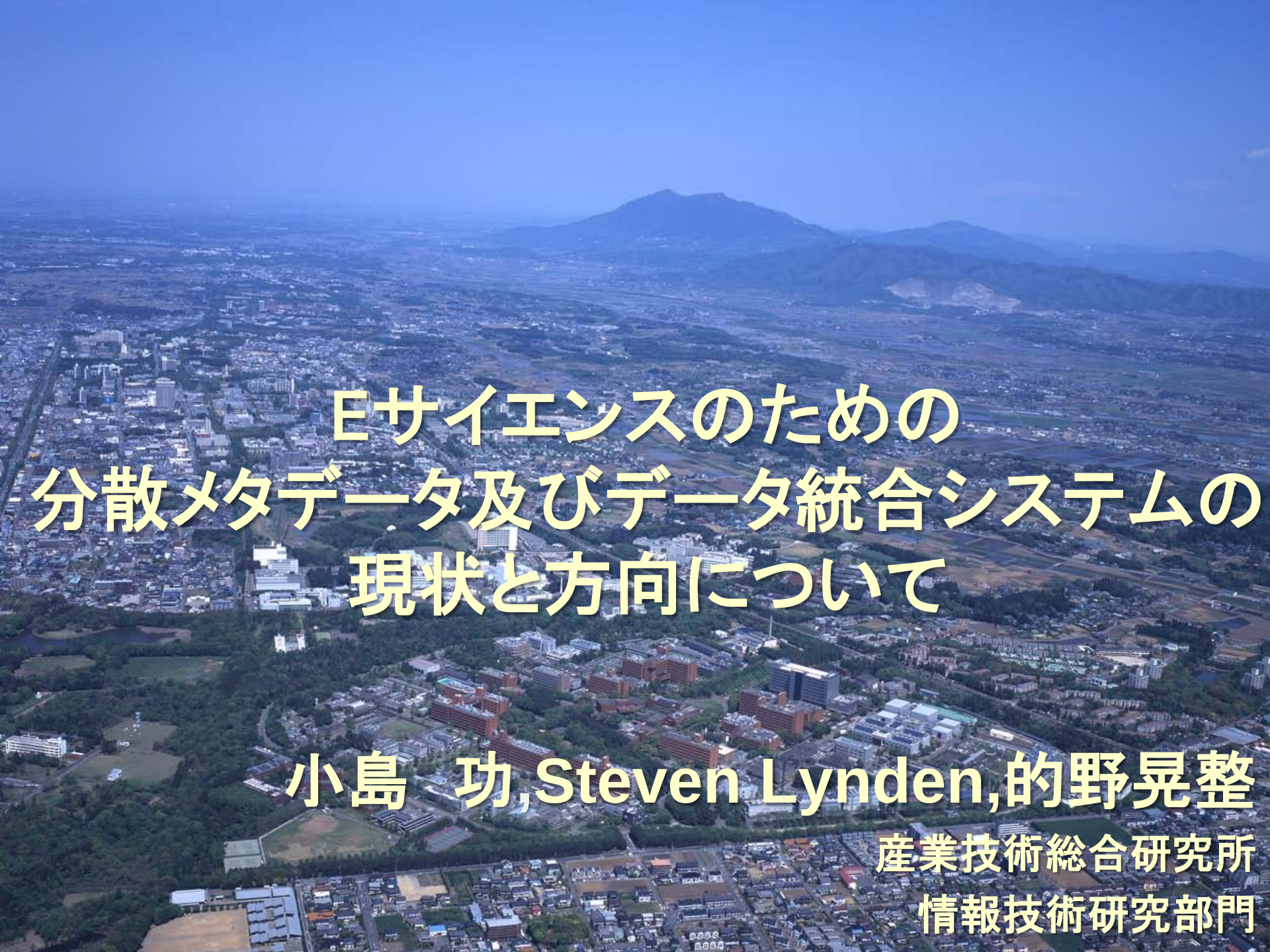


Eサイエンスのための分散メタデータ及びデータ統合システムの現状と方向について

小島　功,Steven Lynden,的野晃整

産業技術総合研究所

情報技術研究部門

まとめ

ファイルもいいけど、DBも重要

データもいいけど、メタデータも重要

協力・連携が重要

概要

- 分散データベース統合でやっていること
  - Federated SPARQL system
  - Demo
- メタデータ統合・検索でやっていること
  - AIST CSW(Catalog Service Web)
  - Demo in GEO Grid
- おまけ
  - 動機と課題など
  - @SC10

