# Ⅰ－35 情報の相互運用性を高めるためのコア用語の選定手法に関する研究

## Research on Method to Select Core Terms for Improving Interoperability of Various Information

北野光一[1]・中村健二[2]・田中成典[3]・古田均[3]

Koichi KITANO, Kenji NAKAMURA, Shigenori TANAKA and Hitoshi FURUTA

**抄録：**近年，建設分野において様々な情報システムが構築され，様々な形式の情報が流通している．これらの情報を統合的に扱うためのデータフォーマットとして XML が用いられている．しかし，新たな XML スキーマを作成する際に，既存スキーマとの統合や相互運用を円滑に行うためにはどのような語を要素名として選択すべきかの判断基準が存在しない．そこで，我々は，建設分野において広く一般的に利活用可能な用語セットを建設分野のコア用語と定義し，それらのコア用語を選定する研究を行ってきた．本論文では，土木用語大辞典から見出し語間ネットワークを構築して各見出し語の特徴を算出し，機械的にコア用語を選定する手法を提案する．

**Abstract:** Recently, various information systems have been built up and exchanged various types of data in construction field. The XML is used as a data format to deal with various type of data integratively. However, there is no criterion to select word as element name for smooth data integration and interoperability in creating new XML schemas. Thus, we define the set of terms as core terms that should be used for general purposes in construction field and study on method for selecting core terms. In this paper, we propose the method for selecting core terms mechanically by creating network of entry words based on Dictionary of Civil Engineering Terminology and calculating characteristics of entry words.

**キーワード：**言語処理，スキーマ統合，コア用語，見出し語間ネットワーク
***Keywords*** **：*Language Processing，Schema Integration，Core Term，Network of Entry Words***

## １．はじめに

近年，情報化の進展に伴い，建設分野において様々な情報システムが構築され，多種多様な情報が流通している．しかし，これらの情報システムで利用されるデータは，それぞれ個別のシステム毎に特化しており，他の情報システムとの相互運用は考慮されていない．そのため，データが一元的に管理されておらず，同じ対象のデータを複数のシステム間で利用する場合，システム毎に情報を管理する必要があり，無駄なコストが発生している．このような状況を鑑み，CALS/EC では，アクションプログラム 2005[1)]の目標 13 や目標 16 において，システム間の情報連携や工事施工中の情報交換・共有を目標として掲げ，情報共有基盤の構築を目指している．この目標は，アクションプログラム 2008[2)]の目標 1，目標 2 や目標 4 に引き継がれており，情報連携基盤を前提とした指針が掲げられている．これらの情報連携の一環として，特定のシステムに依存しないデータ形式である XML を用いたデータの流通が行われている．XML は，データの意味や構造を任意に設計可能なメタ言語であり，データを記録するスキーマの設計が容易である．また，XML で記述されたデータは，同じ文法で機械的に解釈が可能であるため，様々なシステム上で利用可能である．しかし，異なるスキーマ間で，データの意味やスキーマの構造が異なると，機械による意味的な解釈が困難となり，情報の相互運用に問題が発生する．そのため，情報システムで用いられるデータのスキーマは，他のスキーマとの高い相互運用性を確保した設計を採用する必要がある．

スキーマの相互運用性を確保する方法は，「スキーマ同士の類似部分を利用する方法」，「スキーマに従って蓄積されたデータを利用する方法」と，「スキーマの作成時に相互運用性の高い語を選定する方法」の 3 つがある．1 つ目と 2 つ目の方法は，既存のスキーマを対象にしてデータ連携を行う方法である．また，3 つ目の方法は，新たにスキーマを定義する際の方法である．既存の取り組みとして，1 つ目の方法は，スキーマの木構造に着目し，他のスキーマとの共通部分木を

１：学生会員　修士(情報学)　関西大学大学院　総合情報学研究科
(〒569-1095　大阪府高槻市霊仙寺町二丁目一番一号, Tel :072-690-2154, E-mail : noahany@gmail.com)
２：正会員　博士(情報学)　関西大学　ポスト･ドクトラル･フェロー　総合情報学部
(現，立命館大学　助手　情報理工学部)
３：正会員　工学博士　関西大学　教授　総合情報学科（〒569-1095　大阪府高槻市霊仙寺町二丁目一番一号）

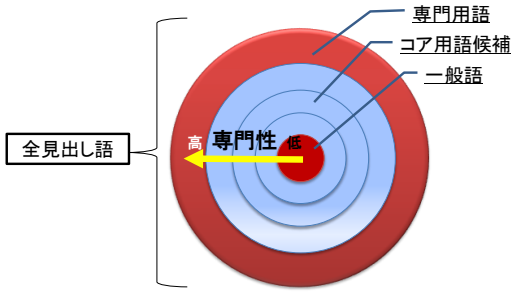
**表－１ 各語の定義**

| 語の種類 | 定義 |
|---|---|
| 専門用語 | 定義 1：専門性が高い用語<br>定義 2：複数のコア用語もしくは複数の語からなる複合語であり概念が明確な用語<br>定義 3：利用用途が専門分野に特化しており利用頻度が低い用語 |
| 一般語 | 定義 4：一般性が高い単語<br>定義 5：コア用語や専門用語などの複数の語において頻繁に利用されている単語<br>定義 6：1 つの概念をあらわす語であるため，形態素数が少ない語 |
| コア用語 | 定義 7：建設分野において一般的に利用可能な用語<br>定義 8：専門性が高い専門用語ではなく，やや専門性を帯びた一般性が高い用語<br>定義 9：他の分野と比較して建設分野での利用頻度が高い用語 |

探索し，共通部分の要素同士を対応付ける手法[3)-6)]が研究されている．2 つ目の方法は，スキーマ間で，決定木を用いて，類似したデータを持つ要素対を特定する手法[7),8)]が研究されている．3 つ目の方法は，設計者が既存の関連するスキーマを収集し，そのスキーマで利用されている語や構造を参考にして，独自に定義している状況である．

これらの研究状況を整理すると，1 つ目と 2 つ目の方法により，既存のスキーマのデータの相互運用性に関する問題については，活発に取り組まれている状況であり，一定の成果が得られている状況である．しかし，これらの方法は，スキーマを実際に運用した後に，必要となった時点で相互運用性を確保するという対応であり，新たな分野や語が増加するたびにルールを更新する必要があり，今後の情報共有を行うための技術基盤としては課題が山積している状況である．そのため，3 つ目の手法について，設計者の能力に依存することなく高い相互運用性を持つスキーマを作成するための仕組みが，情報共有基盤の確立に必要となる．

