異種データベース統合に基づく 研究者情報検索

2007年8月28日(火) 高久 雅生(Masao Takaku) masao@nii.ac.jp



新領域融合研究センター

Transdisciplinary Research Integration Center



自己紹介

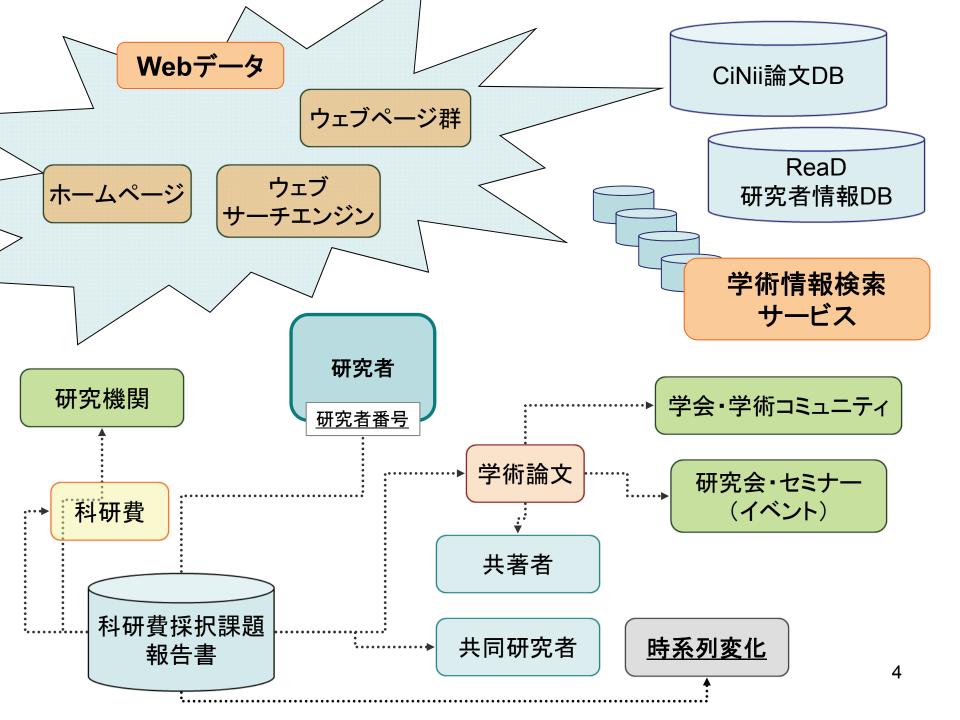
- 情報科学を専攻
- 大量データ情報構造の把握、提供システムに興味
 - (Japan/MARC, NDC, 用語体系)
 - NTCIR-WEB (2004-2005)
 - 学術文献データベース (2006-)
- 現在は、分野を越えた学術情報の提供、学術動向の把握のためのシステムの研究開発に従事
 - ポスドク研究員
 - 国立情報学研究所(NII)に勤務



目次

異種データベース統合に基づく研究者情報検索

- 背景、概略
 - 研究者情報とは?
- 対象データ
- ・研究者情報サーバ
- 考察、応用可能性
- ・まとめ



背景

- 研究成果
 - 研究→学術論文→出版→流通 (学術コミュニケーション)
- 学術研究の発展と論文データベース
 - (研究の現場)いち早く他の研究者による報告内容を知りたい
 - (社会学) 研究理論や学問研究の発展の様子を知りたい
 - (社会) 研究成果の公開、社会的応用可能性を知りたい
- 出版プロセス自体の電子化ともあいまって、論文データ ベースは不可欠のツールに

学術論文の例

論文

情報知識学会誌 Vol.14, No.1

掲載誌

Web サービスによる用語体系データの提供と その応用システム タイトル

Providing Data of Terminological Systems
Based on Web Services and Its Application Systems

高久 雅生 *,† 江草 由佳 * 石塚 英弘 ‡

著者

Masao TAKAKU Yuka EGUSA and Hidehiro ISHIZUKA

本研究では、用語体系の再利用性の向上と効率的な提供を目指して Web サービス (Web Services) の枠組みを利用した用語体系提供システムを構築した。さらに、この用語体系提供 Web サービスをバックエンドとする 2 つの応用システムを構築し、それによって、その適用 可能性を示した。2 つの応用システムは、用語体系のブラウジングシステムと、Google による検索 Web サービスと連携するシステムである。用語体系提供 Web サービスは、それぞれ 一つのサービスが一つの用語体系に対応することとし、見出し語検索と見出し語が持つ構造の取得を行う 2 つのインタフェースをもつ。用語体系を Web サービスの枠組みに基づいて提供することにより、アクセス方式が統一でき、用語体系を提供する側、利用する側が相互に分散的に利用することが可能になった。

We developed a providing service system of a terminological system, using a framework of Web Services, since we aims at improvement in the reusability and at efficient providing of a terminological system. The following two applications are successfully developed with the providing service system as a back-end service; one is a browsing system of several terminological systems and the other is a cooperating system with Web service of Google: search engine of Web pages. We used EDR technical term dictionary and a hierarchical structure data by Open Directory Project, as typical terminological systems. One service corresponds to one terminological system in our implementation, and has two interfaces; one searches word entries and the other acquires the structure of the entry. Various