大量データに基づく科学的知見の創出

- 膨大なデータに基づいた解析
  - 有益な知見を絞り出し、再利用

  → データに基づく科学のライフサイクル

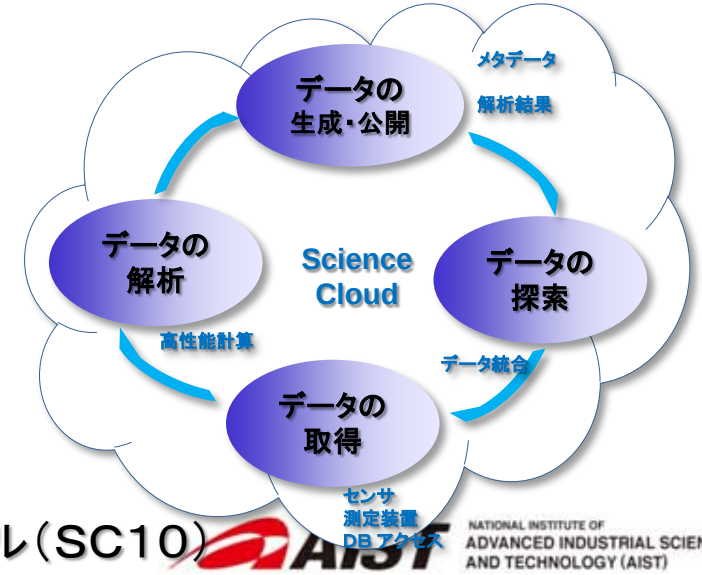
  メタデータが適切に付与されて管理される必要がある

- 解析処理のワーキングセット
  - 膨大なデータ集合のうち、解析に必要なデータの絞り込みのかなりはDBでできる・やる場合も多い。

  DBによる（メタデータ＆）データ管理と、解析（計算）との効果的な連携

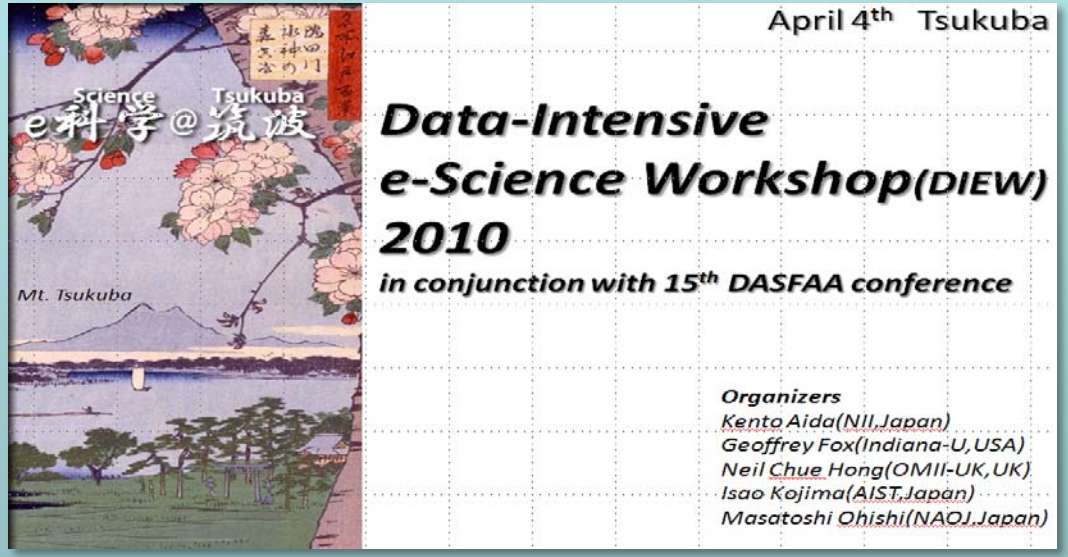
GEO Grid

- ファイル　on Gfarm 約800TB、約180万シーンのASTERが中心
  - DB=PostGIS&独自のカタログ実装(メタデータおよび構造化データの格納)
  - 全量の同時処理・解析＝あることはあるが、あまりない
    - 固定したワークフローが多く、あまり改造・変更がない(維持コストがかかるが、、)
    - パラメータなどの変更による再処理。
      - 再処理化のパラメータ等のメタデータの生成・管理・次の検索に対する利用が重要
  - 検索した結果の解析、興味のある対象を選択しての処理＝とても多い
    - DBを使ったデータ・メタデータ検索との連携が、使い勝手に非常に効く
- データに基づいたライフサイクル
  - 個々の要素(ワークフローや解析)も
    大事だが、それらを使って【回る】
    基盤にしないといけない。
  - メタデータの生成と管理
    - Provenance
    - メタデータ作成と連動したワークフローツール(SC10)



（異種）分散データベース統合

AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)



Malcolm Atkinson (Uk-eScience)
今後の連携につながる議論が行われた

Data-Intensive e-Science Workshop
In conjunction with DASFAA conference
@筑波大

ご協力ありがとうございました。

アプリケーション

Semantic Data Infrastructure to support a Scientific Dataspace for Breath Gas Analysis(jointly presented at AHM 2010)

Ibrahim Elsayed, **Steven Lynden, Isao Kojima** and Peter Brezany

AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)