そこで，著者らは，（財）日本建設情報総合センターの社会基盤情報標準化委員会情報連携基盤小委員会建設 XML 検討 WG と一部共同で，相互運用性の高い XML スキーマの作成を支援するための研究を行ってきた．相互運用性の高いスキーマを作成するためには，「1)命名規則に従ったスキーマ要素名の決定」と，「2)相互運用性の高い語のスキーマ要素への割り当て」の２つの尺度を考慮する必要がある．1）については，既に一定の結論を得ており，その成果を建設分野における XML 記述仕様の考え方（案）[9)]としてまとめ，公開している．そのため，本論文では，2）についての研究成果をまとめ，得られた知見について述べる．ここで，建設分野において，相互運用性の高い用語とは，建設分野で既に広く利用されている用語であり特定の業種や工種に特化した専門用語ではないという特徴がある．専門用語は，建設分野内で特定の業種や工種に特化した概念であり，異業種間での共通概念として認識できず，情報の相互運用性が損なわれる．本論文では，これらの特徴を持つ用語をコア用語として定義し，コア用語を機械的に選定する手法について論じる．




**図－１ コア用語のイメージ**

## ２．建設分野におけるコア用語の定義

本研究では，コア用語という表現の意味を明確にするため，「語」，「単語」，「用語」，「複合語」，「見出し語」と「コア用語」の 6 つの語について，それぞれ次に示す概念を表現する際に利用する．語は，単語もしくは用語を区別なく総称する際に用いる．単語は，使用する分野を限定せず，一般的な概念を表現する際に用いる．用語は，使用する分野を限定し，専門性を帯びた概念を表現する際に用いる．複合語は，複数の語から構成された語を表現する際に用いる．見出し語は，土木用語大辞典に掲載された語もしくは複合語を表現する際に用いる．ここで，土木用語大辞典の見出し語には，他の見出し語が含まれていることから，本論文では複合語も含むものとする．コア用語は，用語の中から一般語と専門用語を除いた特定の分野で広く使用される用語を表現する際に用いる．本研究は，土木用語大辞典[10)]から，建設分野において広く利用される用語セットをコア用語として選出することを目的とする．著者らが考えるコア用語のイメージを**図－１**に示す．著者らは，コア用語を専門用語と一般語の間に存在する用語群であると考えている．本論文における専門用語，一般語とコア用語の定義を**表－１**に示す．専門用語とは，専門性が高く，特定の分野内の狭い範囲内でのみ利用可能な用語である．一般語とは，一般性が高く，その語のみでは概念を具体化できない語や，様々な分野で使用される単語である．コア用語は，一般語に比べて特定の分野でのみ使用される用語であり，専門用語に比べて，特定の分野内で広く使用される用

**表－２ 分析に用いる指標**

| 指標名 | 説明 | 関連する定義 |
|---|---|---|
| 形態素数 | 見出し語を構成する形態素の数 | 定義 2，定義 6 |
| 上位概念数 | 見出し語間ネットワーク中の見出し語の直接の上位概念の数 | 定義 4 |
| 下位概念数 | 見出し語間ネットワーク中の見出し語の直接の下位概念の数 | 定義 1 |
| 説明文長さ | 辞書に掲載されている見出し語の説明文の長さ | 定義 2 |
| 検索結果数 | 見出し語を検索エンジン Google[11)]で検索した結果の件数 | 定義 5 |
| 見出し語間ネットワーク階層数 | 見出し語の最上位概念からの距離＋最下位概念からの距離 | － |

語である．また，コア用語は，複数階層で構成されており，一般性の高いコア用語と，専門性の高いコア用語が存在する．

## ３．土木用語大辞典収録見出し語の分析

本研究では，土木用語大辞典に収録されている全見出し語から，建設分野のコア用語を選出する．コア用語は，第 2 章で定義したように，専門用語と一般語の間に存在する用語である．そのため，本研究では，全見出し語から専門用語と一般語を特定して削除し，一般語と専門用語に分類されない用語を選定することによりコア用語を抽出する．本章では，全見出し語の特性を複数の指標に基づいて分析し，一般語と専門用語の統計的な特性を取得する．そして，取得した統計的な特性をもとに，一般語と専門用語を抽出するための手法を提案する．

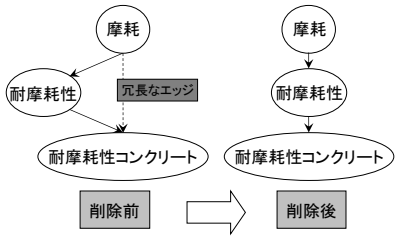
### （１）分析に用いるデータ

本分析では，土木用語大辞典の情報を用いる．土木用語大辞典は，土木・建設分野に関わる語を一般語から専門用語まで幅広く収録しており，一般語と専門用語の中間に位置するコア用語を多く収録していると考えられる．本実験で使用したデータは，「見出し語」と「見出し語の英語名」，「説明文の文字数」で，これらの情報を人手でデータベースに入力した．

### （２）分析手法

見出し語の特性の分析指標として，形態素数，上位概念数，下位概念数，説明文長さ，検索結果数と見出し語間ネットワーク階層数の 6 つを用いる．分析に用いる指標を**表－２**に示す．**表－２**の関連する定義では，本分析で用いる指標名，説明およびその指標が，**表－１**で示す各語のどの定義と関連するかを示している．

見出し語間ネットワークとは，見出し語をノード，ノード間の関係をエッジとする有向グラフを指す．本分析処理では，まず，見出し語 A が見出し語 B の部分文字列である時，A を B の上位概念，B を A の下位概念と呼び，A から B へのエッジを作成する．ここで，「摩耗」という見出し語を上位概念とすると，「耐摩耗性」という見出し語がより専門性の高い下位概念となる．また，本研究では，コア用語の選定指標として，




**図－２　冗長なエッジの例**

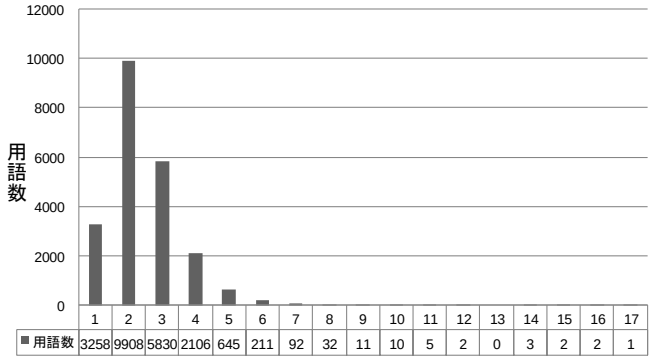
見出し語間ネットワークにおけるネットワークの階層数及び各見出し語が位置する階層の深さを使用する．そのため，冗長なエッジを削除し，見出し語のノードの正しい深さを算出する．冗長なエッジの例を**図－２**に示す. これは, 冗長なエッジの削除前と削除後のノード同士の関係を表している．まず，削除前には，「磨耗」のノードから，「耐磨耗性」と「耐磨耗性コンクリート」の 2 つのノードに対するエッジと，「耐磨耗性」のノードから「耐磨耗性コンクリート」のノードに対するエッジが存在する．削除後には，「耐摩耗性コンクリート」から「摩耗」へのノードが削除され，「摩耗」と「耐摩耗性」間，「耐摩耗性」と「耐磨耗性コンクリート」間のみにエッジが存在する見出し語間ネットワークが構築される．

