

産総研グリッド研究センターにおける データベース・グリッド関連研究の紹介

小島 功

kojima@ni.aist.go.jp

<http://www.gtrc.aist.go.jp/dbgrid/>



National Institute of Advanced Industrial Science and Technology

概要

- **グリッド&データベース関連の研究紹介**
 - **OGSA-WebDB & OGSA-DAI 関連研究**
 - OGSA-WebDB V2.0
 - アプリケーション
 - 実用的なDB連携
 - Dspace & OGSA-DAI/OGSA-WebDB
 - **セマンティックWebを支えるデータベース・グリッド技術**
 - オントロジを支援するグリッド情報サービスの研究開発
 - OWL-S/MDS on GT4/Continual Query
 - スケーラブルなRDFデータベースの研究開発
 - P2P

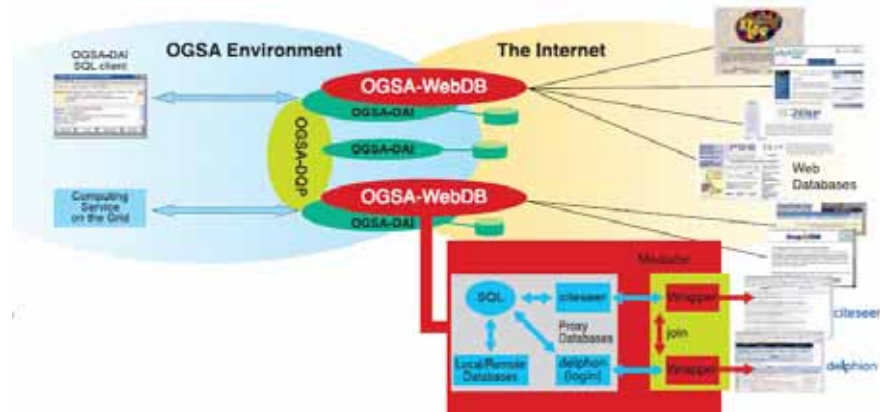


National Institute of Advanced Industrial Science and Technology

OGSA-WebDB: 既存のインターネットリソースをとにかくグリッドにつなぐ。

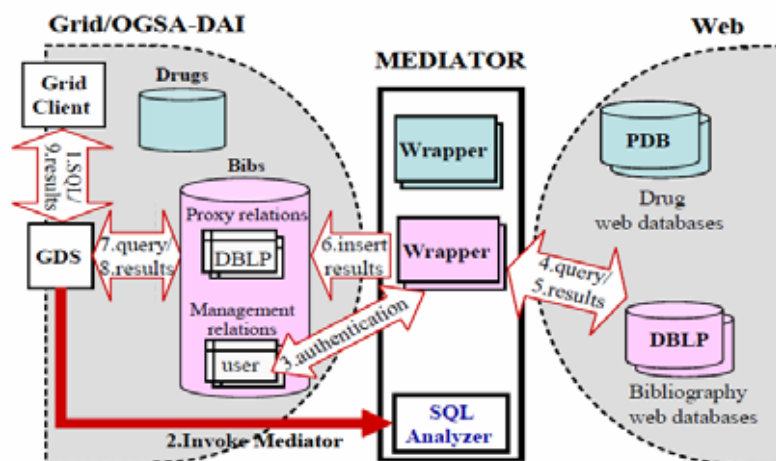
● OGSA-DAIの拡張

- 既存のWebデータベースに対するOGSA-DAI SQLアクセスを可能
- WebDBとOGSA-DAIテーブルの統合
- 他のOGSA-DAI関連ツールと連携



