

# プロジェクト名: 機能と帰納: 情報化時代にめざす科学的推論の形

プロジェクトディレクター: 樋口 知之

## 1. 研究目標

情報社会の実現によって様々な分野で複雑なシステムに関する大量データに基づく予測と、そのリスク評価の方法の確立が重要な社会的課題となっている。本プロジェクトでは、地球、生命、社会等の4研究所の融合分野において戦略的研究を推進しながら、複雑なシステムの理解のための、帰納的手法、あるいは帰納的手法と演繹的手法との融合的手法による、システムの機能のモデル化に関する研究開発を行う。ここで機能のモデル化とは、対象そのものを実体的に精緻にモデル化するのでなく、対象に関する情報の入出力関係に代表されるような、機能自体を模倣する数理モデルを構築することを意味する。これにより、統計的モデル構築法と予測アルゴリズム、情報抽出・知識発見のための情報統合の方法など、分野に共通の道具を生み出すことを研究の目的とする。

またこれらのモデリングの方法を基盤として、リスクの(科学的)評価と管理のための方法論の確立をすすめる。そして、情報とシステムという視点から不確実性に関わる研究の新分野を開拓し、現代社会が直面する諸問題の解決を通じて、安心・安全な国家社会の構築、地球環境の改善など持続的な繁栄を目指す。

## 2. 研究概要

本プロジェクトでは、地球、生命、社会等の4研究所の融合分野において、統計的モデル構築法と予測アルゴリズム、情報抽出・知識発見のための情報統合の方法など、分野に共通のツールを生み出すことを目標としている。

2.1 背景: 中抜きデータ処理

2.2 成果物: メタウェア

2.3 分野に共通のツールとは: 具体的用語

2.4 特徴1: 機能のモデル化

2.5 特徴2: Personalization とベイジアンモデリング

### 2.1 背景: 中抜きデータ処理

今、インターネットに代表されるIT革命により、人間がかかわるあらゆるシステムが大きく変容しつつある。例えば物流システムが理解しやすい。生産者と消費者双方が求める情報の即時的交換を可能にする総合情報サービスの登場により、生産者と消費者が直結された"中抜き"流通

が普通となりつつある（図1）。実は同じことが、一般に普遍的だと思われる、研究の方法（やり方）自体にも起きつつある。データの組織的超大量取得と、そこからの当該分野における有益な情報・知識の獲得、そして対象の制御能力の向上である。もちろんこの一連の作業には、物流システムと同じく機動性と自動化が強く求められる。従前はデータの質、量とも著しく不足していたため、新しい知見の獲得には、非常に大胆かつ奇抜なアイデアの投入が必要であった。そのため研究現場においては、自由な発想の重要性は深く認識されつつも、既存の知識に拘束された研究のすすめ方が多かったことも事実である。

このようなデータをとりまく環境の激変が、情報・システム研究機構の誕生をもたらしたといえよう。実際、機構設立の理念を説明した文（ホームページに掲載）の一部に、『情報・システム研究機構は、4研究所が連携することにより、生命、地球、環境、社会などに関わる複雑な問題を情報とシステムという立場から捉え、実験・観測による大量のデータの生成とデータベースの構築、情報の抽出とその活用法の開発などの課題に関して、分野の枠を超えて融合的な研究を通して、・・・』と示されている。