### （３）分析結果と考察

#### a）形態素数

土木用語大辞典の見出し語の形態素数の集計結果を**図－３**に示す．図から，全体の約 90%の見出し語が，形態素数 4 以下で構成されていることが分かった．形態素数に基づく分類では，次に示す 4 つの特徴がみられた．

- 形態素数が 1 の見出し語は,「橋」,「力」,「港」,「坂」などの一般的な単語と，「橋桁」や「交角」などのややコア用語に近い専門性を帯びた用語の割合が高い．
- 形態素数 2～4 までの見出し語は，「防水コンクリート」，「鉄筋コンクリート」や「無筋コンクリート管」など，一般的な用語と用途を表す語を組み合わせた複合語が多く，コア用語に近い用語の割合が高い．
- 形態素数 5～8 の見出し語は,「移動式大型架設桁」,「空港構造物設計荷重」，「半潜水型石油掘削リグ」や「軽荷重スラブ橋用プレストレスト橋桁」




**図－３ 形態素数の集計結果**

などの用語が多く専門性の高い用語の割合が高い.

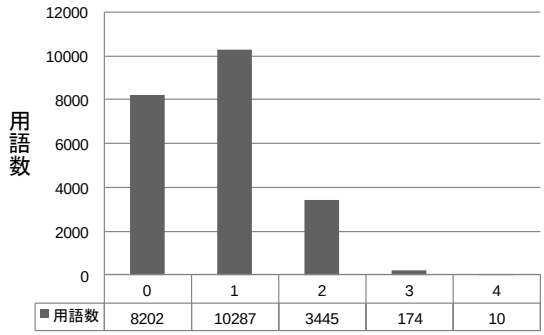
- 形態素数が 9 以上の見出し語では，「地方拠点都市地域の整備及び産業業務施設の再配置の促進に関する法律」や「産業廃棄物の処理に係る特定施設の整備の促進に関する法律」などの法律名が含まれ，コア用語の抽出処理においてはノイズとなるような用語の割合が高い.

これらの特徴をまとめると，形態素数と用語の専門性はある程度の相関関係があると考えられる．形態素数が少ない語は，語に含まれる概念数が少なく，一般語である可能性が高い. また, 形態素数が多い用語は，用語に含まれる概念数が多く，専門性が高い専門用語である可能性が高い．そのため，形態素数が多い見出し語を専門用語として，形態素数が少ない見出し語を一般語として，ある程度抽出できると考えられる．しかし, 形態素数を用いて一般語を決定すると,「橋桁」などのコア用語に近い用語も除外され，コア用語を正しく抽出できないことが分かった．また，形態素数 9 以上の見出し語には，法律や指針などの名称が多く含まれており, ノイズとなるような用語が多くみられた.

**b）上位概念数**

土木用語大辞典の見出し語の上位概念数の集計結果を**図－４**に示す．図から，上位概念数は，最大で 4 であることが分かった．上位概念数に基づく分類では，次に示す 3 つの特徴がみられた.

- 上位概念数が 0 の見出し語には，「橋」，「駅」や「鋼」などの他の語の一部として広く使用される単語と，「NTT 無利子貸付金制度」，「気候変動に関する政府間パネル」や「共同溝の整備等に関する特別措置法」などの，他の語と関連を持たない用語が存在することが分かった.



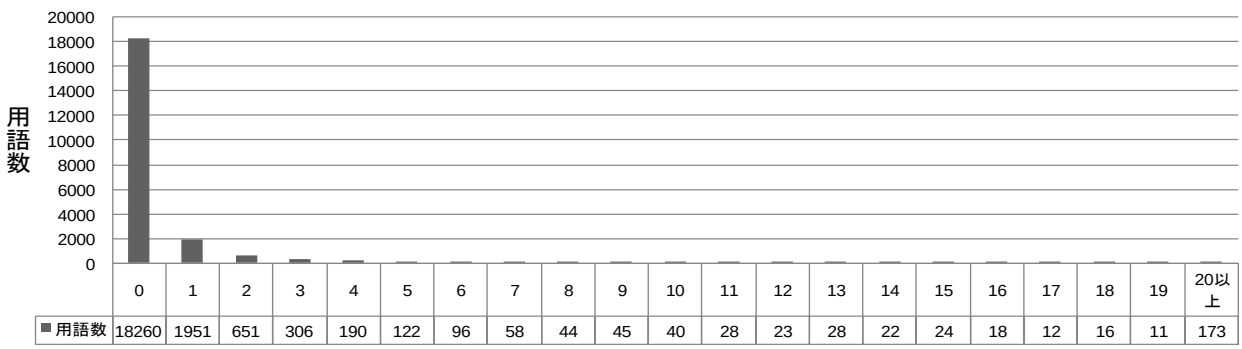

**図－４ 上位概念数の集計結果**

- 上位概念数 1～2 の見出し語には,「橋脚」,「ボックスカルバート」や「鉄筋コンクリート」などコア用語候補となる用語が多く存在することが分かった.
- 上位概念数 3 以上の見出し語には,「遠心力鉄筋コンクリートポール」,「ポストテンション方式遠心力プレストレストコンクリート杭」や「大都市地域における宅地開発及び鉄道整備の一体的推進に関する特別措置法」などの専門用語や法律名などのノイズとなる用語が含まれていることが分かった.

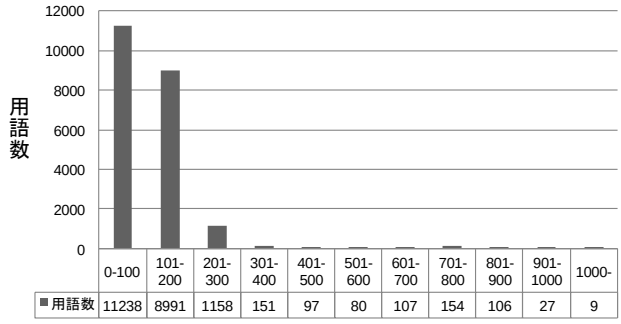
これらの特徴をまとめると，上位概念数が 0 の見出し語には単語やノイズとなる用語が多く含まれていることが分かる．また，上位概念数 3 以上の見出し語には，専門用語やノイズとなる用語が含まれていることが分かる．これは，上位概念数が語に含まれる概念の数を示しており，含まれる概念数が増加するにつれて専門性が高まるためであると考えられる．この結果から，上位概念数 0 および上位概念数 3 以上の見出し語をノイズとなる用語として除去する処理を行うことで, それぞれ，一般語および専門用語の候補となることが分かる.