National Institute of Advanced Industrial Science and Technology

概要



National Institute of Advanced Industrial Science and Technology

利用機能上の特徴

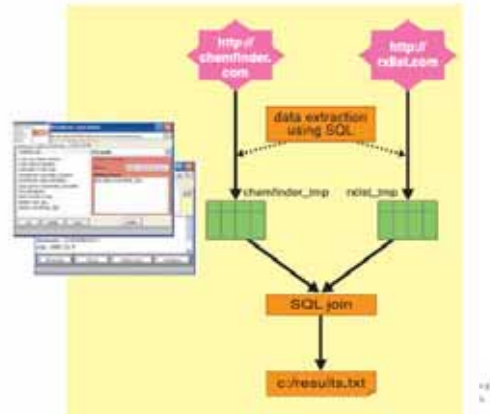
SQLGUI

- 簡単なSQLによるDB統合
- HTMLビュー



アクティビティ・エディタ

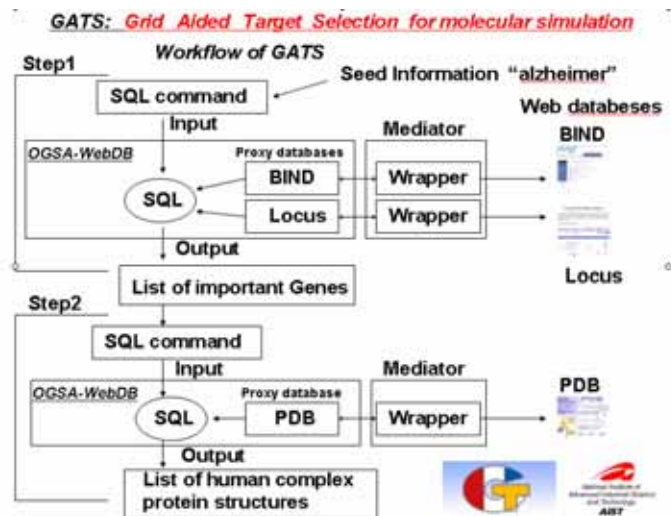
- DB処理ワークフローの構築



National Institute of Advanced Industrial Science and Technology

Application (平田他、RECOMB 2004)

SQLベースのWebデータベース統合でも、それなりに有益な情報がわかるという例



National Institute of Advanced Industrial Science and Technology

現状

- 2005年3月にV1.0 を公開
 - 現在延べ90程度のダウンロード数
 - OGSA-DAIが1500なので、その10分の1が目標
 - 図書館系、医薬・バイオ系で興味をもたれている。
- 研究的には特徴のあるソフトだと考えているが、実用的なプロダクトとしての質の点からはいろいろな課題がある。
 - インストールや設定が直感的でない、設定が手間など。
- より積極的に協力できるよう努力中
 - より多様な応用求む。



National Institute of Advanced Industrial Science and Technology

V2.0

- **Will be Presented at SC2005**
- 地道な改良
 - GT4 & WS-RFに対応
 - OGSA-DAI WSRF 1.0 ベース
 - GT4/Tomcat両コンテナに対応
 - データサービス・リソースの閲覧機能等の細かい対応
 - Postgresqlの支援
 - 支援プラットフォームの拡大
 - **ラッパ登録・導入の手間などの大幅な軽減**



