指定して条件を一致させることができるようになる。必要な範囲を検索し，必要な情報だけを提示することができるようになる。また，見せる情報と見せない情報など，個別に管理できるようになる。

情報を項目毎に管理するという面では，リレーショナルデータベース（以下，RDB と記載）によるデータ管理が，適しているように思えるかもしれない。しかし，本システムの基本は，全情報からの検索である。

このような検索を RDB で実現することは，非現実的だ。なぜなら，全テーブルの全フィールドを個別に検索し，OR 演算することになるからである。

また，性能面でも RDB は適当ではない。本システムが扱う学術研究情報は，その殆どが不定長の文字列情報であり，その中間一致検索を行うことになる。研究者は，千名を越える。膨大な情報量になることが予測される。つまり，リレーショナルデータベースが最も苦手とする検索形態なのである。

こうした場合，RDB による情報管理と，全文検

[4] 全文検索エンジンは，文書を管理する仕組みを持たない。従って，全文検索エンジンしか持たないシステムでは，特定した文書をそのまま提示する程度の処理しかできない。

索による情報検索を組み合わせる手法が，最も有効である。しかし，この手法はコスト面で不利である。管理する情報が変化した際には，データ構造の定義を変更する必要が生じる。初期コストだけでなく，継続的なメンテナンスコストが発生する可能性があるのだ。

そこで，本システムでは SAVVY/EFS を採用した。SAVVY/EFS は文書管理システムであり，RDB のように，個別の項目情報を扱うことを目的としていない。

しかし，本システムでは，文書（ドキュメント）の扱いを工夫することによって，個別項目の管理を実現し，また，低コストで柔軟な情報管理をも実現している。

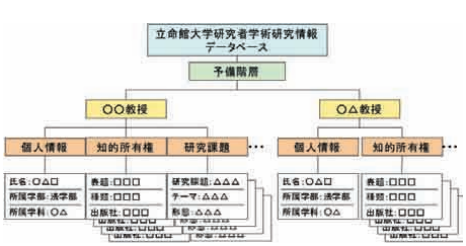
## 5. 特　　徴

本章では，本システムの特徴的な点について，概略を解説する。

### 51 個別項目管理を実現する工夫

SAVVY/EFS は，階層名につけたラベルを使用して，検索範囲を限定することができる。また，ドキュメントの検索対象ページを指定することもできる。この機能を利用して，RDB でいうところのテーブル，レコード，フィールドを，階層構造上にマッピングする。

具体的には，フォルダをテーブル，ドキュメントをレコード，ページをフィールド，として扱う。さらに，研究者毎にドロワを割り当てる。RDB で例えると，研究者毎にデータベースを分割するイメージである。




1 フィールドは，ドキュメント内の 1 ページであるため，不定長・不定形式のデータを収めることができる。もちろん，テーブル内のフィールド数も不定である。つまり，管理する情報が変化してもデータ構造の定義を変更する必要がない[5]。

この構造は，複数の言語に対応する上でも有利である。

各フォルダを，二組づつ用意する。データ構成は同様とし，一方は日本語，一方は英語を管理する。対応言語が増加した際には，フォルダを三組，四組と増やしていけばよい。研究者毎にドロワを分けているため，このような手法が比較的簡単に実現できる。