**c）下位概念数**

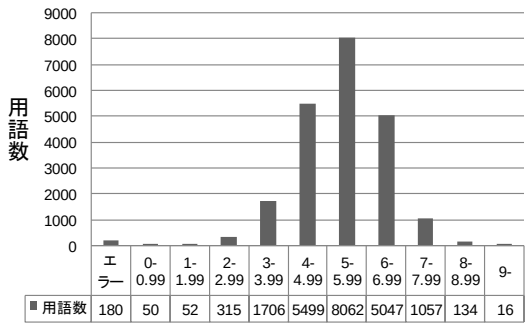
土木用語大辞典の見出し語の下位概念数の集計結果を**図－５**に示す．図から，全体の約 90%の見出し語が




**図－５ 下位概念数の集計結果**




**図－６ 説明文長さの集計結果**




**図－７ 検索結果件数の対数の集計結果**

下位概念数 1 以下で構成されていることが分かった. また, 下位概念数 0 の見出し語が 18,260 件あり, 全体の約 82.5%の見出し語に下位概念が存在しないことが分かった. 下位概念数に基づく分類では, 次に示す 3 つの特徴がみられた.

- 下位概念数 0 の見出し語には,「ボックスカルバート」,「アーチカルバート」や「鉄筋コンクリート」などコア用語候補となる用語や,「スラリー循環型高速凝集沈殿池」や「ソイルセメント合成鋼管杭工法」などの専門性の高い用語が多く含まれていることが分かった.
- 下位概念数 1～19 の見出し語には,「媒体」,「土性」や「無筋コンクリート」など, 一般語やコア用語候補となる用語が混在して存在することが分かった.
- 下位概念数が 20 以上の見出し語には,「渦」,「データ」や「力」など, 単語が多く含まれていることが分かった.

これらの結果から, 下位概念数の値のみを利用して専門用語や一般語を抽出することは困難であることが分かった.

**d）説明文長さ**

土木用語大辞典の見出し語の説明文の長さによる集計結果を**図－６**に示す. 図から, 全体の約 90%の見出し語が, 200 文字以下であることが分かった. また, 土木用語大辞典の編集方針を確認すると, 説明文の長さに応じて見出し語の重要度を設定しており, 800 文字程度が特Aとなり, 土木分野で「幹」となる極めて重要な概念であるとしている. そして, 残りの見出し語について, 200 字程度がA, 100 字程度がB, 指示文のみがCと分類している. この編集方針を考慮して, 説明文の長さに基づく分類を確認した結果, 次の 5 つの特徴がみられた.

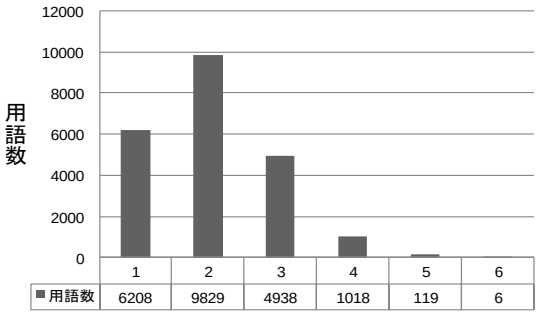
- 説明文の長さが 1～100 で, 同義語のみを示している見出し語を確認したところ,「コンター＝等高線」や「自動車駐車場＝駐車場」など, 語句の言い換えや正式名称などを示していることがわかった.
- 説明文の長さが 1～100 で, 参考語のみを示している見出し語を確認したところ,「取水量→取水」や「無作為抽出法→標本抽出法」など, 語句を理解するための引用や, 上位概念を紹介するための引用などがあり, 参考語の意味が統一されていない状況であった.
- 説明文の長さが 1～100 で, 解説文が正しく記載されている見出し語を確認したところ,「アーチカルバート」,「ボックスカルバート」や「コンクリート管」などのコア用語候補となる用語が存在することが分かった.
- 説明文の長さが 100～200 の見出し語の中では,「桁橋」や「カルバート」などが存在し, 重要な用語であるが一般語に近い用語などが存在することが分かった.
- 説明文の長さが 200 以上の見出し語の中には,「ランドマーク」,「浜」,「標高」,「土」や「大規模地震対策特別措置法」などが含まれており, 単語や専門性の高い用語が混在することが分かった.

これらの結果から, 説明文の長さに応じて一般語や専門用語を分類することは困難であり, 説明文の長さを指標として利用できないことが分かった.

**e）検索結果数**

土木用語大辞典の見出し語を Google[11] で検索し, その検索結果件数の対数により集計した結果を**図－７**に示す. 図から, 検索結果件数に基づく分類では, 次に示す 3 つの特徴がみられた.

- 検索結果件数の対数が 0～3.99 の見出し語には,「最小コンプリメンタリーエネルギーの原理」,「鋼繊維補強コンクリート舗装」や「アップデーティッドラグラジアンアプローチ」など専門性の高い用語が多く含まれていることが分かった.
- 検索結果件数の対数が 4～6.99 の見出し語には,「締固め曲線」,「アーチカルバート」や「リスクマネジメント」などコア用語候補となる用語が含まれていることが分かった. その一方で,「辻村太郎」などの固有名詞も含まれており, コア用語候補選出において, ノイズとなる語があることが分かった.
- 検索結果件数の対数が 7 以上の見出し語には,「都




**図－８ 見出し語間ネットワーク階層数の集計結果**

心」，「公道」や「システムエンジニア」などの一般語が含まれていることが分かった．特に，Google の検索結果という点で，一般的な文書に多く含まれる単語が分類されていることが分かった．

これらの結果から，検索結果件数が多い見出し語は，一般的な文書に利用される単語であり，共通で利用される一般語である可能性が高いことが分かった．一方，検索結果件数の少ない見出し語は，限定された範囲で利用される用語であり，利用頻度が少ないことから専門用語である可能性が高いことが分かった．