National Institute of Advanced Industrial Science and Technology

例

● 従来のラッパ・WebDBの登録作業

- GUIによる やや煩瑣。

● WebDBリソースを定義するXMLスキーマの定義

- URL,ラッパスクリプト&クラス名,マッピング,ログインのあるなしなど。
- このスキーマに基づくWebDBリソースの定義

- WebDBインストーラはこれを読み取ってURLから導入する。

```

<?xml version="1.0" ?>
<webdb xmlns="http://www.gtrc.aist.go.jp/namespaces/ogsa-webdb/config"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.gtrc.aist.go.jp/namespaces/ogsa-webdb/config
http://www.gtrc.aist.go.jp/dbgrid/ogsa-webdb/schema/resource-config.xsd">
  <creation-info>
    <creator>webdb-admin</creator>
    <date>4-10-2005</date>
  </creation-info>
  <mediator-info>
    <proxy-database>
      <data-resource>http://localhost:8080/vsrf/resources/ogسادai/drug</data-resource>
      <resource-id>drug1</resource-id>
    </proxy-database>
    <proxyrelation>rdist</proxyrelation>
  </mediator-info>
  <wrapper type="webf">
    <class>
      <name>aist.webdb.wrapper.Rdist</name>
      <location>http://www.aist.go.jp/~webdb/rdist.jar</location>
    </class>
    <script>http://www.aist.go.jp/~webdb/rdist.webf</script>
  </wrapper>
  <mediator-info>
    <webdb-info>
      <url>http://www.rdist.com</url>
      <authentication>false</authentication>
      <description>A web database that provides drugs related information</description>
    </webdb-info>
  </mediator-info>
  <mapping-info>
    <field>genericname</field>
    <column>
      <name>generic</name>
      <type>varchar</type>
      <length>255</length>
    </column>
  </mapping>
  <field>brandname</field>
  <column>
    <name>brand</name>
    <type>varchar</type>
    <length>255</length>
  </column>
  </mapping>
</mapping-info>
</webdb>

```

National Institute of Advanced Industrial Science and Technology

WebDBレジストリの構築と公開

● 定義XMLファイルのレポジトリ

- Webデータベースに対するUDDI!、とまではいかないが、ある種同等の方向性と狙い。

● 定義情報やラッパのスクリプトなど、あらゆる情報を格納・管理

- 将来的にはデータリソースの資源ディレクトリとして提供。
- オントロジの構築など、意味情報の拡大

● 試験版をSCにデモ公開

- 産総研作成のラッパ群



National Institute of Advanced Industrial Science and Technology

実用的なDB連携の実現

- **OGSA-DAI & OGSA-WebDBは使えるか？**
- **我々自身の何らかの環境でやってみる。**
 - 実用的なDB連携の環境
 - 産総研の研究成果情報のデータベース



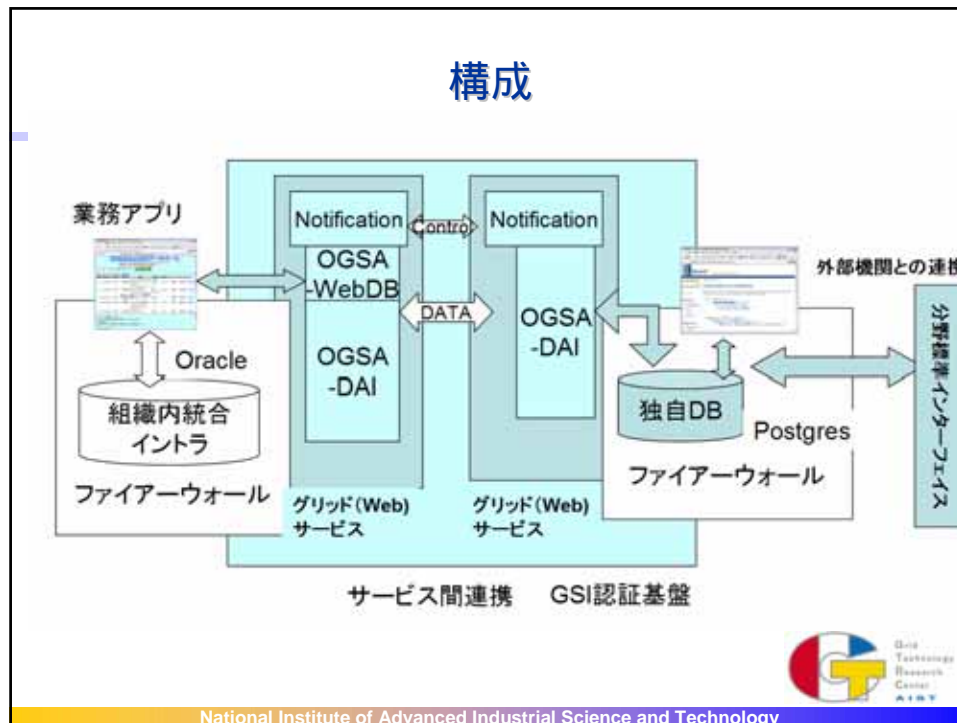
National Institute of Advanced Industrial Science and Technology