内容 キーワード

論文検索サービス: CiNii



■収録データベース: NDL NII-ELS

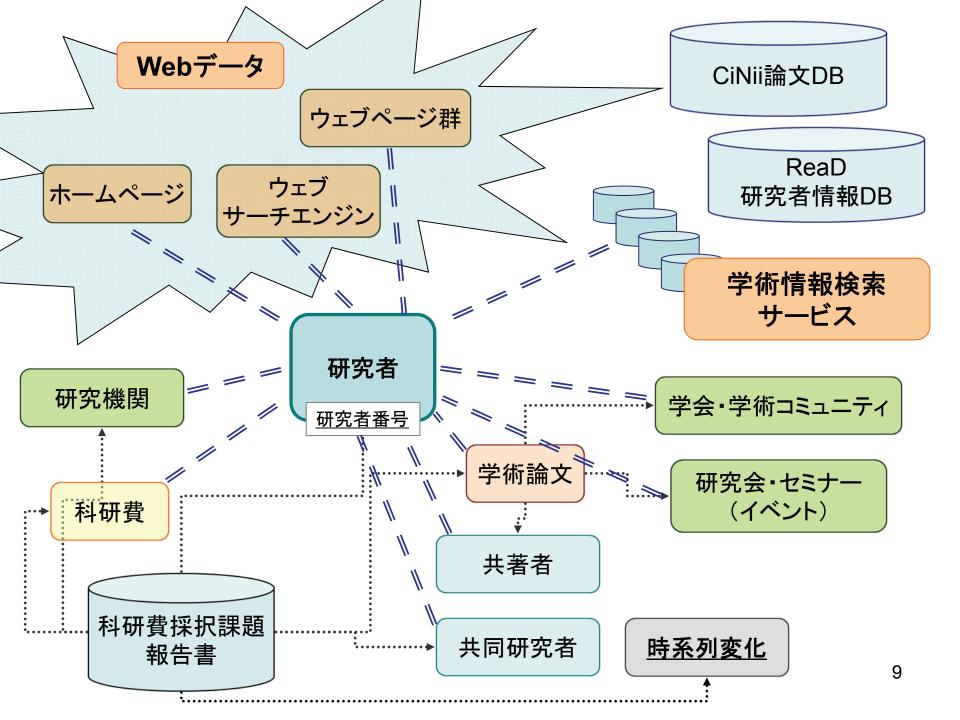
■ 抄録あり

■本文: ☐ CiNii @ D

http://ci.nii.ac.jp/

背景(2)

- 研究助成金
 - 競争的資金(研究費、外部資金)を獲得
 - 大規模実験
 - 研究成果の報告
- 学術サービスの高度化
 - (より多くの情報へのアクセスへ)
 - 論文本文
 - 「研究者」情報



研究者情報サービス

- 研究者の情報を提供
 - 出版物(論文や書籍)、研究分野、その他関連する情報
- 利用者、利用用途
 - 自身:履歴情報のレポジトリ的な使い方(CV)
 - その他: 研究動向、共同研究、専門家探し、評価
 - 査読者、講演者、専門委員...

今日のお話:

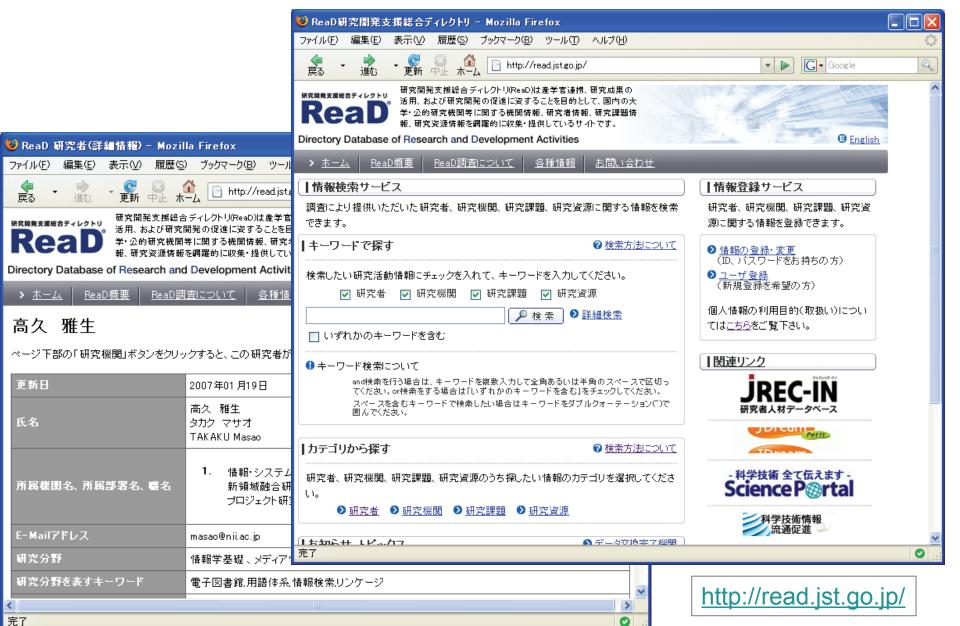
- 「科研費データベース」の内容を元にした研究者情報サーバ
 - 研究者番号をキーにして再編成

既存の研究者情報関連サービス

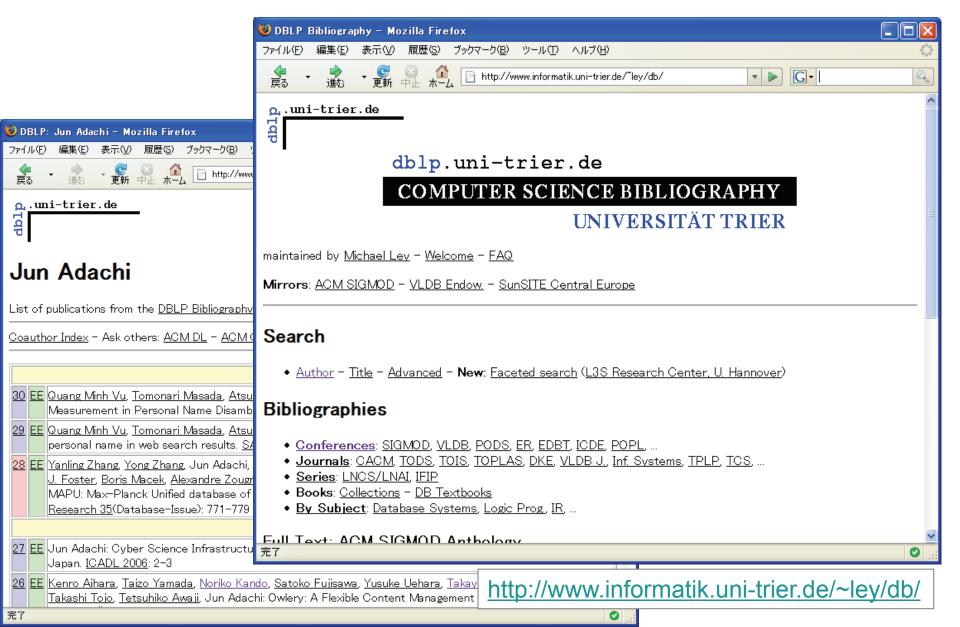
- 関連サービス
 - ReaD
 - 科学技術振興機構(JST)が運営
 - 研究者情報の収集と活用(本人による登録)
 - DBLP
 - 計算機科学分野の書誌情報DB
 - 発表文献、共著情報の提供