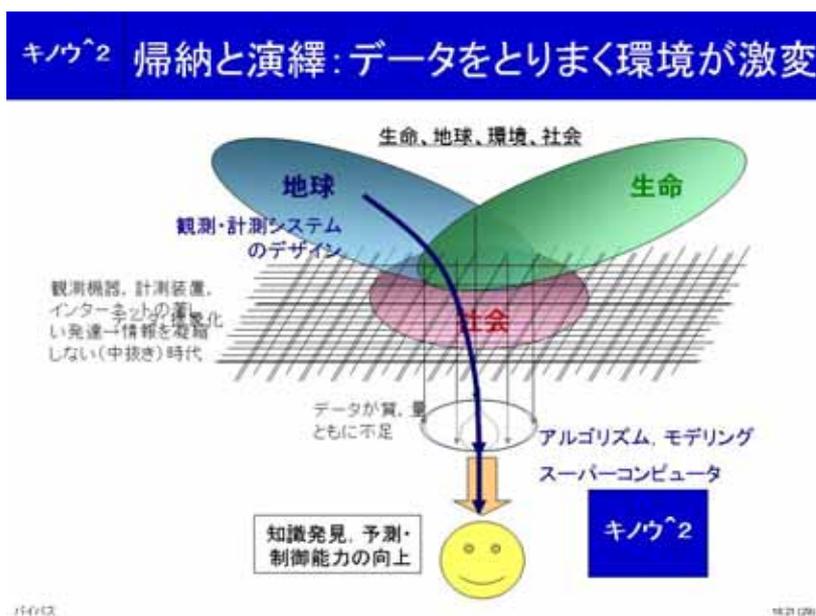


図. 1

## 2.2 成果物：メタウェア

本プロジェクトにおいても、地球（青：図1）、生命（緑）、社会（赤）の3分野に共通の---3原色を混ぜると白色光になる---、いわば無色透明とも結うべき性質の“道具”の研究開発が目標である。ここで道具とは、具体的にはアルゴリズム、モデリング技術、情報収集のデザインといったものである。いま我々はそれを、ハードウェア、ソフトウェアと対峙する、情報学をささえる第3の要素---メタウェア---と呼ぶことにする。近年、知財の様式がモノ主体から変化し、アイデア、ビジネスモデルといった無形物にシフトしているが、これらもメタウェアと位置づける。

## 2.3 分野に共通のツールとは：具体的用語

研究代表者(プロジェクトディレクター)のリーダーシップのもとにプロジェクトを機動的に推進するため、7つのサブ研究テーマを策定している。各サブテーマ名と、各々が関連する手法---分野に共通、あるいは他分野へ転用が平易なツール---を以下に示す。

-- 具体的なサブプロジェクト --

1. 予測と発見のためのモデリング技術とアルゴリズム開発  
**関連する手法**：データマイニング，テキストマイニング，Web マイニング，グラフモデリング，機械学習，ゲノム配列解析，バイオインフォマティクス，DNA アレイデータ解析手法
2. 統計的モデルに基づく地球科学における逆問題解析手法  
**関連する手法**：オーロラ画像解析，映像処理，画像合成，ベイジアンモデリング，オーロラ構造 3次元復元，地球環境長期予測，巨大次元の数値解法，時系列解析手法，画像解析手法，シミュレーション技術，時空間モデリング，データ同化
3. 計算機による帰納的モデリングのための環境  
**関連する手法**：グリッド計算技術，乱数に関する総合研究（物理乱数，擬似乱数），並列計算技術，統計ソフトウェアの共通基盤化，諸ゲノム解析手法の R への実装，数値的最適化技法
4. マルチモーダルデータからの不変情報の発見とその方法論の研究  
**関連する手法**：音声・画像解析，音声・話者認識，自然言語処理，多言語処理，ロボティクス，対話技術
5. 複雑システム理解と予測のためのアレイデータの帰納的解析手法開発  
**関連する手法**：レーダー観測信号処理，諸ノイズ除去手法，異常値処理，時空間モデリング，三次元形状のモデル化とシミュレーション
6. 雑音と帯域制限を有する高速データシステムのモデル化と試作システムによる評価  
**関連する手法**：デジタル・アナログ信号処理技術の融合，マルチパス干渉，マルチユーザ干渉，超高速データのコヒーレント伝送技術
7. リスク評価のための帰納推論と機能的統計解析のデザイン  
**関連する手法**：金融・経済システムの研究，ファイナンス数理，保険数理，医薬品・食品安全性の研究，疫学，サンプリング調査法，環境リスク研究，災害リスク研究，極値分布の研究，公開鍵暗号の安全性解析，電子商取引の制度

## 2.4 特徴1:機能の帰納的手法によるモデル化

本プロジェクトにおいて生み出す分野共通のツールには、その開発原理に3つの言葉で語られる特徴がある。

## 本プロジェクトにおける研究開発原理

1. 機能のモデル化
2. 帰納的手法（帰納と演繹の融合）
3. Personalization

機能のモデル化とは、対象そのものを実体的に精緻にモデル化するのではなく、対象に関する情報の入出力関係に代表されるような、機能自体を模倣する数理モデルを構築することを意味する。例えば、近年の著しい発展があるロボティクスをささえるのも、機能のモデル化の概念である。例えば、悲しい顔を見た時に相手を元気づけるようなロボットを作りたいとする。その目的のために従前は、眼の生理的構造、そこからの信号伝達、脳の信号処理の理解、そして運動方程式に基づくロボットの制御など、すべての研究作業が、素過程の積み上げ方式、つまり演繹のプロセスのつなぎ合わせであった。残念ながらこのような研究アプローチでは目的の達成にはなかなかいたらない。