グリッド研究センター機関レポジトリの構築と、 産総研の研究成果DBとの連携

- **産総研全体の研究成果データベース**
 - Webデータベース
 - 独自インターフェイス
 - イントラ(ファイアウォール内)
- **グリッド研究センターの研究成果レポジトリ**
 - Dspace
 - OAIなどの標準プロトコル
 - 論文本体など、センター独自の情報
- **組織内分散DBの環境**
 - サービスベースのデータベース連携で実現




National Institute of Advanced Industrial Science and Technology



特徴

- **GlobusToolkit/OGSA-DAI/OGSA-WebDB**
- **Notificationを使ったサービス間更新同期の有用性**
 - Notificationによる更新通知
 - OGSA-DAI・OGSA-WebDBによる研究成果DBのラップと差分検索
- **GSIによるアカウントの一元管理(試験中)**
 - イントラ・Dspace・DBアカウントの対応関係
- **サイドエフェクト:日本語化Dspaceの改良**
 - 検索機能の拡大
 - メタデータスキーマの拡張など

● <https://www.gtrc.aist.go.jp/dspace/>



National Institute of Advanced Industrial Science and Technology

Semantic Web & Database Grid

● はやりもの = 何でこんなことをするのか？

例: OGSA-WebDB: 膨大なWebリソースがグリッドに取り込まれたとする。

- リソースの管理は大変。(だからリソースのディレクトリ・DB化をいくつかある)

● つくばWANの公開Webデータベース調査

- つくば地域の研究機関が公開する研究関連Webデータベース群

- 2003年 現在1000URLほどの収集・管理

- 他のディレクトリと差別化するため、概要文や説明を抽出
- 人手をある程度関与させて、知的な収集・整理

(現在改良中)



- 多様な構造、意味、アクセス手段

- 現在は単に全文検索ポータルとして提供

- 将来的にグリッドなどの環境でうまくデータ統合を図りたい。
- 分散データマイニングなど

- これらをうまく結びつけるのは、それだけではとても難しい。

● ボトムアップ的に作られたリソースをどう管理するか？

- 意味記述など、意味表現の必要性
- オントロジなど、整理の階層の有用性



National Institute of Advanced Industrial Science and Technology

データベース統合

● 産総研内のDB統合でも、スキーマはずいぶん違う。

- MIX(XQuery-XQuery変換ツール)

- XMLスキーマレベルでの統合の限界

- プログラミング的には有益だが、意味的にはわかりにくい。
(データベースの全容がつかみにくい)

● Semantic Web技術との関連

- Semantic Grid
- 次世代のSOA基盤？

● グリッド技術による(ある程度の)下支えの必要性

- Semantic Web技術で画期的に状況が変わるわけではない(と思う)が、ある種のインターフェイスや技術の適用・導入はやはり必要。



National Institute of Advanced Industrial Science and Technology

問題と成果

- **グリッド上でのリソースの発見・検索**
オントロジを支援するリソース情報サービス

1st International Conference on Semantic and Knowledge Grid
 (11月28 - 29、北京)で発表予定
<http://kg.ict.ac.cn/skg2005/>

- **RDF(Resource Description Framework)の大規模なレポジトリ**
P2PベースのRDF問い合わせ処理技法

DBWeb2005 (11月21日-22日、日本未来館@お台場)で発表予定
<http://hikendbs.eei.metro-u.ac.jp/dbweb2005/>