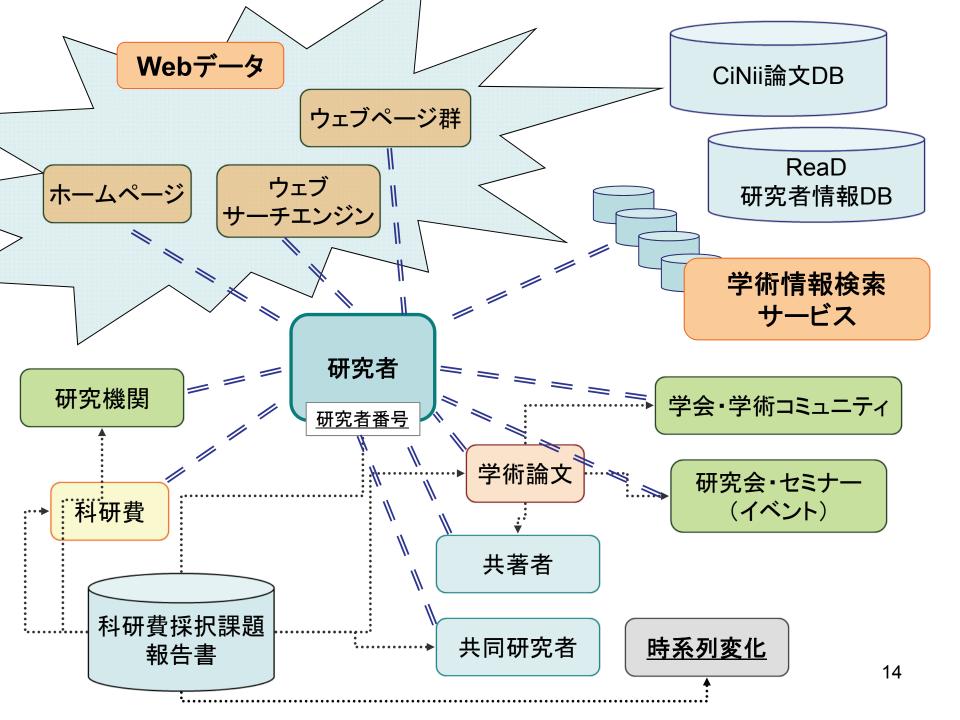


ReaDデータベース



DBLP国際会議データベース





科研費 (科学研究費補助金)

- 日本最大の公的研究助成(文部科学省所管)
 - 年間総額・約1,900億円
 - 年間研究課題数·約55,000件
- 主に学術研究機関に所属する研究者に対して補助 金を支出
 - 研究者個々の研究計画を分野ごとの専門家による審査
 - 人文系から自然科学系まで全学問分野を対象として、年間 数十万円~数億円規模まで、多様な研究種目、応募枠

科研費の例: 自身の例

一 右 ナ	研究	な肥入してください。 A · (B)	1				接整理番号	
亦	分 野	分科		組 日		細目	番号(4ケタ)	分割番号
45	総合領域	情報学	研究分 里		ティベース		1004	(A)- B
•					総合・	新聞地名	のみ(「作成・記	入腰領」2.を参照)
	*9 # \$	A44 21						30 ph
研究代表者	氏 名	氏	名	ED (年 齢 H18.4.1 現	在)	S. 51年	1 月生まれ
	消滅研究機関 部局・職	题所	情報学資源 クト研究』		位 博士 報学)	il il	性の専門	 「
研究課題	WWW 上・ の研究	サイトマップ	研究課	題的	N型用語	体系提	供システム	171-h
	年 度	研究経費	使	H	内 訳	(千 円)
研究経費	** BL	(千円)	設備備品費	消耗品費	籐	茸	謝金等	その他
	平成18年度	948	200	58	38	100	60	0
	平成19年度	1,110	研究予	算	0	400	60	0
千円未開の 無数は切り	平成20年度	980	450		10	400	60	30
		3.038	1.300	60	98	900	180	30

科研費データベース

- http://seika.nii.ac.jp/
- 国立情報学研究所が提供
- 科研費採択課題、報告書の情報を提供

- 科研費研究者番号
 - 国内研究者を一意に同定するためのIDキーとして有望
 - 基本的に1人に1つの番号が付与
 - 所属を移っても有効
 - 大規模

課題の情報(科研費DB)

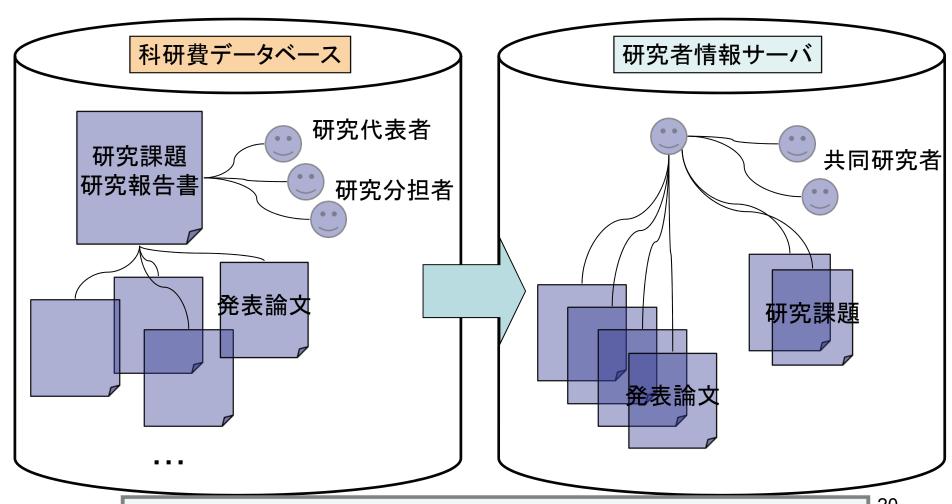


研究者情報

- 履歴的情報
 - 氏名
 - 年齡
 - メールアドレス
 - 住所
 - 所属機関
 - 職位
 - 研究分野
 - 研究歴
 - 学歴
 - 学位
 - 研究業績(出版物)
 - ホームページ

