大学研究者総覧 DB を対象とした横断検索のための概念抽出・統合の試み

A trial on extracting and integrating concepts of university staff directories for federated search

蔵川 圭[*1] Kei Kurakawa　Aman Shakya[*2]　武田 英明[*1,3] Hideaki Takeda

[*1] 国立情報学研究所 National Institute of Informatics
[*2] 総合研究大学院大学 The Graduate University for Advanced Studies
[*3] 東京大学 The University of Tokyo

Most of Japanese universities have provided their own staff directories on the Web. These directories are consisted of their own original attributes of the staff and multifunctional search services. This paper aims at providing a service for users to search Japanese researchers across universities. One of the ways for providing the service is to extract concepts of researcher entity from staff directories different from each other, and define relationships among them. The StYLiD we've developed is a tool to define the relationships among different concept definitions. We conducted a case study of defining concepts of several Japanese university staff directories with the StYLiD, so that it makes it possible to search researchers across universities.

1. はじめに

論文執筆を主体とした学術情報流通が Web という媒体を通して行われることが一般的になった今日，その主要な主体である研究者の業績も研究教育機関によって Web 上に公開されてきている．ほとんどの大学では，組織に属する教員や研究スタッフの研究業績を含めた紹介を研究者総覧と称し，データベースを通して Web 上に公開している．大学や職員の評価に関連して業績を公開している場合もあれば，組織の広報として公開している場合もある．

大学の研究者総覧データベースは，利用者の立場に立てば様々な目的で利用されている．一般の利用者が，ある研究者自身がどのような業績をあげているのかを知るための目的もあれば，組織に属する研究者の主要な研究分野を知ることも目的の一つである．研究の主要な原資が税金である組織にとって研究成果を公開することは義務であり，納税者が研究成果を閲覧することもある．また，論文誌の編集者が投稿された論文の査読者を探すときにも利用できる．

国内では類似の Web 上の公開データベースとして，科学技術振興機構の運営する研究開発支援総合ディレクトリ ReaD[*1] がある．多くの研究者は所属する機関の研究者総覧と同様の業績データを，ReaD に直接アクセスすることによって更新している．いくつかの先進的なシステムを備えた大学では，機関の研究者総覧のデータを自動的に ReaD に登録する仕組みを備えているが，すべての研究者がその恩恵を受けているわけではない．

その他に，自由登録制の産学プラザ[*2]，SNS の機能を備えた ResearchMap[新井 09]などが国内の研究者の紹介をおこなったサイトとしてあげられる．

このようにほぼ同様の内容を紹介する複数のサイトがある状況では，研究者は自身の業績データを複数個所で更新しなければならず，複数のサイトにアクセスして同様の作業を行うことは非効率的であると感じあまり積極的ではない．そのため，必要性のある身近なサイトから更新作業を行っているようである．所属組織の研究者総覧の更新が最も頻繁であり，正確性を担保していることは容易に想像できる．

したがって，日本の研究者を総覧するためには，現状では更新の頻度や正確性を最も担保した大学研究者総覧を対象として研究者のデータを閲覧すればよいことになる．実際には，すべての大学の研究者総覧に対して横断的に検索できることが必要となろう．

個々の大学の研究者総覧におけるデータは，列挙された項目やその表現が若干異なる．横断的で対象項目を指定した検索を実現するためには，個々の研究者総覧のデータ項目の概念を見比べてつなげ，統一的な検索項目として関係を明示する必要がある．そうすれば，横断的に氏名で検索することや研究分野で検索することが可能となる．

本論文では，複数の大学研究者総覧に対し，研究者データの項目を一つの概念系としてとらえ，われわれがすでに開発した StYLiD[*3] という概念定義と個別データの入力が可能なプラットフォームを用いて概念系同士の関係づけ，研究者データを統一的に検索が可能となることを示す．

以下，次章では StYLiD について紹介する．つづけて，第 3 章では実際に複数の大学研究者総覧を例として StyLiD を用いたケーススタディについて述べる．これによって異なる項目や表現の研究者総覧を統一的に検索閲覧可能であることを示す．最後に，本論をまとめる．

2. StYLiD

2.1 システム概要

プラットフォーム StYLiD[Shakya 08a,08b,08c]は，簡便なユーザインタフェースから概念定義と個別データ(インスタンス)の入力ができるシステムであるが，概念定義において一つの概念に対して複数の異なる定義を付与できるところが特徴である(図 1)．

一般にオントロジーを構築する場合，個々の概念に一意の定義を与える．しかし，実際にはこのような汎用の概念を定義することは容易ではない．そこでこのシステムでは各個人で自由に