図. 2

ところが最近のロボティクスの主たる作業は、インプットデータとアウトプットデータの関係にのみ注目し、その機能（入出力関係）を近似する数理モデルの構築である。数理モデルに含まれる諸パラメータは統計的学習手法によって決定できる。粗過程の積み上げによる演繹的手法より、データにもとづく帰納的手法のほうが目的の達成には効率がよいことは、子供がニュートンの運動方程式を知らなくても自転車に乗れたり、また逆上がりできるようになる事実を想起すれば明らかであろう。

## 2.5 特徴2: Personalization とページアンモデリング

社会からの要請に目をむけると、無駄を省く（低価格化，低コスト）ために資源の有効利用，つまり資源利用の選択と集中化が焦眉の急である，また，価値観の多様化などを受け，“コ”（個人，個性，固有，個別）に特化したサービスが求められている．例えば，オーダーメイド医療，副作用の研究，テーラーメイド教育，マイクロマーケティング（One-to-One marketing, Situation marketing），環境に優しい製品などすべて“Personalization”という情報技術で概括できる，“コ”に特化したサービスあるいは製品である．21世紀は，20世紀の大量生産・大量消費をめざした科学から，個人に焦点をあわせる科学へ確実にシフトしつつある．

では Personalization を達成するために必要な技術は何であろうか．図3左上に示された表のように，さまざまな医療診断項目（横項目）に対する結果（数値，カテゴリーデータ）の各個人データが縦に大量に並べられたデータセットを考えてみる．大量のデータが得られるようになったとしても，健康な人とそうでない人は自ずと検査項目が異なり，その結果，すべての項目にデータがある場合は極めてまれである．ある病気の因子を探ろうとして，条件をそろえる操作---Conditioning---，例えば身長，体重，中性脂肪値，... が似た値をもつ人の仲間で病気の発症に大きく影響する特徴量を探そうとしても，条件数が大きくなると似た値をもつケース数が著しく減少する．そもそも，すべての特徴量がそろっているデータなど希であるのが通常である．図3の上段右に示したように，地球上のさまざまな観測点で得られる地球物理データに対しては，どの観測地点においても必ず得られる，緯度，経度のような種類の特徴量でデータを並べ替えてみる．地理上近い距離にある観測点群間で他の特徴量を比較しようにも，観測された特徴量がそろっていない場合も多い．このようなデータの有限性，表でいえばごく一部分にしかデータが埋まっていないような，情報の欠損を前提として，“コ”に特化したサービスを実現しなければならない．

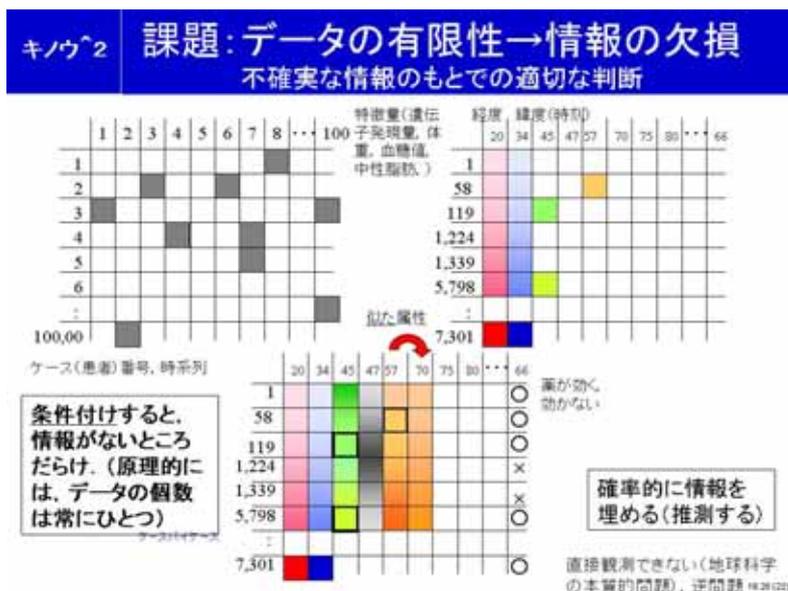


図.3

そのためには、ある特徴量（属性）で似た値をとるものは、他の特徴量でも似た値をとることが期待できるといったような先見の知識を活用することで、表でデータが抜けているところを埋めていく作業が必要である（図3下段）。それも決定論的に埋めるのではなく、確率的に情報を埋めていく。この作業---imputation---が終われば上述の条件付け(Conditioning)の作業を行っても、ケース数がゼロになることは避けられる。