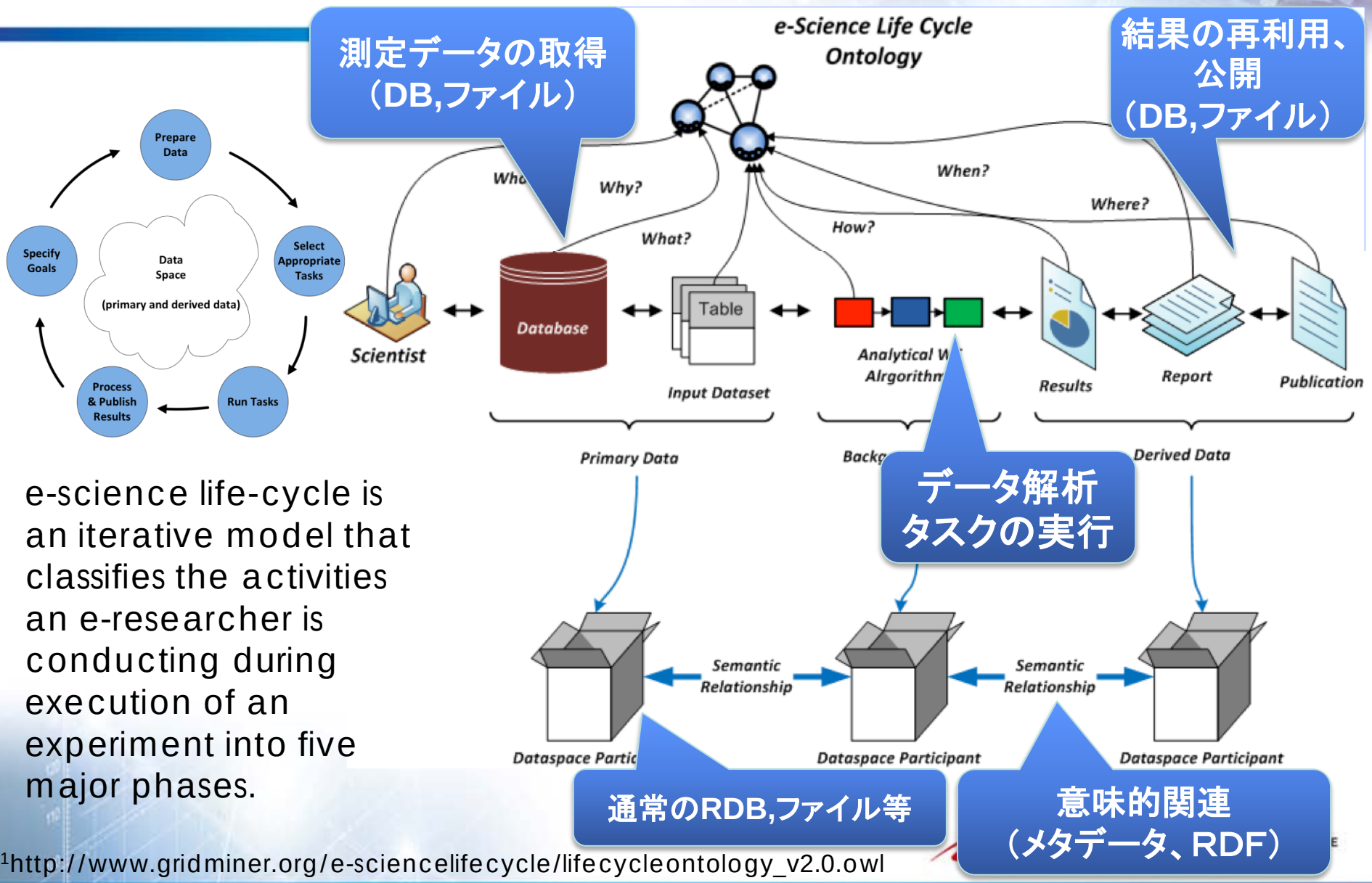


Breath Gas Analysis (呼気分析)

- Emerging new scientific field with a growing international community
- Strong evidence to detect specific cancers using the concentration pattern of volatile organic compounds
- Investigating and screening for hundreds of compounds in exhaled breath gas

AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

e-Science におけるライフサイクルの支援

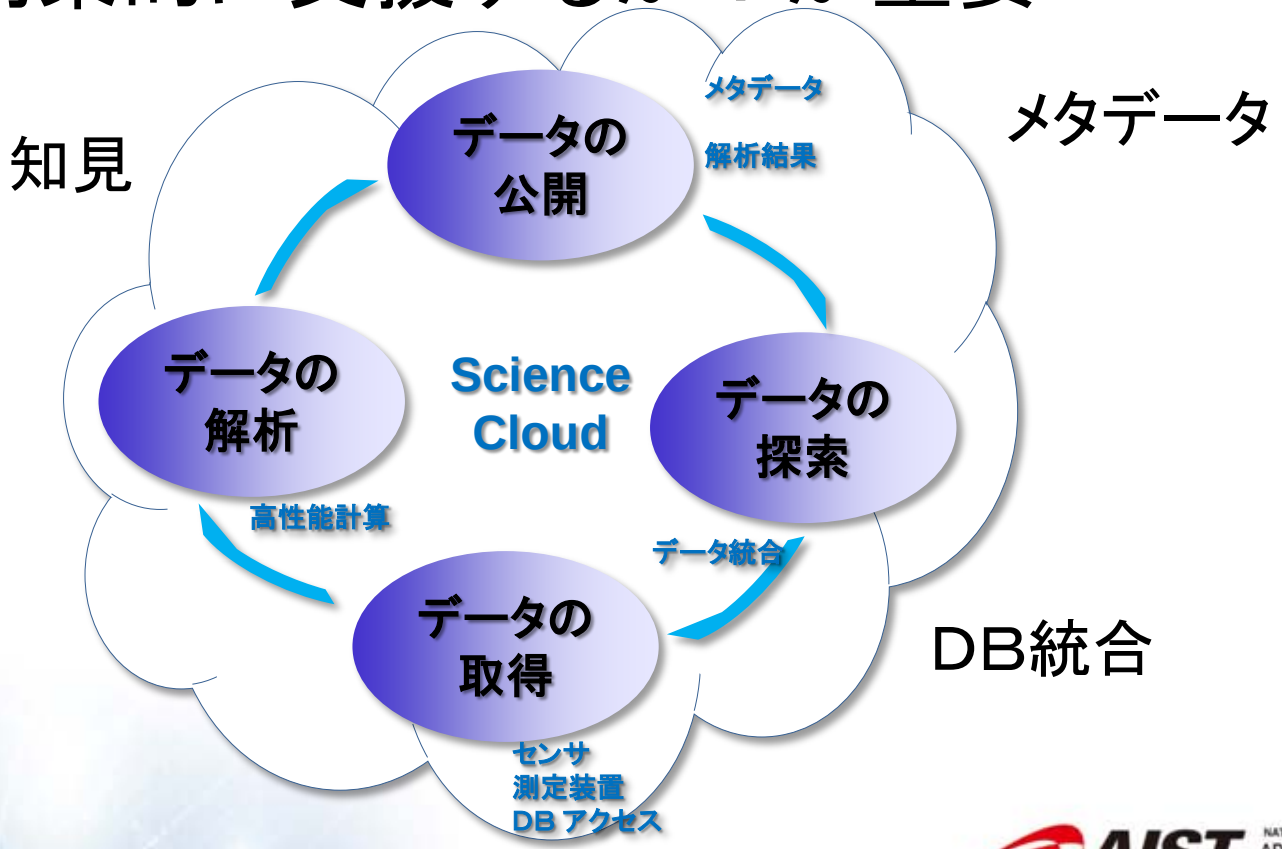


e-science life-cycle is an iterative model that classifies the activities an e-researcher is conducting during execution of an experiment into five major phases.