連絡先：蔵川圭，国立情報学研究所 学術コンテンツサービス研究開発センター，〒101-8430 東京都千代田区一ツ橋 2-1-2 ， TEL:03-4212-2372 ， FAX:03-4212-2374 ， kurakawa@nii.ac.jp

[*1] http://read.jst.go.jp
[*2] http://www.sangakuplaza.jp/
[*3] http://www.stylid.org/

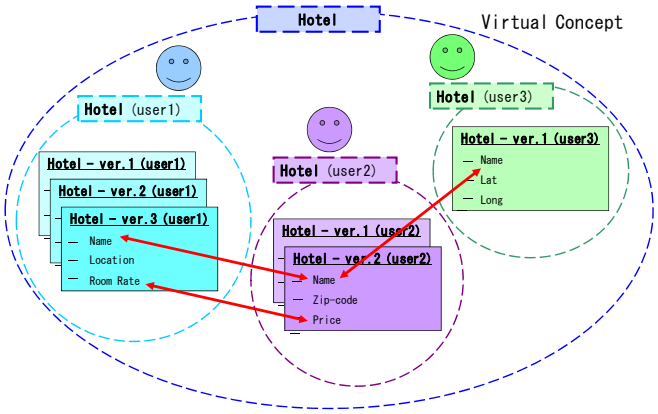

図 1 StYLiD を用いた概念定義

概念定義ができるようになっている．その際，同じ概念に異なる定義があってもよい．そしてその自分の定義した概念に合わせてデータ(インスタンス)を入力することができる．

StYLiD では同じ概念を異なる定義で定義した概念を集めたものを統合概念(consolidated concept)と呼んでいる．統合概念は個別に定義された概念集合とのアライメントルール(alignment rule)集合を持つ．このアライメントルールは異なる概念の属性間で共通なものを指定する．このルールを指定することで検索時に共通の属性として検索することができるようになる．

2.2 システムの操作手順

(1) 概念の定義

概念定義とは，概念の名前，属性名と属性の値域の組のリストを与えることである．属性の値域とは属性として適切なクラスを指定することである．属性の値域はなくてもよい．インスタンス定義とは概念定義に基づいて個体(個別の情報)を入力することである．

個々のユーザは自由に概念を定義することができる．定義に当たってはほかのひとの定義を使っても構わないし，自分の既存の概念を拡張してもよい．

(2) インスタンスの定義

インスタンス定義においては，適切な概念を選び，その概念の属性の属性値を入力する．属性値の入力においては属性の値域として指定されているクラスのインスタンスが推奨されるが，それ以外も入れることができる．値域のクラスのインスタンスであれば，現在のデータベースを検索して選んで入力することができる．また Wikipedia のページを属性値として指示することができる．その場合は，実際の入力としては DBpedia の URI が入力される．

(3) 統合概念の検索と編集

概念はタグクラウド方式で検索することができる．このタグクラウドは普通のタグクラウドと異なり，階層的に構造化されている．まず一階層目においては統合概念のみが表示されている．なお一つしか定義をもたないものも統合概念である．多くの定義があるものは大きく表示される．次に複数定義が存在する統合概念をクリックすると，その概念は展開され，作成者と概念名の組のリストが表示される．複数のバージョンをもつ作成者の概念は大きく表示される．さらにこの一つの組をクリックすると，バージョンが表示される．

このタグクラウドを使うことで，ユーザは概念を絞り込んでインスタンスを検索することができる．

また，このタグクラウドから統合概念を選んで，アライメントルールを定義することができる．個別概念の属性のどれが統合概念のどの属性に対応するかを指示する．なお，システムは文字列の類似度を計算することで蓋然性の高いルールは自動的に構成する．この類似度の計算においては WordNet を用いて表記上異なるものでも意味的に近いものは発見できるようにしている．

ルールが付与された統合概念は個別概念のインスタンスをそのインスタンスであるかのように表示することができる．また検索においても同様である．

3. 大学研究者総覧 DB を対象とした概念定義

3.1 大学研究者総覧 DB のクローリング

StYLiD を用いて，複数の大学研究者総覧 DB のデータを統一的に検索できることを示す．そのための基礎実験として，大阪大学と名古屋大学の二つの研究者総覧 DB をピックアップして対象とした．2009 年 4 月現在，大阪大学研究者総覧[*1] では 1881 名の研究者が登録公開され，名古屋大学研究者総覧[*2] では 1106 名の研究者が登録公開されている．