- [活用方法]
- ・ 本人が利用
 - 広報·普及活動
 - 応募・申請書類に記載
- 専門家探し
 - 講演者
 - 査読者
 - 有識者 etc.
- 人事•評価担当者
 - 本人の経歴を確認
 - 業務活動の情報源

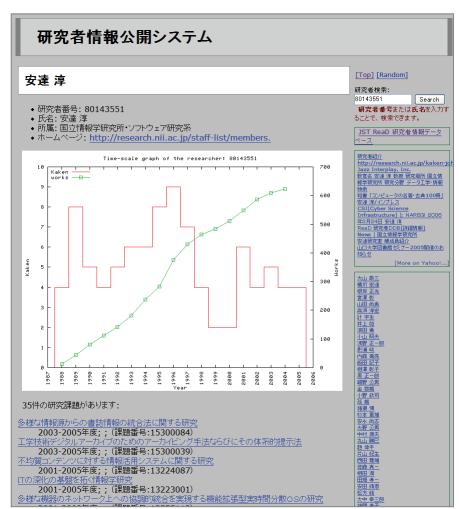
科研費データベースから 研究者情報サーバへ

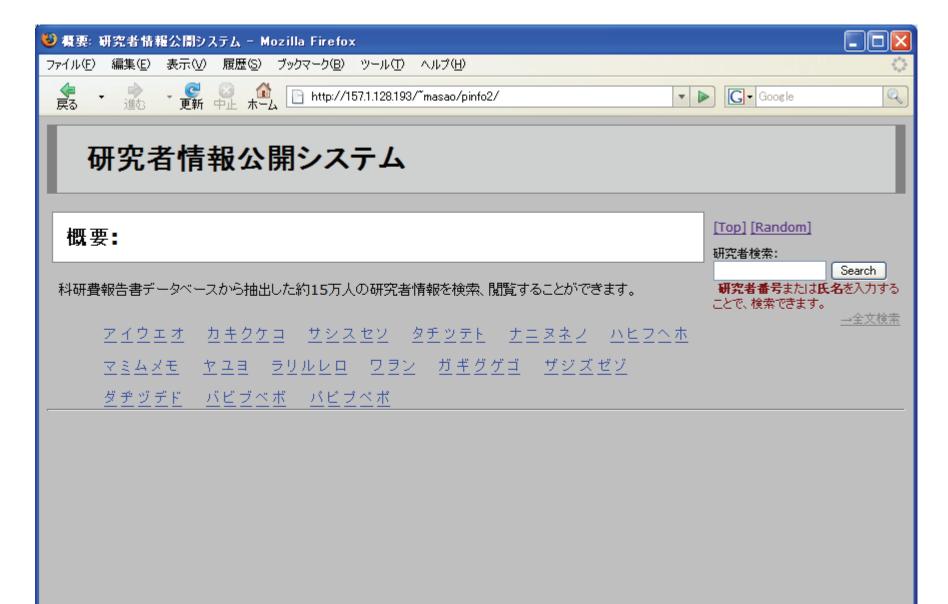


20

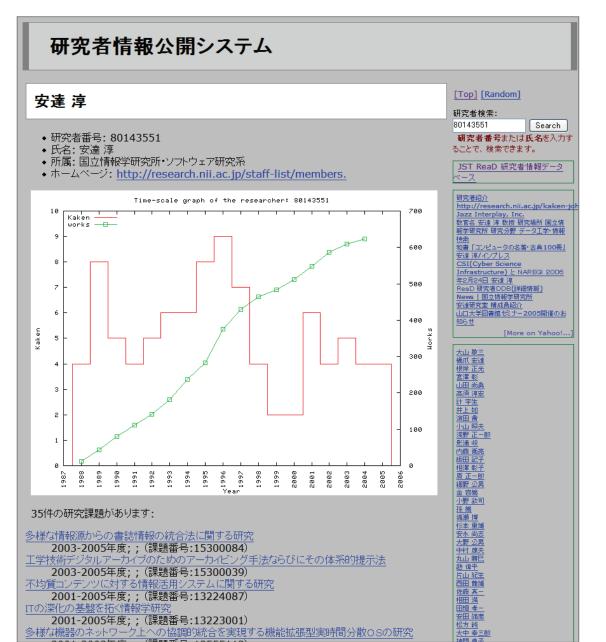
研究者情報サーバ

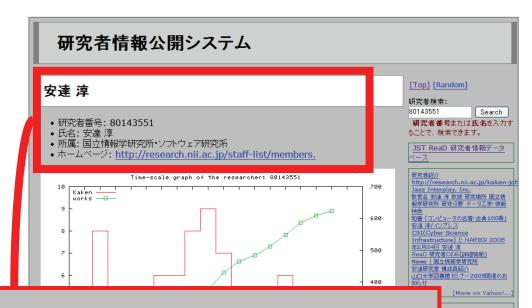
- 元データ(科研費DB)
 - (科研費採択課題) 247,745件
 - 主に1989年度から2004年度報告分まで
 - (研究代表者·分担者) 133,067名分
 - 個別研究者情報
 - (報告書記載文献数)1,896,070件(重複あり)
- 研究者の基本情報
 - 氏名、所属、職位
- 採択課題情報
- 関連する発表文献情報
 - 報告書記載分
- Web情報源とのリンク
 - ReaDデータベース
 - Yahoo!
- 経年的な活動の様子を可視化
 - 採択課題·関連発表論文数
- 共同研究者ネットワークを可視化
 - 過去の研究課題における共同研究者
 - GoogleMaps連携











安達 淳

- 研究者番号: 80143551
- 氏名:安達 淳
- ◆ 所属: 国立情報学研究所・情報学資源研究センター・教授, センター長
- ホームペーン: http://research.nii.ac.jp/staff-list/members.