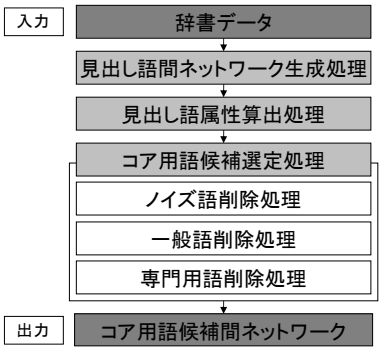
**f）見出し語間ネットワーク階層数**

土木用語大辞典に含まれる見出し語の見出し語間ネットワーク階層数による集計結果を**図－８**に示す．図から，土木用語大辞典には，他の見出し語と関連を持たないネットワーク階層数 1 の語が 6,208 件存在することが分かった．見出し語間ネットワーク階層数に基づく分類では，次に示す 2 つの特徴がみられた．

- 見出し語間ネットワーク階層数 1 の見出し語には，「福祉のまちづくり」，「特定商業集積の整備の促進に関する特別措置法」や「中小企業新技術体化投資促進税制」などのノイズと認識される用語，「大工」や「入江」などの一般語が存在し，一様に分類することが困難であることが分かった．
- 見出し語間ネットワーク階層数が 2～6 の見出し語には，コア用語，専門用語，単語が混在しており，この指標のみで判別することが困難であることが分かった．

これらの結果から，見出し語間ネットワークの階層数のみで分類することが困難であることが分かった．しかしながら，本指標と上位概念数および下位概念数の指標を組み合わせることで，次の 2 つの特徴がみられることが分かった．

- 見出し語間ネットワーク階層数が 2 以上で上位概念数が 0 の見出し語には，「変形」，「地域」，「車両」，「力」や「土」などの単語が多く含まれることが分かった．これは，ネットワークの頂点は一般語に近いことを示していると考えられる．
- 見出し語間ネットワーク階層数が 4 以上で下位概念数が 0 の見出し語には，「流域総合治水対策協議会」，「インターロッキングブロック舗装」や「波力発電式防波堤」などの専門的な用語が多く含まれることが分かった．これは，ネットワークの末端は専門用語になりやすいことを示していると考えられる．




**図－９ コア用語候補選定の流れ**

これらの結果から，見出し語間ネットワーク階層数が深く，そのネットワーク階層の頂点および末端は，それぞれ一般語および専門用語となることが分かる．

## ４．コア用語候補の選定手法

本章では，土木用語大辞典の見出し語の分析結果に基づき，コア用語候補を選定する手法について述べる．本研究では，土木用語大辞典の見出し語の中から，ノイズ語，一般語と専門用語を削除することで，コア用語候補を選定する．コア用語候補選定の流れを**図－９**に示す．

まず，見出し語間ネットワーク生成処理にて，見出し語を構成する部分文字列に着目し，見出し語間ネットワークを生成する．次に，見出し語属性算出処理にて，各見出し語について 6 つの分析指標を求め，コア用語候補選定処理で用いるための見出し語の特徴を取得する．そして，コア用語候補選定処理にて，見出し語間ネットワークからノイズ語，一般語と専門用語を削除し，コア用語の候補を選定する．

著者らは，土木用語大辞典の解析結果に基づき，ノイズ語，一般語と専門用語の条件を**表－３**のように定義した．本研究では，**表－３**に示す通り，固有名詞や法律に関する見出し語を一般語，コア用語や専門用語に分類できない見出し語としてノイズ語と定義した．ノイズ語の傾向を確認すると，助詞を含んだ見出し語や形態素数が 9 以上の見出し語は，公式名，法律名やイベントに関する用語などが多く含まれていた．また，見出し語の英語名の 1 文字目が大文字で始まる語は，人名や地名などの固有名詞が多く含まれていた．一般語は，**表－１**の定義 4～6 の内容に基づき，一般性が高く，下位概念が多い見出し語を抽出した．抽出条件としては，前章の分析結果に基づき一般性が高いと判断される上位概念や検索結果件数の多さなどを指標と

**表－３ ノイズ語，一般語と専門用語の条件**

| 用語種類 | 条件 | 説明 | 件数 |
|---|---|---|---|
| ノイズ語 | 1 | 形態素数が 9 以上の見出し語 | 36 |
| | 2 | 固有名詞 | 1,623 |
| | 3 | 法律名と助詞を含んだ見出し語 | 696 |
| 一般語 | 4 | 見出し語間ネットワーク階層数が 2 以上で上位概念数が 0 の見出し語 | 1,994 |
| | 5 | 下位概念数が 20 以上の見出し語 | 173 |
| | 6 | 検索結果件数の対数の値が 7 以上の見出し語 | 1,207 |
| | 7 | 形態素数が 1，検索結果件数の対数の値が 6 以上，上位概念数が 0 の見出し語 | 1,120 |
| 専門用語 | 8 | 形態素数が 5～8 の見出し語 | 980 |
| | 9 | 形態素数が 4，上位概念数が 2，下位概念数が 0 の見出し語 | 528 |
| | 10 | 上位概念数が 3 以上の見出し語 | 184 |
| | 11 | 見出し語間ネットワーク階層数が 4 以上で下位概念数が 0 の見出し語 | 465 |
| | 12 | 検索結果件数の対数の値が 0～3.99 の見出し語 | 2,123 |
| | 13 | 見出し語間ネットワーク階層数が 3 で下位概念数が 0，形態素 4 の見出し語 | 681 |

して用いた．専門用語は，**表－１**の定義 7～9 の内容に基づき，専門性が高く，利用場面が限定的な見出し語を抽出した． 抽出条件としては，前章の分析結果に基づき専門性が高いと判断される下位概念や検索結果件数の少なさなどを指標として用いた．

## ５．実証実験

### （１）実験内容

実証実験では，土木用語大辞典の見出し語について，前章で定義した指標に基づき，「ノイズ語」，「一般語」と「専門用語」を抽出する．そして，抽出したそれらの見出し語が，実際に「ノイズ語」，「一般語」と「専門用語」であるかを分析する．最後に，土木用語大辞典の全見出し語から，これら 3 種類の見出し語を除外し，コア用語候補を抽出した後に，抽出したそれらの用語について考察する．また，全見出し語中で，アルファベットのみで構成される見出し語 323 件は，他の見出し語の略語であり，コア用語候補としてふさわしくないため，処理対象外とした．

### （２）不要語の抽出に関する実験結果

#### a） 土木用語大辞典の分類結果の概要

土木用語大辞典の見出し語をコア用語候補と，その他の語である「ノイズ語」，「一般語」及び「専門用語」（以下，不要語と略記）に分類した結果の件数を**表－４**に示す．

**表－４ コア用語候補選定処理で抽出された用語**

| | 内容 | 抽出件数 |
|---|---|---|
| 不要語 | ノイズ語 | 2,249 |
| | 一般語 | 2,253 |
| | 専門用語 | 2,684 |
| コア用語候補 | | 14,609 |
| 合計 | | 21,795 |

