様式５

平成２５年　３月３１日

京都大学人文科学研究所共同研究実績・活動報告書

（　３年計画の　３年度目）

１．研究課題

（和文）情報処理技術は漢字文献からどのような情報を抽出できるか ― 人文情報学の基礎を築く

（英文）What information can be extracted from Kanbun texts with computational methods ?
-- A contribution to fundamental research in Digital Humanities --

２．研究代表者

（氏名）山崎直樹(関西大学)

３．研究期間

平成　22年　　7月　から　平成　25年　　3月　まで

４．研究目的（400字程度）

本課題は，現代の情報処理技術が東アジア古代社会の遺産として残された漢字文献からどのような情報を抽出できるかという問題に対し，複数の角度からその可能性を探り，人文情報学の基礎を築くことを最終目的とした。そして、そのために、現状に対する認識を踏まえ、下記の課題(甲)-(丙)を検討をすることを作業目的とした。課題(甲): 非文字データのように、これまでメタデータを付与することが困難なだと考えられてきたデータへのメタデータ付与にはどのような問題点があり、どのような可能性があるか。課題(乙): 文字データにせよ、非文字データにせよ、メタデータを付与することにより、単なる機械可読ではなく、機械が意味を読み取ることが可能となったデータは、そのデータ相互をどのように関連づけたらよいのか。「セマンティックウェブ」という枠組みで示された方向でよいのか。課題(丙): 我々が分析の対象とする文字データの集まり＝テキストは、どのような構造をもっているものとしてモデル化できるか、また、非文字データの集積体の基本的なデータフォーマットは、画像や動画であると考えられるが、それらの集積体（例: 記録映画など）はどのような構造をもっているものとしてモデル化できるか。

５．本年度の研究実施状況（400字程度）

5月25日、7月6日、9月14日、10月12日、11月2日、11月16日に研究班を開催し、これまでの研究およびシンポジウムやセミナーでの議論を総括した。また、本研究の課題について確固たる見通しを得るべく、2月16日に公開シンポジウム『すべてをコンピュータの中に(繋がってしまったデータとその未来)』を開催した。発表タイトルおよび発表者は以下の通りである。

(1) 漢字構造情報のRDF化の試み／守岡知彦（京都大学）

(2) 国立国会図書館のメタデータ標準– データを繋げるメタデータ: DC-NDL／柴田洋子

（国立国会図書館）

(3) PageRank と学術論文の評価: ノーベル賞の窓を探そう／藤田裕二（(株)ターンストーンリサーチ, 日本大学）

６．研究成果の概要（400字程度）

これまでの研究期間に行ったシンポジウムやセミナーおよびそこでの議論を通じて、下記の認識を得た。

> データを繋げること自体は容易であるし、すでにある種のデータは繋がってしまっている。しかし、データがリンクすることと、その繋がったデータの集積体が特定の目的をもった探索にふさわしい構造をもっているかは別の問題である。機械可読、とくにセマンティックな面での機械可読を実現するためには、データにどのような構造をもたせ、繋げられたデータの集積体にどのような構造をもたせるのかについて、理論的な基盤が必要である。

この認識に基づき、公開シンポジウム『すべてをコンピュータの中に(繋がってしまったデータとその未来)』を開催した。以下に各発表で得られた知見を示す。

(1)「漢字構造情報のRDF化の試み」は、漢字構造情報を機械可読にする表現の代表的な形式であるIdeographic Description Sequence（IDS）を、セマンティックウェブを支える基幹技術として、メタデータやオントロジーを表現するに開発されたRDFで表現した場合、どのような結果を招くのか、その可能性と問題点を論じた。

(2)「国立国会図書館のメタデータ標準」は、書誌情報のメタデータ形式として標準化されているDublin Coreから出発したものの「もはや、Dublin Coreではない」と評される、国立国会図書館によるメタデータ形式DC-NDLについて、その変遷や、概念モデル、データフォーマットについて解説がなされた。