この種の、手元にはない情報を先見情報を利用して確率的に埋めていく作業は、原理的に逆問題解法と同じである。有限の観測データから直接観測できない量を推測する、いわゆる逆問題は地球科学の本質的問題である。先端領域における発見のためには、その未知の部分に対する我々の期待を積極的にモデルの形で表現し、能動的に情報抽出を行うことが必要である。この枠組みは、ベイズ統計、あるいはベイズ的アプローチと総称され、そのさらなる特徴として対象の表現方法が階層的であることも挙げられる。これらの特徴により、異種の情報（マルチモーダルデータ）を自然に組み合わせることが可能なのである。

### 3. 年次計画

テーマ	16年度	17年度	18年度	19年度	20年度	21年度
	予備研究	プロジェクト初年度		中間評価		
研究体制の編成						
情報収集・整理						
研究会，チュートリアル，ワークショップの開催						
研究体制の見直し						

PDが統括する内容に関してのみ年次計画に記載した。

#### 平成16年度(予備研究)

平成16年度に新領域融合研究センター研究課題として採択された以下の研究プロジェクトを、17年度スタートした本プロジェクトは引き継いでいる。

- ・オーロラ科学における画像解析と逆問題

研究代表： 佐藤 夏雄（国立極地研究所）

- ・帰納機械による動的なマルチモーダル情報の検索と認知の研究

研究代表： 松井 知子（統計数理研究所）

- ・南極大型大気レーダーによる高級観測アルゴリズムと高速データ処理システムの開発

研究代表： 江尻 全機（国立極地研究所）

- ・磁気圏・電離圏・大気圏複合システムの定量的解析に向けた研究

研究代表： 樋口 知之（統計数理研究所）

- ・科学研究における計算機によるモデリング環境

研究代表： 中野 純司（統計数理研究所）

### **平成17年度**(プロジェクト開始)

統計的モデル構築法と予測アルゴリズム等分野に共通の道具を生み出すために、まずサブテーマを選定し集約的に効率よく研究をすすめる体制を整える。特に、予測と発見のためのモデリング技術とアルゴリズム開発、計算機による帰納的モデリングのための環境開発、マルチモーダルデータからの不変情報の発見とその方法、高速データシステムのモデル化技術、リスク解析とその評価技術など分野横断的な研究の情報収集、整理を網羅的に行う。

### **平成18年度**

スーパーコンピュータ上での並列化と（物理及び疑似）乱数の利用が比較的簡単な操作で行えるようなシステムの開発や、高速データ通信モデル化用の準備的なハードウェアを試作する。マルチモーダルデータに含まれる不変情報の発見や、大規模アレイデータからの構造抽出に関する、帰納的手法を用いた既存手法の整理をする。個別科学におけるリスク解析の現状とのギャップをはかるため、各分野で比較的小規模のワークショップを複数開催し分野間連携の準備を行う。

### **平成19年度**(中間評価)

データのコーディングモデル、アルゴリズムやモデルなどについて、分野横断的な考察を行う。また、分野間の交流と統計解析手法の水準のボトムアップを引き続いて図る。スーパーコンピュータ上での並列化と（物理及び疑似）乱数の利用が比較的簡単な操作で行えるようなシステムの開発を引き続いて行う。システムを R に限定せずにデータの可視化や解析を行うための研究も行う。マルチモーダルデータに含まれる不変情報の発見や、大規模アレイデータからの構造抽出に関する、帰納的手法を用いた既存手法の発展を図る。また、マルチパス、フェーディング環境下で高速・高性能を実現するための無線システムモデル化の研究を推進する。アレイ観測データの効率的なノイズリダクション法や、地球科学データのダイナミック逆問題解法の研究をすすめる。

### **平成20年度**

複数のスーパーコンピュータやパーソナルコンピュータが有機的に協力してモデリングを行えるような環境の研究を行う。帰納的メタウェア、データのコーディングモデルに関する考察結果の整理や、開発した観測アルゴリズムによる試験観測などを通して、帰納的手法の体系化を行う。リスク解析に関わる外国人客員の招聘等により、チュートリアルセミナーを開催することによって、研究者養成に資する。