### 52 検索結果の分類

本システムと同様な目的をもつシステムは通常，検索結果を順に出力するか，研究者をキーとして結果を出力するものが多いと思われる。

下図のような状態である。本システムも，フリーワードによる簡易検索では，この表現を採用している。

しかし，本システムの，項目条件を設定できる詳細検索では，任意な条件で検索した結果を，ユーザが指定した項目をキーとして分類した結果を，出力することができる。

5 情報の形式や数が変わるため，インターフェイスプログラムやデータコンバートプログラムの変更は，必要になることがある。

| 氏名 | 氏名(カナ) | 所属 | 職名 | 専門分野 | 基本情報 |
|---|---|---|---|---|---|
| 織田/信長 | オダ/ノブナガ | 法学部<br>法学研究科 | 教授 | 政治学<br>日本史 | 表示 |
| 徳川/家康 | トクガワ/イエヤス | 経営学部<br>経営学研究科 | 教授 | 経営学 | 表示 |
| 豊臣/秀吉 | トヨトミ/ヒデヨシ | 文学部<br>心理学科心理学専攻 | 教授 | 心理学 | 表示 |
| 前越/浦三 | サイトウ/ドウサン | 経済学部<br>経済学研究科 | 教授 | 財政学<br>金融論 | 表示 |
| 武田/信玄 | タケダ/シンゲン | 法学部<br>法学研究科 | 教授 | 政治学 | 表示 |
| 浅井/長政 | アサイ/ナガマサ | 文学部<br>史学科日本史学専攻 | 教授 | 日本中世史 | 表示 |
| 上杉/謙信 | ウエスギ/ケンシン | 文学部<br>史学科日本史学専攻 | 教授 | 日本中世史 | 表示 |
| 毛利/元就 | モウリ/モトナリ | 文学部<br>史学科東洋史学専攻 | 教授 | 東洋史 | 表示 |
| 豊臣/秀長 | トヨトミ/ヒデナガ | 経営学部<br>経営学研究科 | 教授 | 会計学 | 表示 |
| 伊達/正宗 | ダテ/マサムネ | 法学部<br>法学研究科 | 教授 | 行政学 | 表示 |

下図は，同じ条件で詳細検索を実施した結果である。

00001 | 法学部

| 氏名 | 氏名(カナ) | 所属 | 職名 | 専門分野 | 基本情報 |
|---|---|---|---|---|---|
| 織田/信長 | オダ/ノブナガ | 法学部<br>法学研究科 | 教授 | 政治学<br>日本史 | 表示 |
| 武田/信玄 | タケダ/シンゲン | 法学部<br>法学研究科 | 教授 | 政治学 | 表示 |
| 伊達/正宗 | ダテ/マサムネ | 法学部<br>法学研究科 | 教授 | 行政学 | 表示 |

00002 | 経営学部

| 氏名 | 氏名(カナ) | 所属 | 職名 | 専門分野 | 基本情報 |
|---|---|---|---|---|---|
| 徳川/家康 | トクガワ/イエヤス | 経営学部<br>経営学研究科 | 教授 | 経営学 | 表示 |
| 豊臣/秀長 | トヨトミ/ヒデナガ | 経営学部<br>経営学研究科 | 教授 | 会計学 | 表示 |

00003 | 文学部

| 氏名 | 氏名(カナ) | 所属 | 職名 | 専門分野 | 基本情報 |
|---|---|---|---|---|---|
| 豊臣/秀吉 | トヨトミ/ヒデヨシ | 文学部<br>心理学科心理学専攻 | 教授 | 心理学 | 表示 |
| 浅井/長政 | アサイ/ナガマサ | 文学部<br>史学科日本史学専攻 | 教授 | 日本中世史 | 表示 |
| 上杉/謙信 | ウエスギ/ケンシン | 文学部<br>史学科日本史学専攻 | 教授 | 日本中世史 | 表示 |
| 毛利/元就 | モウリ/モトナリ | 文学部<br>史学科東洋史学専攻 | 教授 | 東洋史 | 表示 |

00004 | 経済学部

| 氏名 | 氏名(カナ) | 所属 | 職名 | 専門分野 | 基本情報 |
|---|---|---|---|---|---|
| 前越/浦三 | サイトウ/ドウサン | 経済学部<br>経済学研究科 | 教授 | 財政学<br>金融論 | 表示 |

この機能は，一度検索した結果から，分類キー項目の一覧を作成し，キー毎に検索した結果と，先の検索結果の AND 演算によって実現している。

## 6. おわりに

インターネット上には，無数のホームページが存在する。情報が氾濫している。ホームページを検索して，何らかの情報を得ようと試みたことのある読者であれば，情報を取捨選択することの重要性と難しさを，ご理解いただけることと思う。

本稿では，『膨大な情報の中から，目的の情報を手に入れる手助けをするシステム』を構築する過程の，一例を紹介した。紙面の都合上，表面的な解説にとどまったが，本稿が，同様なシステムを考察する際の一助となれば幸いである。