安達 淳

◆ 研究者宙っ・

研究者情報公開システム

[Top] [Random]

研究者検索:

80143551

Search

[More on Yahoo...]

研究者番号または**氏名**を入力すること で、検索できます。

JST ReaD 研究者情報データベース

研究者紹介

Amazon.co.jp: コンピュータの名著・古典

100冊: 本: 石田 晴久,青山 幹雄,安達 淳,塩

田 紳二,....

______ Amazon.co.jp:コンピュータの名著・古典100

<u>册:本</u>

オンライン書店ビーケーワン:安達 淳

安達淳 作品一覧:紀伊國屋書店BookWeb

情報基盤の整備をどうすすめるか

2001年 12月18日 国立情報学研究所 安達

安達淳とは - はてな

Contact

研究者紹介

• 氏名: 安達 淳 ことで、検索できます。 所属: 国立情報学研究所・ソフトウェア研究。 JST ReaD 研究者情報データ • ホームページ: http://research.nii.ac.jp/staff-list/men Kaken works -金 容鰻 小野 欽司 35件の研究課題があります: 多様な情報源からの書誌情報の統合法に関する研究 2003-2005年度;;(課題番号:15300084) L学技術デジタルアーカイブのためのアーカイビング手法ならびにその体系的提示法 **2003-2005年度; ; (課題番号:15300039)** 不均質コンテンツに対する情報活用システムに関する研究 2001-2005年度;;(課題番号:13224087) ITの深化の基盤を拓く情報学研究 2001-2005年度;;(課題番号:13223001) 多様な機器のネットワーク上への協調的統合を実現する機能拡張型実時間分散OSの研究 26

Top] [Random]

研究者番号または氏名を入力す

Search

研究者情報公開システム			
安達 淳	[Top] [Random] 研究者検索:		
 研究者番号: 80143551 氏名: 安達 淳 所属: 国立情報学研究所・情報学資源研究センター・教授, センター長 ホームページ: http://research.nii.ac.jp/staff-list/members. 	80143551 Search 研究者番号または氏名を入力すること で、検索できます。 JST ReaD 研究者情期データベース 研究者総介		

145件の関連発表文献があります:

- 9th International Online Information Meeting. (1985)
- A. Matsumura, A. Takasu, J. Adachi: "Structured Index at IREX"Proceeding of the IREX Workshop. 57-60 (1999)
- A. Matumura, J. Adachi, A. Takasu: "Structured Index System at NTCIRI"Proceedings of the 1st NTCIR Workshop. 117-122 (1999)
- ADACHI, Jun: "Dissemination of Japanese Academic Journal over the Internet" The Proceedings of ISDL'97.
 32-35 (1997)
- ADACHI, Jun.: Fourth U.S. Japan Conference on Libraries and Information Scienece in Higher Education, Racine, Wisconsin. (1988)
- Adachi, Jun: "Dissemination of Japanese Academic Journals over the Internet" Proceeding of ISDL'97. 32-35
 (1997)
- · Akiko Aizawa: "A Co-evolutionary Framework for Clustering in Information Retrieval Systems"Proceedings of



研究者情報公開システム

12件の研究課題があります:

不均質コンテンツに対する情報活用システムに関する研究

2001-2005年度; 特定領域研究(C); (課題番号:13224087)

インデクス構造化による多言語情報検索の実現

1999-2001年度; 基盤研究(C); (課題番号:11680432)

分散情報ネットワークにおける情報資源発見機構の開発研究

1996-1997年度; 基盤研究(B); (課題番号:08458091)

異種分散型データベース用アクセスプロトコルの開発と実装

1993-1995年度; 試験研究(B); (課題番号:05558039)

電子図書館における原文書デリバリーシステムの開発

1991-1993年度; 一般研究(B); (課題番号:03452292)

```
1989-1990年度; 一般研究(C); (課題番号:01580038)
- パーコンピュ-タの高度利用に関する総合的研究
1988-1989年度; 総合研究(A); (課題番号:63302063)
結合型回書館ネッケランにおける統合業務ソステムの研究
1988-1989年度; 総合研究(A); (課題番号:63301102)
学的媒体を用いたデータベース・ジステムに関する研究
1987-1987年度; 一般研究(A); (課題番号:62410015)
格情報ネットワークにおける密結合型の書館システムの書機能の高度化とその応用
1986-1987年度・総合研究(A)・(課題番号-61301084)
```

1985年度; 試験研究; (課題等 0880005)

ウラに対ける。自然研究、(課題番号:60810009)

145件の関連発表文献があります:

- 9th International Online Information Meeting. (1985)
- A. Matsumura, A. Takasu, J. Adachi: "Structured Index at IREX"Proceeding of the IREX Workshop. 57-60 (1999)
- A. Matumura, J. Adachi, A. Takasu: "Structured Index System at NTCIRI"Proceedings of the 1st NTCIR

 Workshop, 117-122 (1990)
- ADACHI, Jun: "Dissemination of Japanese Academic Journal over the Internet" The Proceedings of ISDL'9 32-35 (1997)

Adachi, Jun: "Dissemination of Japanese Academic Journals over the Internet" Proceeding of ISDL'97. 32-35

- ADACHI, Jun.: Fourth U.S. Japan Conference on Libraries and Information Scienece in Higher Education, Racine, Wisconsin. (1988)
- (1997)
 Akiko Aizawa: "A Co-evolutionary Framework for Clustering in Information Retrieval Systems" Proceedings of

研究者情報公開システム

密結合型図書館ネットワークにおける統合業務システムの研究

[Top] [Random]

研究者検索:

九伯牧林。

研究者番号または氏名を入力することで、検索できます。

Search

- ◆ 1988 研究実績報告
- ◆ 課題番号: 63301102
- 1988-1989年度:総合研究(A)
- ◆ 研究代表者: 猪瀬 博 (学術情報センター・所長)
- ◆ キーワード: 学術情報システム / 目録所在情報サービス / データベース / LAN / DBMS / 図書館機械化
- ◆ 共同研究者:
 - ◇ 井上 如: 学術情報センター・研究開発部・教授
 - ◇ 上田 修一: 慶應大学・文学部・助教授
 - ◇ 山本 毅雄: 図書館情報大学・図書館情報学部・教授
 - ◇ 安達 淳: 学術情報センター・研究開発部・助教授
 - ∘ 根岸 正光: 学術情報センター・研究開発部・教授
- ◆ 概要:

本研究は、ネットワーク分散型の図書館システムをプロトタイプに基づき検討して、「密結合型統合図書館業務システム」を、学術情報システムにおける大学図書館向けサービスの今後の展開の方向として示そうとするプロジェクトである。本年度は、2カ年にわたる研究計画の第1年次であるので、ネットワーク分散型システムの実験に要する機器類を整備するとともに、システムの利用者である個別図書館からみた、この種のシステムに対する具体的、実務的要求およびその現実的適用性の分析と、こうしたシステムの実現に関わる技術的動向と可能性の評価という、いわば制度論と技術論の両面からの検討を行った。

(1)学術情報センターの目録所在情報サービスやその他の電質化システム既導入図書館(本館・分館)に対して、アンケー・調査を実施し、310件の回答を得た。その機器構成調査によれば、学内共用電質機、図書館専用機、パソコンの3者が3分の1ずつであり、従って、これらそれぞれに適したシステムを構成する必要があることなどがわかる。また、図書館への訪問調査の過程で、学術情報センター・システムにおけるデータベース蓄積の進行を背景として、これを利用した実務経験の中から、受入と目録業務など、図書館業務間の新たな関係の提案なども出されてきており、検討に値する。

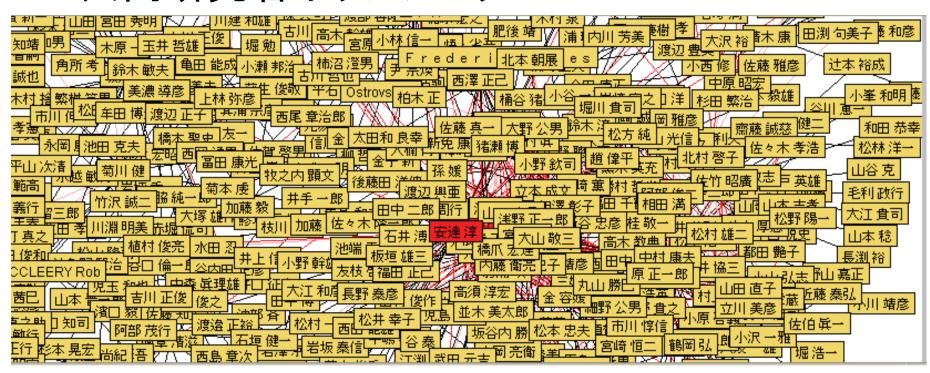
(2)大学図書館の業務システムについては、昨今の大学における情報化や全学LANの導入という潮流を踏まえて、新世代のシステムを提案するものとした。この点については、LANにおける分散処理方式とワークステーションの普及・低廉化・高機能化を考慮する必要がある一方で、図書館業務に要する大型データベース処理に向けて、DBMSの分散方式やMMLの利用可能性についても検討する必要がある。このため、当面2つのシステム・モデルを作成して、その評価等を実施しつつある。

発表文献

- ADACHI, Jun.: Fourth U.S. Japan Conference on Libraries and Information Scienece in Higher Education, Racine, Wisconsin. (1988)
- ◆ 内藤衛亮: 学術情報センタ-紀要. 2. (1989)
- NEGISHI, Masamitsu.: Online '88 Conference Proceedings, New York, Online Inc. 114-118 (1989)
- ◆ 猪瀬博: 知識情報の世界を拓<. 8-17 (1988)



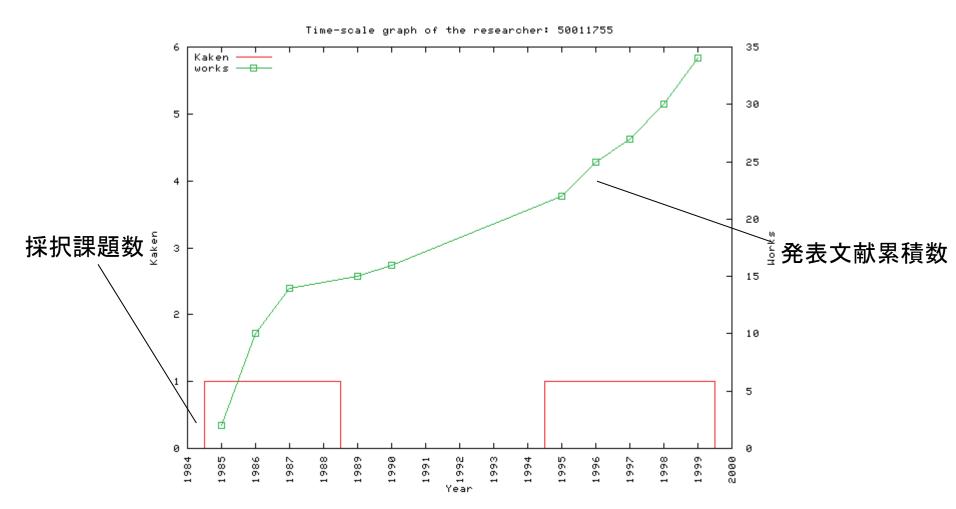
• 共同研究者ネットワーク



考察

- 可視化
 - 時系列での研究履歴
- ・ 文献の同定
 - 論文データベースにリンクを張りたい
- 研究者のネットワーク
 - 共同研究の関連研究者

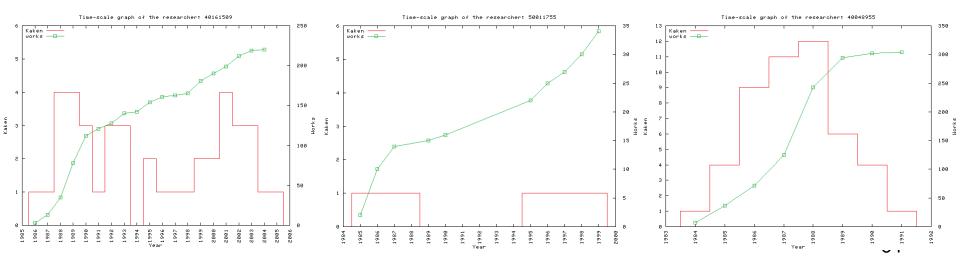
時系列情報の可視化(1)