分類した結果，不要語は 7,186 件，コア用語候補は 14,609 件となった．この結果は，全体の約 67%がコア用語候補であることを示している．土木用語大辞典の規模と，**図－１**に示す通り，コア用語候補が複数階層で構成されていることを考慮すると，問題のない数字であると考えられる．

#### b） 不要語の抽出結果

本研究にて，不要語と設定したノイズ語，一般語，専門用語とコア用語候補の集合から抽出した 50 件をそれぞれ，**表－５**，**表－６**，**表－７**と**表－８**に示す．**表－５**から，ノイズ語として抽出された結果には，著者らが提案した通り，公式や法律名，固有名詞が含まれていることが分かった．しかし，一部には硬質塩化ビニルパイプカルバートなどが含まれており，コア用語ではないが，専門用語に近い見出し語が抽出されていることが分かる．**表－６**を確認すると，一般語として抽出された結果には，**表－１**に示した定義の通り，畑や土など分野を横断して広く利用される一般的な概念が大多数を占めていることが分かる．**表－７**を確認すると，専門用語として抽出された結果には，**表－１**に示した定義の通り，専門性が高い用語を取得できていることが分かる．**表－８**は，全見出し語から，負用語を削除したコア用語候補を表す．コア用語候補の見出し語間ネットワークの分析については，本章の（３）で行う．次に，抽出した不要語についてそれぞれ詳細に分析する．

#### c） ノイズ語についての分析

本項では，本提案手法で取得したノイズ語について詳細に分析する．ノイズ語の分析では，ノイズ語として判別した 2,249 件を目視で確認し，その見出し語にどのような内容を示す語が含まれているかを集計した．ノイズ語として抽出した見出し語の内訳を**表－９**に示す．**表－９**に示す通り，固有名詞が 1,563 件であり，ノイズ語として抽出した見出し語中の約 69%を占めている．また，集計した結果，固有名詞以外のその他の見出し語が 687 件抽出された．これらの見出し語は，「波の予測」や「浮体の安定」など，対象の概念を表す単一の語がなく，複数の語で構成された複合語であった．このため，本提案手法で取得したノイズ語は，

**表－５ ノイズ語削除処理で抽出した見出し語の例**

| | | | | |
|---|---|---|---|---|
| ランベルト正(等)角円錐図法 | ダルシーの法則 | 関西文化学術研究都市 | 貞山堀 | 社会資本整備の財源 |
| A セグメント | ニュートンの運動方程式 | 近郊整備地帯 | 土木研究所 | 線分の交差判定 |
| H 形桁橋 | ピアソン 3 型分布 | 古第三紀 | 縄張 | 堤防の景観設計 |
| NP 標準活荷重 | ブリネル硬さ試験 | 硬質塩化ビニルパイプカルバート | 尺 | 波のエネルギー |
| T 溶接継手 | ポインティング効果 | 根室半島沖地震 | 北伊豆地震 | 分岐器の番数 |
| アルベーグ式 | ムーディ図表 | 治水事業五箇年計画 | 輪虫類 | 波の発生 |
| オイラーの方程式 | ランキン土圧 | 新幹線保有機構法 | サンフルーの簡略式 | 波の連なり |
| キャブシステム | ワーグナー関数 | 成富兵庫茂安 | 運動量の定理 | 浮体の安定 |
| ゲルバー橋 | 横浜水道 | 騒音規制法 | 橋梁の耐震 | 放射性物質の地中移動 |
| シュタッケルベルグ均衡解 | 火山噴火予知連絡会 | 宅地等水防災対策事業 | 硬化コンクリートの性質 | 有害廃棄物の越境移動 |

**表－６ 一般語削除処理で抽出した見出し語の例**

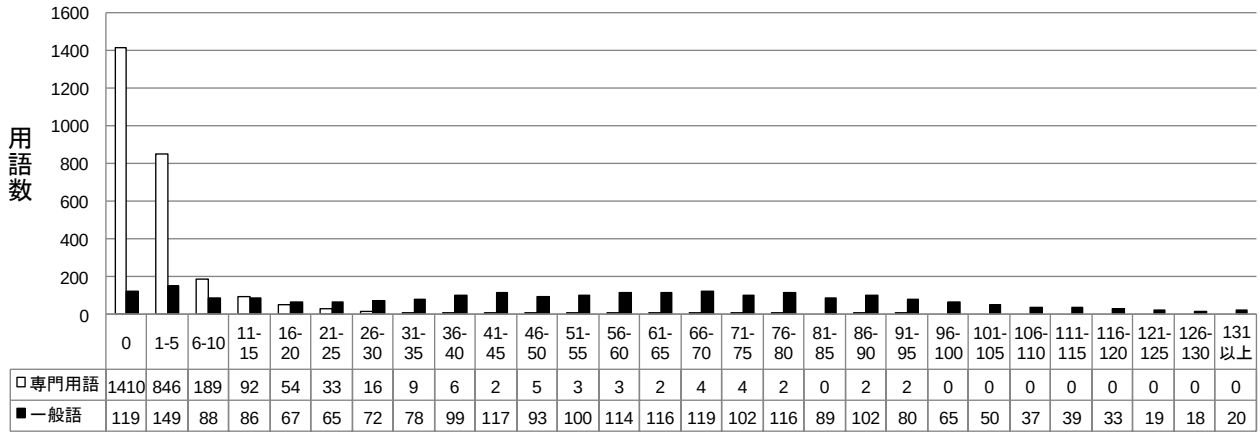
| | | | | |
|---|---|---|---|---|
| バー | 督促 | スノーネット | はね返り | メモリー |
| ウイルス | 捕食 | コンストラクター | 都市活動 | 計量 |
| ウェブ | レラクゼーション | ゼロサムゲーム | 市区改正 | 農業工学 |
| おろし | 水締め | ライトペン | エッジ | スキージャンプ |
| くびれ | 工事運営 | アースドリル | 画素 | 塗料 |
| タール | 買収方式 | オペレーティングシステム | 振幅 | 空気量 |
| ボリューム | アヴェニュー | ベータ線 | 地方部 | 地すべり |
| 絞り | 織込み | コンピューターグラフィックス | プロトコル | 森林 |
| 石英 | 閉鎖工事 | ユニットロード | ダイナマイト | 消化 |
| 石鹸 | オペレーションズリサーチ | 通行止め | 鉄道 | トンネル |