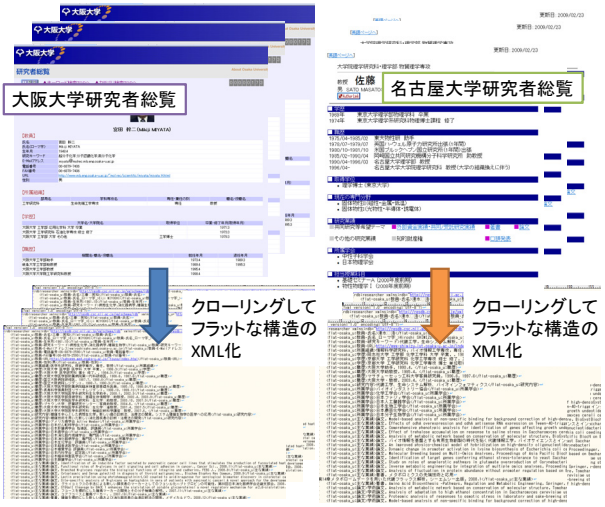

図 2 クローリングによる大学研究者総覧 DB データの取得

図 3 フラットな XML 構造を持つ研究者データ

*1http://kenpro.mynu.jp:8001/scripts/websearch/index.htm?lang=Jhttp://www.spock.com

*2http://www.dma.jim.osaka-u.ac.jp/kg-portal/aspi/rx0011s.asp

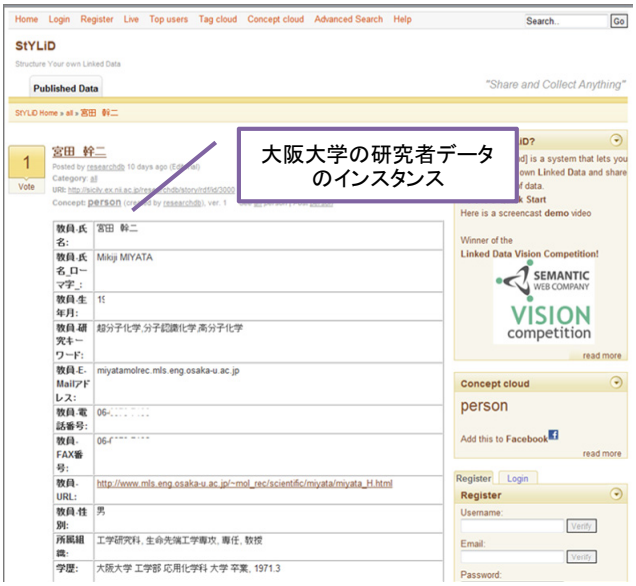

図 4 大阪大学の研究者データの StYLiD 上での表示

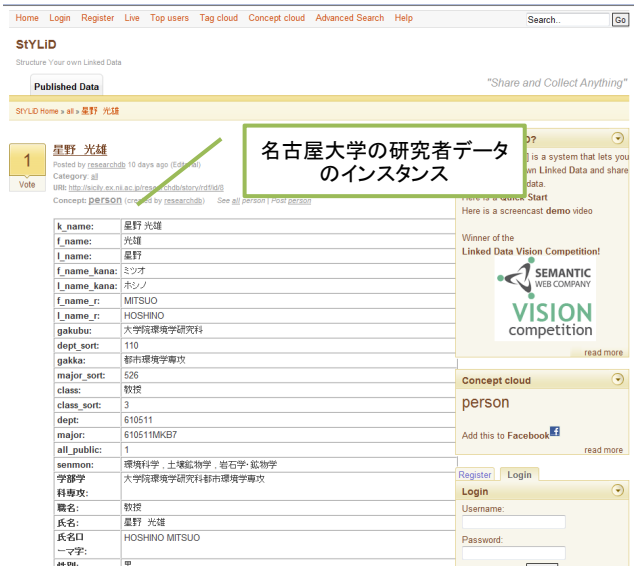

図 5 名古屋大学の研究者データの StYLiD 上での表示

現在の StYLiD はフラットなデータ構造を許容しデータインスタンスを登録できるので，2 つの大学研究者総覧 DB のデータをクローリングしてフラットな構造の XML データを構築した（図 2）．データを作成する際に，各研究者のページで表示されている大項目名や小項目名を連結させて属性名を作り階層化のないデータ構造とし，連結させた小項目に含まれるデータをカンマ区切りで連結させた（図 3）．たとえば，「教員」という大項目に「氏名」や「性別」などの小項目があった時には，「教員_氏名」や「教員_性別」などと項目名を連結して属性名を定義した．