33

時系列情報の可視化(2)

・採択課題と発表文献には、様々なパターンが 見られる(把握が容易)



時系列情報の可視化(3)

- 可視化
 - 時系列的な推移の可視化
 - 類型的なパターンの発見
- どういった情報を可視化するのがよいか?
 - 予算
- ・隠れた発表
 - 科研以外での発表 or 報告書に書かない発表
- ・ 関連発表文献の正確性=???
 - 対象を広げていく必要性
 - 外部データベース(CiNii?)とのリンケージ
 - → 文献同定、論文データベースとの統合

考察: 文献の同定

- ・ 論文データベースCiNiiや、書籍目録データベースWebcatとの間で文献同定
 - 論文同定サーバ (by 相澤教授@NII)
- 表記ゆれ等の問題から科研費報告書記載の 論文情報はそのままでは論文データベース のレコードとマッチングできない
 - 科研費報告書は紙ベースでの報告であるため
 - 人手による参照方式はまちまち

文献同定: KAKEN-CiNii間の場合

- 科研費報告書記載文献は、年度単位での報告なため、誤りがあることもある
 - ノイズ
 - •「(印刷中)」、「(発表予定)」
 - 個々の研究課題ごとの研究者の文献参照方式 の差異
 - 「Vol.15, No.3, 2005, pp.5-25」⇔「15(3)5-25, 2005」
 - ・タイトルを書かない(理系分野の一部)

論文同定サーバ(by 相澤@NII)

- 特徴語抽出+編集距離
- 元データベースレコード とのマッチング
 - CiNii
 - Webcat
 - Crossref



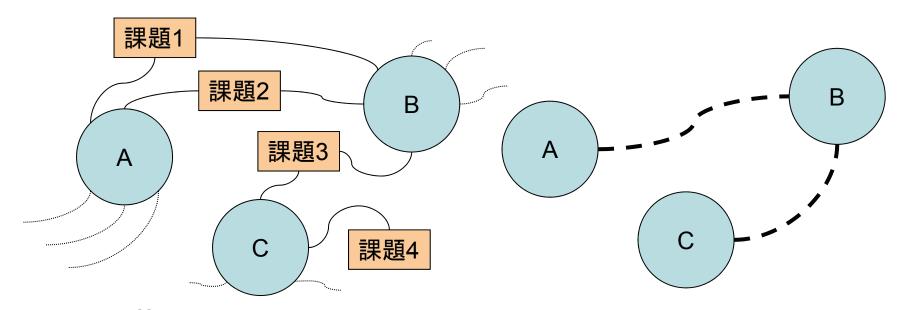
その他情報源のリンケージ

- Web情報とのリンケージ
 - 簡易: ReaDデータベース、Yahoo!
 - 研究者情報に特化したリンケージ手法
 - Wangs
 - 研究者情報に広く関連する他のエンティティ抽出
 - イベント、論文

データベース統合

- 研究者情報サーバと論文データベースの統合
 - 「名寄せ」
 - 同姓同名問題への対応
 - 人手判定

考察:研究者ネットワーク



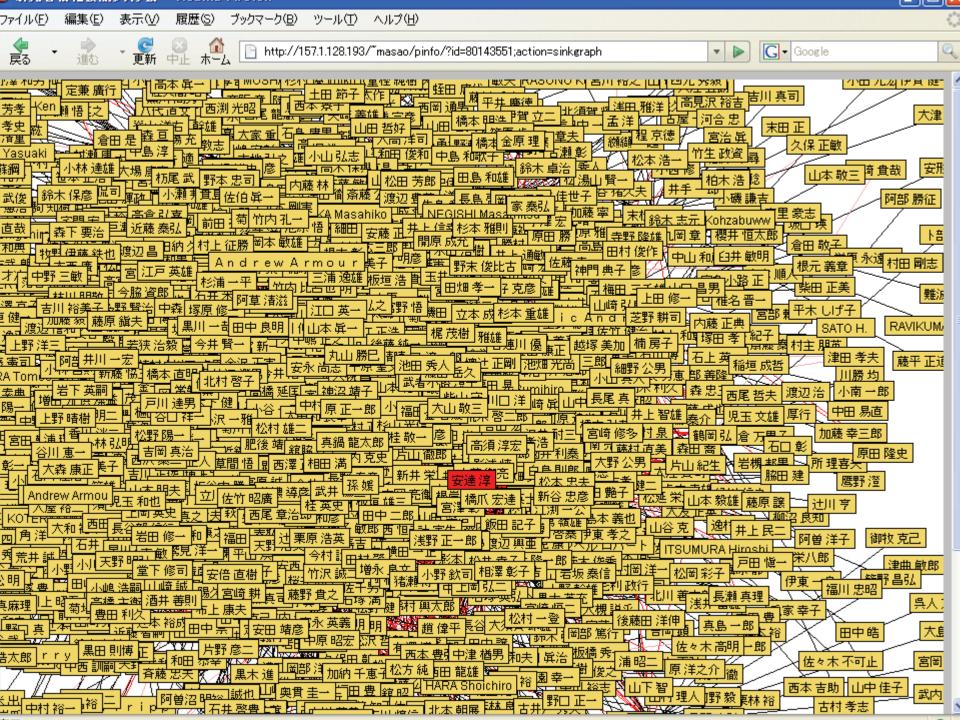
- ・ 科研費の場合
 - ある研究課題に一緒に参画した経験(つながり)
 - 研究代表者 共同研究者