- **どちらも発表前なので、本日はいずれもそれらの背景と基本的な考え方**
などを紹介



National Institute of Advanced Industrial Science and Technology

RDF(Resource Description Framework)とは。

RDF: メタデータ記述の基礎を提供する枠組み

- **基本概念**

- **すべての情報を文(主語, 述語, 目的語)で表現**
- **RDF3つ組(トリプル)**



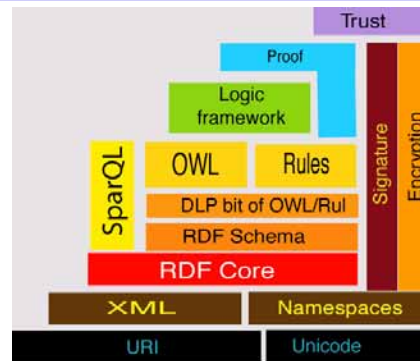
```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person> </rdf:RDF>
```



National Institute of Advanced Industrial Science and Technology

RDF 続き

- Semantic Webのプロトコルスタックにおいて基礎的なフレームワーク
- OWL(Web Ontology Language) も含め、グリッド環境側からの支援の試みがあってよい。



National Institute of Advanced Industrial Science and Technology

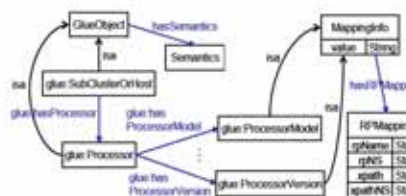
オントロジを支援するリソース管理サービス

- SC2005でデモ展示予定 & SKG2005で発表予定

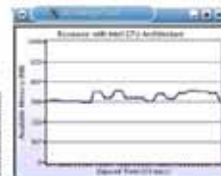
大雑把には、

GlobusMDS4 + オントロジ + Continual Query (CQ)

オントロジ階層による資源の整理



CQによる動的モニタリング



National Institute of Advanced Industrial Science and Technology

動機と背景

- WebDBなどのデータリソースの管理 オントロジなど意味を扱うことが不可欠
- 計算リソースなど、他のリソースにも同様の問題？
 - リソースの高度なマッチメーキングでは、意味を扱うことが不可欠。
 - LinuxもSolarisもUnix系(である種の資源互換性がある)といったように、直接満足する資源がなくても、同等の代替資源を提供する場合など
- **さまざまなグリッド上のリソース(計算リソース、データリソース)を統合して扱い、かつ意味(オントロジ)を扱えるリソース管理サービスの必要性**