3.2 大学研究者総覧 DB 項目からの個別概念の定義

大学研究者総覧 DB からクローリングによりデータを取得した際に作成した XML スキーマを参照して，StYLiD での概念定義とインスタンスデータの登録を行った．今回は，大阪大学研究者総覧 DB から取得したデータと，名古屋大学研究者総覧 DB から取得したデータに対し登録を行った．StYLiD では，それぞれ図 4，図 5 のように表示される．

図 6 大阪大学と名古屋大学の研究者データ属性項目から共通概念を抽出し StYLiD 上で統合概念を構成

3.3 複数の個別概念定義を対象とした大学研究者総覧 DB 統合概念の定義

次に，登録した研究者の 2 つの概念定義に対して，統合概念を定義する．研究者という概念に対して，それぞれ独自に属性概念および属性名を定義しているので，共通する属性概念を探して統合概念のひとつの属性として定義する．StYLiD 上では，図 6 に示すように，すでに定義した大阪大学の研究者データの属性名と名古屋大学の研究者データの属性名がプルダウンメニューから選択可能であるので，共通する属性概念を選択して，右側の統合概念のフィールドに属性名を入力する．たとえば，大阪大学の「所属組織」を名古屋大学の「学部学科専攻」を同一の属性概念とみなし，統合概念として「学部学科専攻」という属性名をつけている．

StYLiD は，ユーザが統合概念を容易に構成できるよう，属性名の類似度を機械的に判定して，初期状態として統合概念の候補をあらかじめ列挙している．ユーザは，ここから属性概念の一致の判断を試行錯誤しながら行い，統合概念を構成することとなる．時に，一方の属性概念は他方の属性概念集合に一致するものがない場合もあり，その場合統合概念として定義することはできないので削除したりする．また，機械的に対応する属性を列挙していないものは追加することとなる．図 6 は試行錯誤後の統合概念である．この例では 19 組の対応のうち 10 組はシステムが提案したものをそのまま採用しており，9 組は追加している．