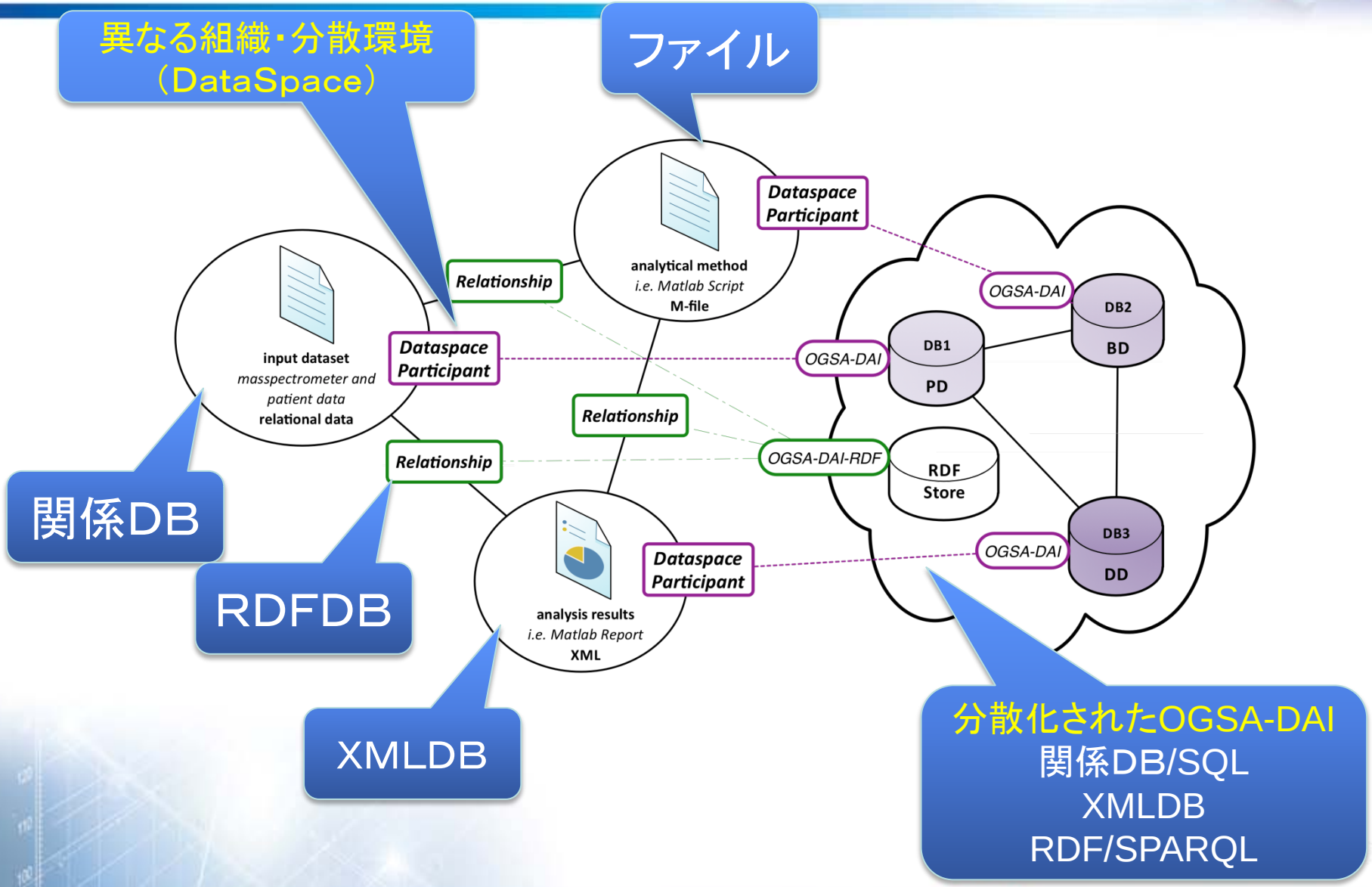
[1]http://www.gridminer.org/e-sciencelifecycle/lifecycleontology_v2.0.owl