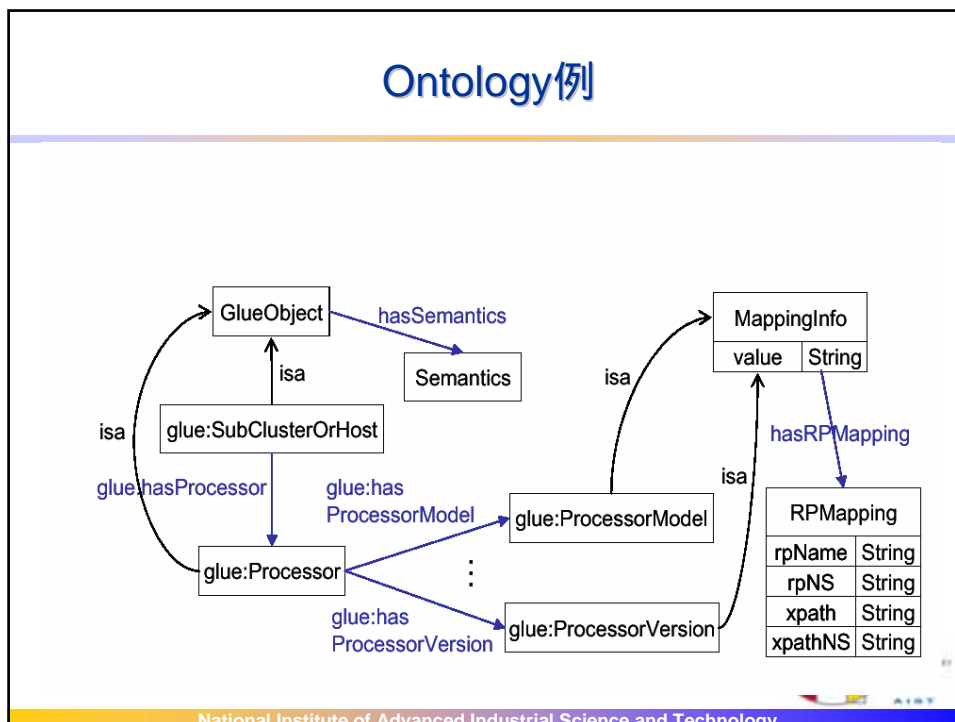
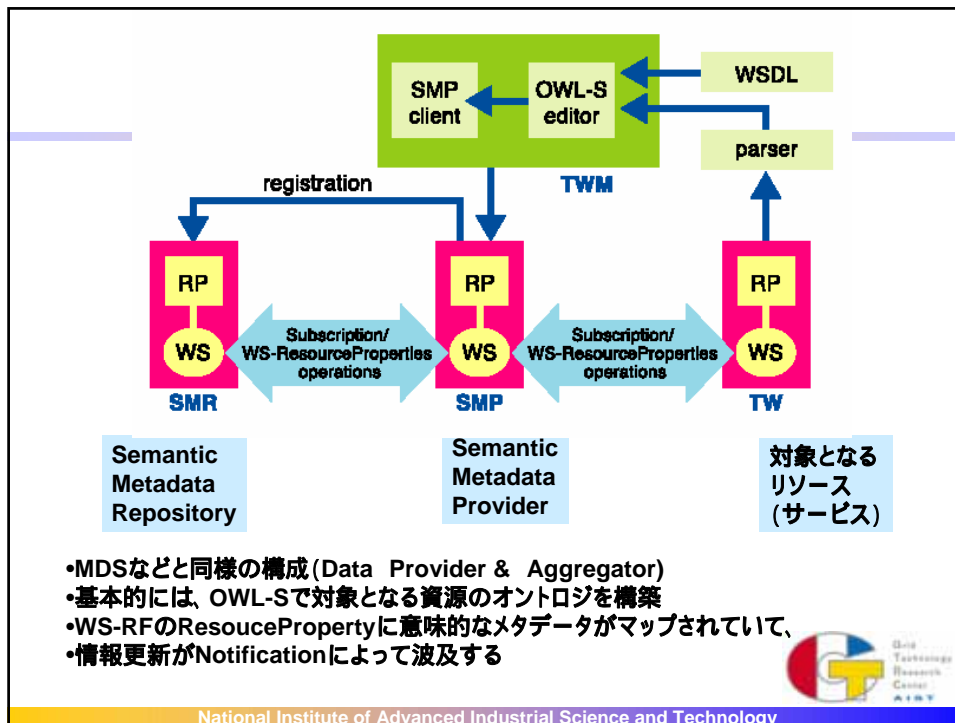
National Institute of Advanced Industrial Science and Technology

概要

- 各リソース(データリソース、計算リソース)をRDF/OWL-Sなど、オントロジを支援するフレームワークで記述・管理
- リソースの発見はRDFの問い合わせ言語などで行う。
- **実装上のポイント**
 - 既存のグリッド・リソース管理の体系との親和性
 - まったく別の体系・システムを作っても仕方ない
 - **WS - RF規格やGT4 MDSとの親和性**
 - オントロジのWS-RF: リソースプロパティへのマッピング
 - MDSとのアーキテクチャ的互換性の実現



National Institute of Advanced Industrial Science and Technology

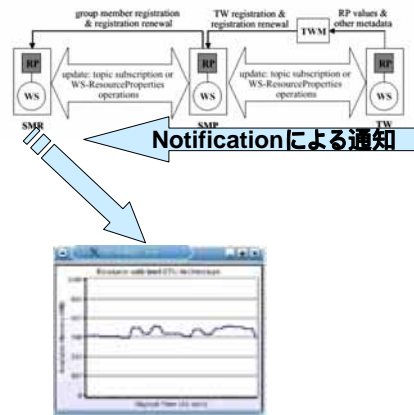


特徴 Continual Query

- **MDSのNotificationによる情報更新通知機能**
リソースの情報を最新状態に維持。

- **RDF問い合わせ言語における、Continual Queryの実現。**
検索結果データの前データが更新されると、結果も自動的に変わる。

- **詳細は、SC & SKG 2005で**



RDFの問い合わせによって処理されている。
(Notificationによって値が動的に変わる)