(3)「PageRank と学術論文の評価」は、代表的な検索エンジンであるGoogle Searchの基盤技術であるPageRankと、その人文科学の他分野への応用について論じた。PageRankは、データとデータのリンクから再帰的にデータの重要度を決定するというアイデアを基にしているが、このアイデアを現実に存在する文書間のネットワークの解析に応用するためにはいくつか実装上の問題点がある。これらの問題がどのように解決されたか、そしてそれを、「学術論文間の引用ネットワーク」というデータ構造の解析に応用するには何が必要かが報告された。

７．共同研究会に関連した公表実績（出版、公開シンポジウム、学会分科会、電子媒体など）

2013年2月16日に公開シンポジウム『すべてをコンピュータの中に(繋がってしまったデータとその未来)』を開催した。また、これに合わせて、予稿集『すべてをコンピュータの中に』(京都大学人文科学研究所、2013年2月)を発行した。なお、このシンポジウムの様子は、以下の2つのURLでUSTREAM動画を公開している。

http://www.ustream.tv/recorded/29319356

http://www.ustream.tv/recorded/29322434

８．本年度の共同利用・共同研究の参加状況

区　　分	機関数	受入人数			延べ人数		
			外国人	大学院生		外国人	大学院生
学内（法人内）	2	4	1		18	4	
国立大学	2	2			2		
公立大学							
私立大学	2	3			11		
大学共同利用機関法人							
独立行政法人等公的研究機関	1	1			1		
民間機関	1	1			1		
外国機関							
その他							
計	8	10	1		33		

研究参加者の所属機関数、参加人数、延べ人数を区分に応じて記入して下さい。

※「学内」の所属機関数は「学部数」等を記入して下さい。

※参加人数及び延べ人数の算出方法は、以下の例に基づき算出して下さい。

（例）・１つの共同利用・共同研究課題で２人を共同研究員として３日間受け入れた（参加した場合）：参加人数２人、延べ人数６人

９．本年度 共同利用・共同研究を活用して発表された論文数

（参加研究者がファーストオーサーであるものを対象）

論文数	7	
うち国際学術誌に掲載された論文数	7 （　3 ）	0 （　0 ）

※下段の（　）内には、拠点外の研究者による成果（内数）を記載。

（注）分野の特性を踏まえて、参加研究者がファーストオーサーである場合の他に、コレスポンディングオーサーである場合や指導した大学院生がファーストオーサーになっている場合など、論文における重要な役割を果たした実績を示す必要がある場合は、その役割を明示の上で論文数を記載。

役割		
論文数		
うち国際学術誌に掲載された論文数	（　）	（　）

※下段の（　）内には、拠点外の研究者による成果（内数）を記載。

※　高いインパクトファクターを持つ雑誌等に掲載された場合、その雑誌名、掲載論文数、そのうち主なものを以下に記載。

※　拠点外の研究者については、発表者名にアンダーラインを付す。

掲載雑誌名	掲載論文数	主なもの	
		論文名	発表者名

（注）インパクトファクターを用いることが適当ではない分野等の場合は、以下に適切な指標とその理由を記載上で、掲載雑誌名等を記載。
拠点外の研究者については、発表者名にアンダーラインを付す。

インパクトファクター以外の指標とその理由		人文科学分野においてはインパクトファクターそのものの定義が困難であるが、学会誌ないし商業誌として信頼性と多くの読者を持つことで高い評価を得ているものに限定した。	
掲載雑誌名	掲載論文数	主なもの	
		論文名	発表者名
じんもんこん 2012	1	古典中国語形態素解析のための品詞体系再構築	山崎直樹・守岡知彦・安岡孝一
すべてをコンピュータの中に	3	漢字構造情報の RDF 化の試み	守岡知彦
東洋学へのコンピュータ利用	3	多粒度漢字構造情報のための包摂規準機械可読化の試み	守岡知彦