## 平成21年度

機能と帰納での科学研究におけるモデリングを行う際の有用なツールとなるシステムの例示や、帰納的メタウェアやマシンのツール化、資料化を行う。開発した帰納的手法のインターネット等を通じた一般公開や、高速・高品質無線伝送システムモデルのフィールド評価を行う。またあわせてこれらの手法の他分野への適用について研究する。成果は適宜国内外の学会、及び論文にて発表する。高速・高品質無線伝送システムのモデルを確立し、国内外へ発表・提案・啓蒙活動を行う。アレイデータ観測実システムへの適用を国内外に対して提案し、成果発表・チュートリアル等を行っていく。過去4年間の研究成果を踏まえ、必要に応じて既存研究サブテーマの見直しと人員の再配置を行い、新規重点分野の開拓を行う。複雑なシステムの理解を加速する、アルゴリズムとモデリング技術のさらなる研究開発の推進により、安心・安全な国家社会の構築、地球環境の改善など国家及び人類の持続的な繁栄に貢献していきたい。

### 4. 平成17年度研究実施体制

研究代表者（プロジェクトディレクター）のリーダーシップのもとにプロジェクトを機動的に推進するため、7つのサブ研究テーマを策定した。機構内研究所および大学共同利用機関の特性を生かしたコミュニティから人選したサブ研究テーマ実行チームを構成している。

1. 予測と発見のためのモデリング技術とアルゴリズム開発  
サブテーマ代表：鷲尾隆（大阪大学 / 統計数理研究所）
2. 統計的モデルに基づく地球科学における逆問題解析手法  
サブテーマ代表：門倉 昭（国立極地研究所）
3. 計算機による帰納的モデリングのための環境  
サブテーマ代表：中野純司（統計数理研究所）
4. マルチモーダルデータからの不変情報の発見とその方法論の研究  
サブテーマ代表：松井知子（統計数理研究所）
5. 複雑システム理解と予測のためのアレイデータの帰納的解析手法開発  
サブテーマ代表：和田 誠（国立極地研究所）
6. 雑音と帯域制限を有する高速データシステムのモデル化と試作システムによる評価  
サブテーマ代表：瀧澤 由美（統計数理研究所）
7. リスク評価のための帰納推論と機能的統計解析のデザイン  
サブテーマ代表：椿広計（筑波大学 / 統計数理研究所）

図4に示したように、サブテーマは、その主として取り扱う応用分野に関連づけて意識的に並べられている。順番の若い方が生命科学を、次に地球科学、情報工学、・・・リスク解析、のように、番号が大きくなるにつれて人間の実生活に関連した社会科学の諸問題を取り扱うサブテーマが並ぶ編成とした。方法論の観点からすると、番号の小さいサブテーマは知識発見的要素が、

また後になるにつれてデザインの要素が強い性格を帯びた手法に関する研究内容となるように並べてある。また、最初と最後（サブテーマ1とサブテーマ7）に、重要な社会的課題に挑戦する戦略的研究組織である、統計数理研究所の戦略研究センター、予測発見戦略研究センターとリスク解析戦略研究センターを配置している。また、サブテーマの2と5は国立極地研の教員をサブテーマ代表とするプロジェクトである。機構内参加者数は55名、機構外参加者数を含めると総68名のプロジェクト体制である。参加した機構外の参加者の所属機関名は