National Institute of Advanced Industrial Science and Technology

大規模なRDFストレージの実現の課題

- **スケーラビリティの提供**
 - リソース記述などのメタデータ=>RDF
 - インターネット上に分散。
 - ボトムアップ的に作成 = 非集中管理
 - 相互に関連し、全体として大きな意味の体系を構築しうる。
 - 非集中管理の下での、スケーラビリティの実現
 - 研究課題:P2P (Peer-to-Peer)技術の適用
- **サービスベースのインターフェイス提供**
 - セマンティックWebの研究活動との連携
 - 規格化活動との連携 W3C,GGF
 - ミドルウェア



National Institute of Advanced Industrial Science and Technology

RDF P2Pの課題

- **検索言語をサポートしないといけない。**
 - **通常のP2P=>オブジェクトのアクセスのみ**
 - たとえば、名前をキーとするオブジェクトの検索など)
 - **集合検索ができない**
 - x RDFPeersなど
 - **RDFの問い合わせ言語**
 - SPARQL(W3Cで標準化しつつある)
 - RDF固有の検索機能
 - クラス(オントロジ)にかかわる検索など



National Institute of Advanced Industrial Science and Technology

RDF P2Pの課題2

- **P2P上の集合演算**
 - **そのままでは非現実的なアクセス量**
- **RDFデータとは？**
 - 主語・述語・目的語からなる三つ組 (Subject, Predicate, Object)
 - 情報単体の構造自体は単純で、限られる
この構造をうまく生かすことで、集合演算など問い合わせ処理の効率化を図る。

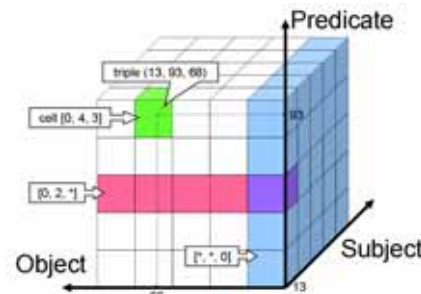


National Institute of Advanced Industrial Science and Technology

アプローチ

● RDFCube(主語、目的語、述語からなる3次元空間)の定義

- 各RDFトリプルはこの空間(ハッシュ空間)のどこかにある。
- 仮想的に部分集合(セル)に分ける。
セル内にデータが含まれるかどうかを示すビットマップを導入する。
 - 一種の索引情報



● DHT(Distributed Hash Table)へ適用

- P2Pにおける一般的な実装手法
- ハッシュの三次元空間をDHTで実装

ビットマップを使って集合演算におけるデータアクセスを減らす。
ビット処理にからむDHTを別途構築し、併用する。



National Institute of Advanced Industrial Science and Technology

結果

● 実験プロトタイプ

- DHTのシミュレーション環境で実装
- SPARQLのサポート
 - 結合的な演算の評価
- DBLP(DB関連の文献データベース)のデータで実験・RDFPeersと比較した場合の有用性を確認

● 詳細・興味のある人はぜひDBWeb2005にて。
実装は、WSDL 1.1バインディングを意識
WS-DAI*とのリエゾンなど、今後の課題



National Institute of Advanced Industrial Science and Technology

まとめ

- **実用的なデータ統合基盤の研究開発**
 - OGSA-WebDB・OGSA-DAIとその周辺の研究開発
 - 応用

- **次世代のデータグリッド基盤のための研究開発**
 - 高機能のグリッド資源管理サービス
 - MDSと互換
 - オントロジの支援
 - Continual Queryの実装
 - 大規模分散RDFストレージ
 - P2P技術の応用
 - サービスベースインターフェイスと実装