共同研究ネットワークの可視化

- 特定の研究者からたどれるネットワーク
- ・ 2ホップ先までの研究者を対象とした表示
- エッジの距離
 - 研究者ペアごとの参加プロジェクト数に反比例した距離を設定

• 課題

- 多くの研究者とつながりを持つ研究者の場合
 - 一度に視認可能な量を超えてしまう限界
- インタラクティブなやりとりができない

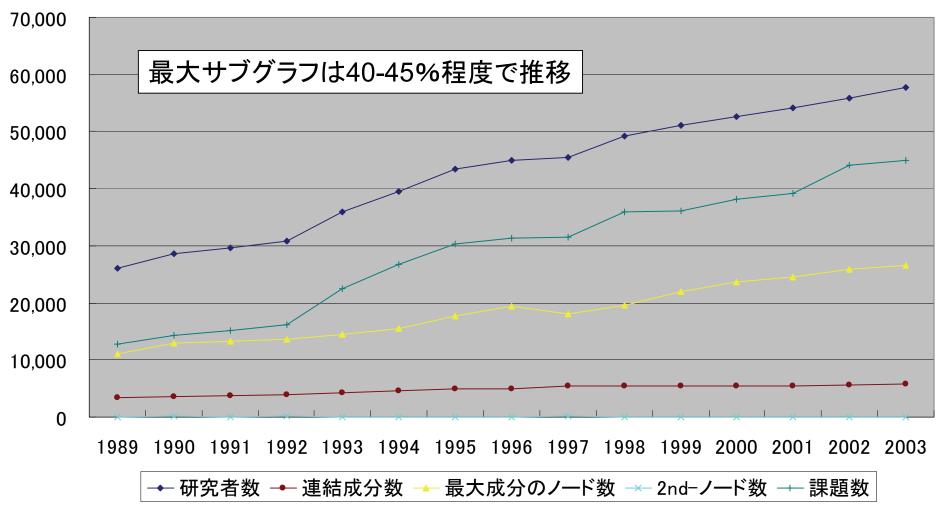


グラフサイズによる研究者ネットワークの 分析

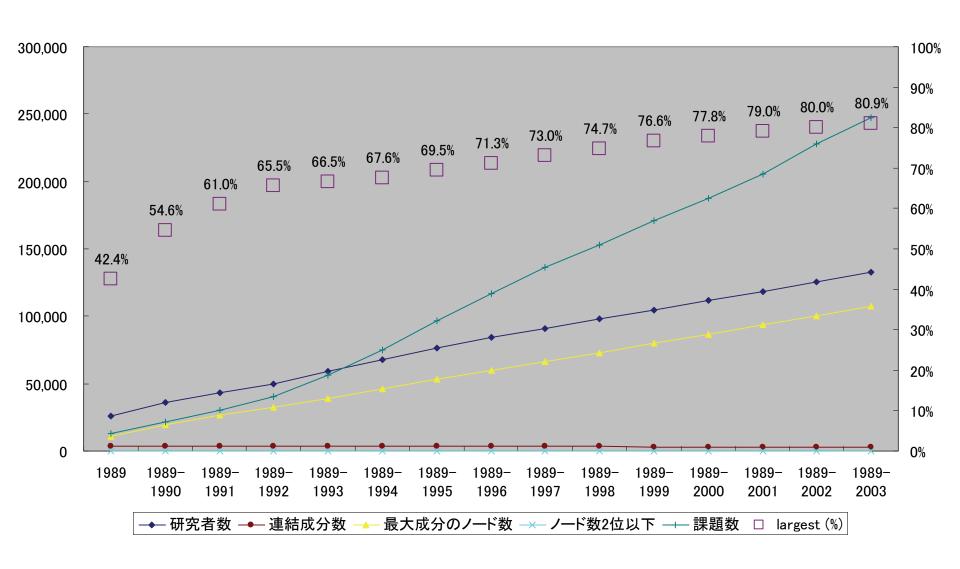
- 連結成分:ネットワークの広が りによる連結の範囲
- 全体のうち多くの研究者 (81.5%)が連結成分となるようなグラフを構成
 - Small-world !?
 - 推測(仮説)
 - 様々な種目があるために、分野を またいだ共同研究が行われやす い側面がある?
 - 競争的資金特有の性質か?
 - → 要分析

研究者数	<u>133,067</u>
サブグラフ数	2,911
1	108,485
2	34
3	31
4	30
5	26
5	26
7	21
8	20
9	20
10	19

各年度における連結成分のサイズ



グラフサイズの年度累計推移



小規模サブネットワークの分析

研究者総数: 133,067

<u>九日</u> 1	応致. 133,007
1	108,485
2	34
3	31
4	30
5	26
5	26

和歌山県立医科大 山梨医科大				
	日本歯科大学			
	愛知医科大学			
	岐阜大学			
	自治医科大学			

7	21
8	20
8	20
10	19
10	19
10	19

	東京慈恵会医科大		
)	長崎大(歯学)		
)	佐賀医科大		
)	島根医科大		
)	東京慈恵会医科大		
)	北里大 広島大 厚生連尾道総合病院		

• 考察

- グラフサイズ2位以降のものを確認したところ、同一大学医学部内の所属研究者のネットワークがいくつか見られる
 - 医学部特有の研究体制が影響か?(白い巨塔??)

共同研究者ネットワーク(今後の課題)

- 共同研究者ネットワークのモデル
 - 1. 同一課題に関わった者同士の関係
 - 2. 同一課題における代表者・分担者の関係
 - → 両モデルでの分析を行ってみる
- ・ 年代、種目、重複申請などによる制限
- ネットワーク特性の分析
 - 中心性、集中度
 - Hub / Authority
 - 分野ごとの分析、年度ごと(時系列)の分析
 - 種目別の分析

まとめ

- 科研費DBの研究者情報をまとめて提示
 - 採択課題と関連発表実績との関連を可視化
 - Web情報源(ReaD, Yahoo!)とのリンケージ、連携
 - 論文データベースへのリンク
 - 共同研究ネットワーク
- 今後の課題
 - 頑健なリンケージエンジン
 - 「エンティティ」Web検索技術
 - 研究者•専門家検索