**表－７ 専門用語削除処理で抽出した見出し語の例**

| | | | | |
|---|---|---|---|---|
| セメント系吹付け法面 | 焼なまし鉄線 | ホットビン | 公開競争入札方式 | 手掘り式シールド |
| 海底鉱物資源開発基地構想 | 動水勾配線 | 井桁擁壁 | 土量換算係数 | 第三者賠償責任保険 |
| 軽荷重スラブ橋用プレストレスト橋桁 | 漏水量測定堰 | 軌間狂い | 振動ローラー締固めコンクリート | プレストレッシング工 |
| 細粒度アスファルト混合物 | 歩道用コンクリート平板 | 行止り式停車場 | ロックボルト引抜き試験 | 白色ポルトランドセメント |
| 状態空間型物理的流出モデル | 常時使用水量 | 縦補剛材 | 浮遊砂濃度計 | 金属二重殻貯槽 |
| 側水路式流入部 | アスファルトレベリング層 | 舌状砂州 | 有効土被り圧 | マイクロ波高度計 |
| 底設導坑先進上部半断面工法 | グラウチングトンネル | 地質縦断面図 | logt 法 | ポルトランドセメント |
| 波力発電式防波堤 | スクリードフィニッシャー | 道床係数 | 制動停止視距 | 浅層安定処理 |
| 方杖付き張出し床版 | ディシジョンクリティカルパス法 | 複剪断継手 | 沿道環境整備制度 | 有効応力解析法 |
| 共回転応力速度 | パラセントル | 眩光防止施設 | 曲げモーメント－曲率関係 | コンクリートブロック舗装 |

**表－８ コア用語候補選定処理で抽出した見出し語の例**

| | | | | |
|---|---|---|---|---|
| インバート | 越境廃棄物 | 広域運営 | 石こう | 濃度計 |
| クラウンバー | 音線 | 高強力鋼 | 潜在水硬性 | 発電効率 |
| サーフェスモデル | 火山性地震 | 最急勾配 | 走査計 | 氷期 |
| スペース機能 | 階層型データベース | 散逸系 | 代替技術 | 伏角 |
| テクスチャー解析 | 環境アセスメント条例 | 時価見積り | 地すべり危険箇所 | 偏平断面 |
| バケットローダー | 希望線図 | 斜杭 | 中央構造線 | 膨張性地盤 |
| ふるい砂利 | 漁法 | 充実率 | 沈降分離 | 目標計画法 |
| マッチキャスト工法 | 空港除雪 | 消失点制御 | 天端沈下 | 螺旋シェル |
| レール用ばね釘 | 結合水 | 浸透説 | 土木会議 | 量反応関係 |
| 移流分散 | 湖面蒸発 | 水道システム | 特定多目的ダム | 艤装 |




**図－10 一般語と専門用語が登場する論文誌数**

**表－９ ノイズ語として抽出した見出し語の内訳**

| 内容 | | 単語数 |
|---|---|---|
| 固有名詞 | 特定の法律・計画・会議・制度・記録等に関する用語 | 350 |
| | 特定の工法・事業名に関する用語 | 79 |
| | 特定の団体に関する用語 | 85 |
| | 特定の物理法則に関する用語 | 227 |
| | 特定の歴史・イベントに関する用語 | 84 |
| | 特定の人名・地名・物・施設名 | 290 |
| | 上記以外の用語 | 448 |
| その他 | | 687 |
| 合計 | | 2,249 |

固有名詞が主であり，その他の見出し語として複数の語で構成される複合語が含まれることが分かった．

これらの結果から，本提案手法で取得したノイズ語は，コア用語候補とは異なり，不要語だと考えられる．

**d） 一般語と専門用語についての分析**

本項では，本提案手法で取得した一般語と専門用語について詳細に分析する．一般語と専門用語の分析では，それぞれの見出し語が技術文書にどの程度含まれているかを尺度として評価を行った．まず，論文検索エンジンである Cinii[12]を利用して，各見出し語が含まれる論文を取得する．次に，それらの論文が掲載された論文誌を取得する．そして，各見出し語が出現する論文誌の種類数を集計する．これにより，専門用語は，技術文書に存在しないか，利用される論文誌の種類が限定される傾向にあり，一般語は，複数の技術文書に出現するため，複数の論文誌で利用される傾向があると考えられる．一般語と専門用語について，登場する論文誌数で集計した結果を**図－10** に示す．**図－10** に示す通り，一般語は，多くの論文誌に存在する見出し語で構成されていることが分かる．一方，専門用語は，まったく技術文書に登場しない見出し語が 1,410 件（専門用語全体の約 52%），1～5 種類の技術文書に登場する見出し語が 846 件（専門用語全体の約 31%）となっており，大半の見出し語が利用される分野が限定された用語であることが分かる．ここで，0 件，1～5 件の区分に一般語が 268 件分類されていることが分かる．これらの区分に分類されている見出し語について，詳細に分析したところ，51 件（約 19%）が土木分野の用語であることが分かった．これらの用語は，**表－３** の条件 4 によって抽出された見出し語であるが，一般語全体の約 2%であるため，抽出精度という点を考慮すると高精度に一般語を取得できていることが分かる．

**e） まとめ**

不要語の抽出に関する実験は，実験結果の分析結果の通り，ノイズ語，一般語と専門用語ともに，適切な結果であったことが分かった．そのため，土木用語大辞典の見出し語から，これらの不要語を除去したコア用語候補は，コア用語に適した見出し語が多く含まれると考えられる．

**（３）見出し語間ネットワークの評価実験**

**a） 実験内容**

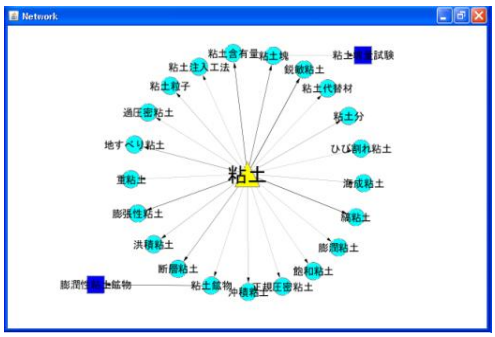
本実験では，本提案手法で分類した一般語，コア用語候補，専門用語のネットワーク構造を CytoScape[13]により可視化し，その構造に基づきコア用語の特性を解析する．本解析により，一般語，コア用語と専門用語の関係を明らかにする．また，本実験では，より見出し語間の関係を明確化するため，第 3 章で構築した見出し語間ネットワークに関係強度を付加する．見出し語間の関係強度は，次に示す 3 つの手順で算出する．まず，各見出し語が含まれる論文の論文誌名とその件数を計算し，見出し語毎に論文誌ベクトルを生成する．次に，見出し語間ネットワークにおいて関連する任意の 2 つの見出し語を取得し，論文誌ベクトルを用いて類似度を算出する．ここで，類似度の算出には，Vector Space Model[14]を用いる．最後に，任意の 2 つの見出し語のすべての組み合わせを対象に，類似度を算出する．