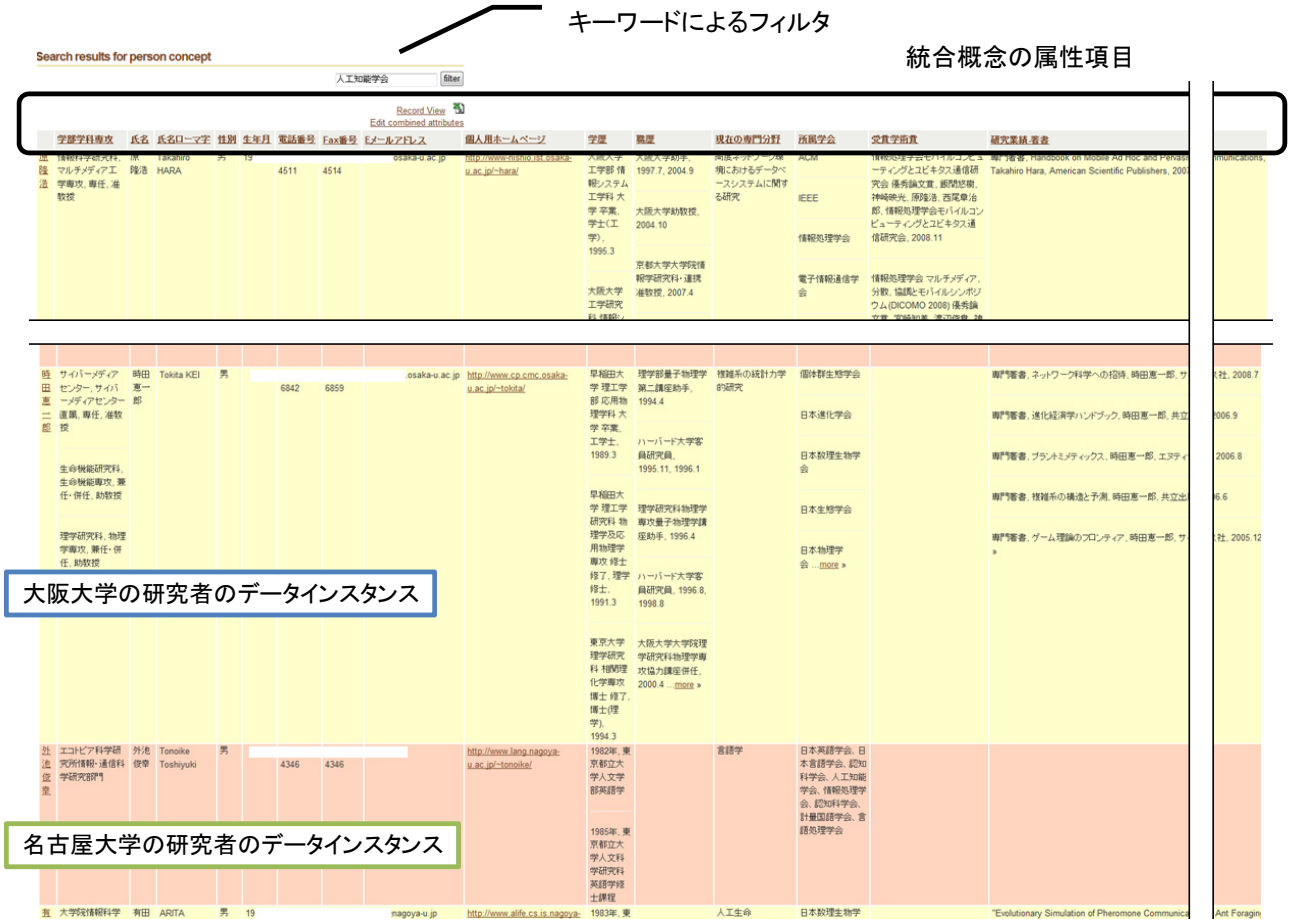

図 7 統合概念によって共通のフレームワークで閲覧可能となった大阪大学と名古屋大学の研究者データ

3.4 大学研究者総覧 DB 統合概念による検索・閲覧

構成した統合概念を用いて，大阪大学と名古屋大学の 2 つの研究者データを共通のフレームワークの中で検索閲覧することができる．図 7 は，その一覧テーブルである．統合概念の属性名がテーブルのトップに表示されている．その下には研究者のデータが配置されており，大阪大学の研究者のデータと名古屋大学の研究者のデータが並列して表示されていることが見て取れる．

一覧テーブルに列挙された研究者のデータは，キーワードによるフィルタによって選択的に表示可能である．図では，「人工知能」とキーワードを入れてフィルタをかけたところ，全 2987 名のうち 71 名がリストされた．

4. おわりに

今回の実験によって，大学研究者総覧という必ずしも同一ではないが類似の属性項目を持つ複数のデータベースを対象として，研究者データ項目の概念を統合した共通のフレームワークのもとで研究者情報を提示できることがわかった．このことは，各大学の研究者総覧を一次情報源とした日本の研究者情報提供サービスを構築できることを示唆する．サイトを限定してデータをクローリングし全文検索サービスを提供する以上の価値は，データ項目間の関係性を維持し意味付けによって統合概念を構成するところから生まれるであろう．

大学研究者総覧は，論文情報を提供する機関リポジトリのような定型的なデータ項目のみによって構成されるデータベースではない．むしろ，大学の特徴とでもいうべき特別な項目が入っていることの方が多い．このようなデータベースの特徴があるとき，属性項目を利用して横断的に検索可能とするには StYLiD のようなデータ項目間の意味付けと統合を行うプラットフォームが有効活用できるのである．

今後は，StYLiD において複数の属性を統合して一つにできるように，統合概念の構成の多様性を目指す．また，実験においても今回取り上げた 2 つの大学研究者総覧だけでなく，より多くのデータベースを対象とし実現性を調査したい．

参考文献

[Shakya 08a] A. Shakya, H. Takeda and V. Wuwongse: StYLiD: Structure Your Own Linked Data, in International Conference on Weblogs and Social Media (ICWSM2008), pp. 220–221, Seattle, Washington (2008), Association for the Advancement of Artificial Intelligence.

[Shakya 08b] A. Shakya, H. Takeda and V. Wuwongse: StYLiD: Social Information Sharing with Free Creation of Structured Linked Data, in Proceedings of the Social Web and Knowledge Management Workshop (SWKM 2008), pp. 33–40, Beijing, China (2008), Located at the 17th World Wide Web Conference (WWW2008).

[Shakya 08c] A. Shakya, H. Takeda and V. Wuwongse: Consolidating User-Defined Concepts with StYLiD, in Proceedings of 3rd Asian Semantic Web Conference (ASWC 2008), pp. 287-301, Bangkok, Thailand (2008)

[新井 09] 新井 紀子:サイエンス 2.0 へようこそ Researchmap.jp について, 情報管理, Vol. 52, No. 1, (2009), 12-19.