大阪大学，早稲田大学，和歌山大学，京都大学，北海道大学，東京大学，地球環境フロンティア研究システム，長岡技術科学大学，岡山大学

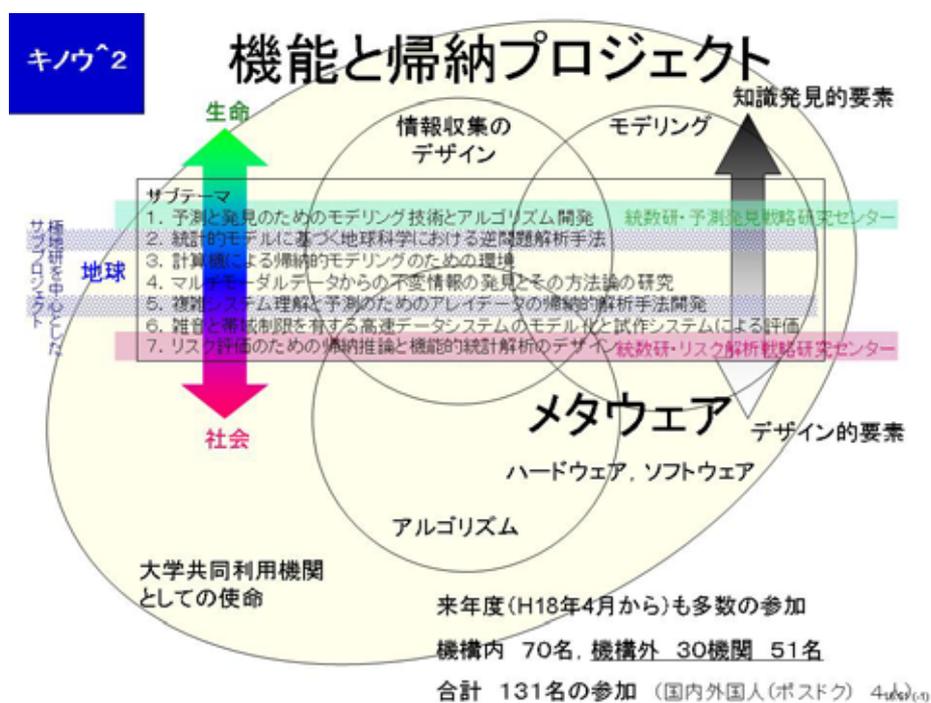


図.4

## 5. 平成17年度の研究進捗

複雑なシステムの理解のための、帰納的手法によるシステムの機能のモデル化に関する研究開発を、統計数理研究所と国立情報学研究所が連携しながら行った。国立遺伝学研究所と統計数理研究所の共同研究により、近年の統計科学周辺分野の新しい考えに基づいた解析手法とゲノム情報学分野における諸解析技術の融合研究を開始した。また国立極地研究所と統計数理研究所を核とした大学等の外部機関を含めた共同研究態勢により、地球科学における先進的解析手法の研究も進めた。地球、生命、社会等の4研究所の融合分野において、具体的には、地震、南極大気、オーロラ、成層圏、電離圏、地球環境、遺伝情報、画像、映像、音声、無線、金融、食品、医薬品等の幅広い具体的な対象の研究のために、統計的モデル構築法、予測アルゴリズム、情報抽出・知識発見のための情報統合の方法など、分野に共通のツールを生み出す目的のもと、各サブテ

マにおいて初期的な成果が得られ始めている．詳細は各サブテーマの成果の欄を参照されたい．

平成 17 年 10 月 7 日に，平成 17 年度の研究打ち合わせ及び事務連絡会を統計数理研究所にて開催した．実質的なプロジェクトメンバー全体の最初の顔合わせということもあり，7 つあるサブテーマ（第 1 サブテーマはさらに 4 つのサブサブテーマに分かれている）の各サブテーマのディレクターから，研究の背景，ねらいを中心に概要と現在の進行状況の説明をしてもらった．

今年度はプロジェクトの初年度ということもあり，プロジェクトディレクターは全体の広報宣伝活動に力を注いだ．プロジェクトのねらいなどを分かりやすくまとめたパンフおよびポスターを作成し，宣伝用に活用してもらうよう各サブテーマの代表者 機構本部にも配付した．またホームページ（図 5）も作成し，一般公開している．URL は以下に示すとおり．

<http://www.ism.ac.jp/kinou2/index.html>



図.5

バナー（図.6）も作成し，機構のセンターからのリンクも分かりやすくすることで，新領域融合研究センターの実体はバーチャルなセンターではあるが，ホームページ上では組織としての一体感をもたせるよう工夫した．



図.6

また機構本部の新領域融合研究センターとの事務連絡を緊密にすることと、2 研究所にまたがるサブテーマ代表者間およびPDとサブテーマ代表者間との連絡を緊密にする目的で、年度末の3月に運営マネージャーを採用した。運営マネージャーは所長直下に新たに設置した運営企画室に所属することとした。また事務連絡を正確かつ迅速に行うために、“キノウ^2”メールマガジンを発行した。

## 6. 平成17年度の研究成果

### (1) 知見・成果物・知的財産権等

各サブテーマの項目4.(1)を参照

### (2) 成果発表及び著書執筆等

各サブテーマの項目4.(2)を参照