本実験では，見出し語間ネットワークの評価を行うためのネットワークとして，「粘土」と「剪断」を起点としたネットワークを採用した．

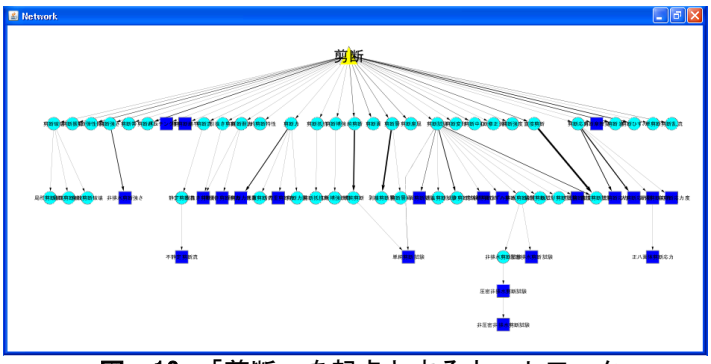
**b） 実験結果**

「粘土」と「剪断」について，見出し語間ネットワークを構築した結果をそれぞれ**図－11**，**図－12** に示す．見出し語間ネットワークでは，一般語を「三角」，コア用語候補を「丸」，専門用語を「四角」で表現する．

まず，**図－11** を確認すると一般語「粘土」を中心として，コア用語が広がっており，コア用語の周りに専門用語が広がっていることが分かる．**図－11** のネットワークでは，「粘土」と一つの語が関連付くことによって，コア用語候補が形成されており，コア用語候補と一つの語が関連付くことにより，専門用語が形成されていることが分かる．この結果から，土木用語大辞典の見出し語は，語の専門性に従って階層的に広がっていることが分かった．次に，**図－12** を確認すると「剪断」を最上位階層として，6 階層のネットワークが構成されている．各ネットワークの経路を確認すると，「一般語→コア用語候補（N 回以上）→（専門用語）」，




**図－11 「粘土」を起点とするネットワーク**




**図－12 「剪断」を起点とするネットワーク**

「一般語→専門用語」となっており，「粘土」の見出し語間ネットワークと同様に一般語に語句を加えることで専門性の高い用語が構成されることが分かる．また，特に階層が深くなるにつれて専門用語の割合が増加しており，2 階層目 10%，3 階層目 43%，4 階層目 80%，5 階層目以降は 100%となっている．これらの結果から，階層を追うごとに専門性が高くなる傾向にあり，場合によっては 2 階層目で専門用語と判断される可能性もあることが分かった．

## ６．おわりに

本研究では，建設分野における XML スキーマの相互運用性を向上させるための基礎研究として，土木用語大辞典の見出し語を複数の指標で分析し，スキーマの要素名に使用すべき用語セットをコア用語として機械的に選定するための手法を提案した．本論文では，土木用語大辞典から「専門用語」，「一般語」，「ノイズ語」を分類する指標を定め，適切に選定可能であることを実証した．そして，専門用語，一般語とコア用語候補を見出し語間ネットワークとして可視化したところ，一般語を上位概念，専門用語を下位概念とする階層構造が構築されており，著者らが分析した通り，中間層にコア用語候補の層が存在することが明らかとなった．抽出された用語は，「建設分野において一般的に利用可能」，「専門性が高い専門用語ではなく一般性が高い」，「他の分野と比較して建設分野での利用頻度が高い」という 3 つの特徴を持つ．これらの特徴から，新しくスキーマを作成する際に，多くの設計者が要素名としてこれらのコア用語を統一的に選択することで，過去のスキーマで用いられてきた概念と共通の概念を同じ用語で表現でき，相互運用性の高いスキーマを作成できると考えられる．本研究は，コア用語の選定指標に建設分野特有の内容を用いていないため，その業界全体の内容を網羅した用語辞書があれば，他の業界においても応用可能であると考えられる．

今後は，本研究で得られたコア用語候補から，社会に公開するべき土木建設業界のコア用語を正しく抽出する手法について検討する．

**謝辞**：財団法人日本建設情報総合センターの方々には大変お世話になりました．ここに記して感謝致します．また，本研究の一部は，平成 20～24 年度私立大学戦略的研究基盤形成支援事業（研究課題「セキュアライフ創出のための安全知循環ネットワークに関する研究」）から助成を受けその成果を公表するものである．

## 参考文献

1) 国土交通省:「国土交通省 CALS/EC アクションプログラム 2005」の策定について, 2006 年 3 月
2) 国土交通省:「国土交通省 CALS/EC アクションプログラム 2008」の策定について, 2009 年 3 月
3) Jeong, B., Lee, D., Cho, H. and Lee, J. : A Novel Method for Measuring Semantic Similarity for XML Schema Matching, *Expert Systems with Applications:An International Journal*, *Pergamon Press*, Vol.34, No.3, pp.1651-1658, 2008.
4) Madhavan, J., Bernstein, A. and Rahm, E. : Generic Schema Matching with Cupid, *Proceedings of the 27th International Conference on Very Large Data Bases*, *Morgan Kaufmann Publishers*, pp.49-58, 2001.
5) 中川剛, 石原靖哲, 藤原融:異なる構造を持つ XML データを統合するための構造間の関係について, データベース・システム研究報告, 情報処理学会, No.2001-DBS-126, pp.145-152, 2002 年 1 月.
6) Sanz, I., Mesiti, M., Guerrini, G. and Berlanga, R. : Fragment-based Approximate Retrieval in Highly Heterogeneous XML Collections, *Data & Knowledge Engineering*, *Elsevier Science Publishers B.V.*, Vol.64, No.1, pp.266-293, 2008.
7) Gama, J. and Brazdil, P. : Cascade Generalization, *Machine Learning*, *Kluwer Academic Publishers*, Vol.41, No.3, pp.315-343, 2000.
8) Zhao, H. and Ram, S. : Entity Matching across Heterogeneous Data Sources : An Approach Based on Constrained Cascade Generalization, *Data & Knowledge Engineering*, *Elsevier Science Publishers B.V.*, Vol.66, No.3, pp.368-381, 2008.
9) 国土交通省国土技術政策総合研究所:建設分野における XML 記述仕様の考え方(案), 2009 年 3 月.
10) 土木学会:土木用語大辞典, 1999 年 2 月.
11) Google<http://www.google.com>,（入手 2010.5.17.）.
12) Cinii<http://ci.nii.ac.jp/>,（入手 2010.5.17.）.
13) Cytoscape Consortium:Cytoscape, <http://www.cytoscape.org/>,（入手 2010.5.17.）.
14) Salton, G., Wong, A. and Yang, C. : A Vector Space Model for Automatic Indexing, *Communications of the ACM*, *ACM*, Vol.18, No.11, pp.613-620, 1975.

（2010.5.28 受付）