eScience基盤の使い勝手

- データに基づくライフサイクルを
  どう効果的に支援するか？が重要



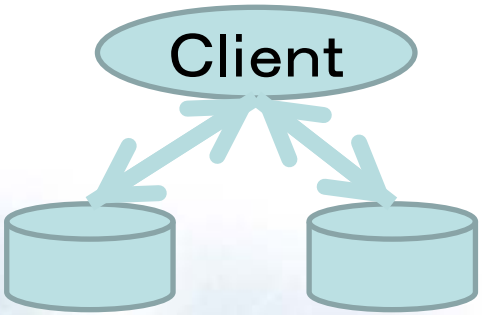
分散環境での異種データの管理

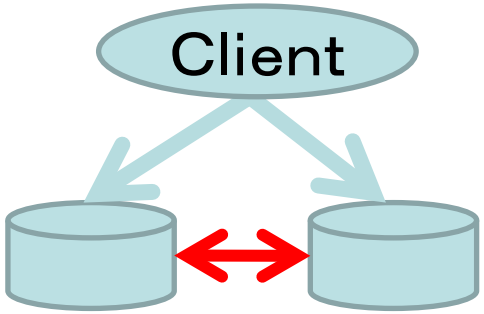


DB統合/連携の要件

- 異種のデータが同じフレームワークで扱えること
  - XML,RDB,RDF,web etc

- サイト「間」処理によるデータ統合
  - 同じ問合せを全サイトへ（簡単）
  - 分散結合（最適化が入るのでややこしい）



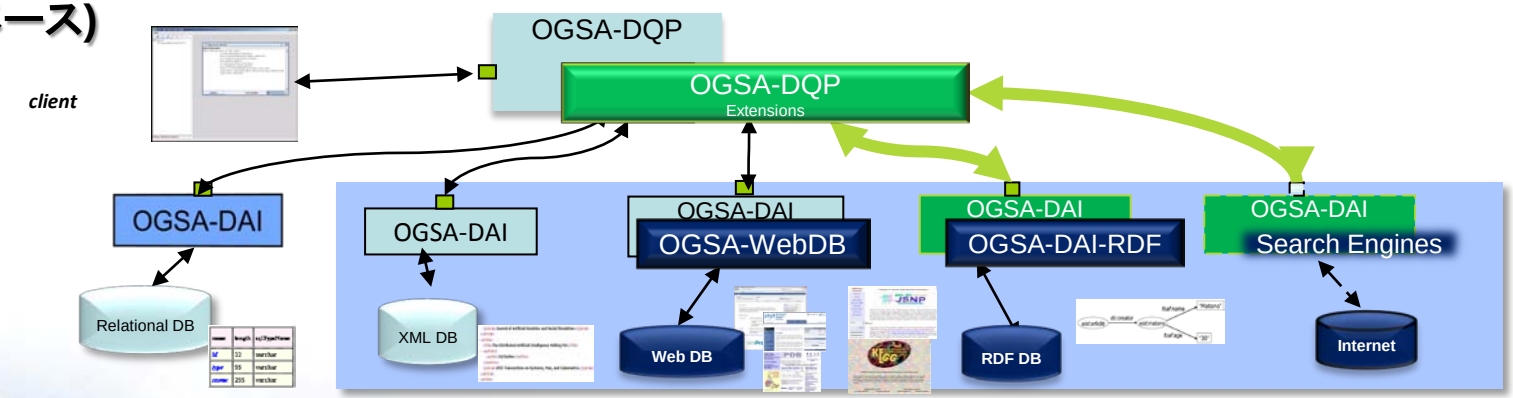


データベース連携・統合技術に関する研究

1.　異種の分散データベースを連携・統合するミドルウェアの研究開発

- 宣言的な言語（SQLやSPARQL)で複数のDBを連携
- DBの異種性（全文検索やRDF)を吸収する機能
- 分散処理の実行順序の（動的な）最適化機能

プロトタイプ (OGSA-DAI ベース)



データベース連携に関する研究

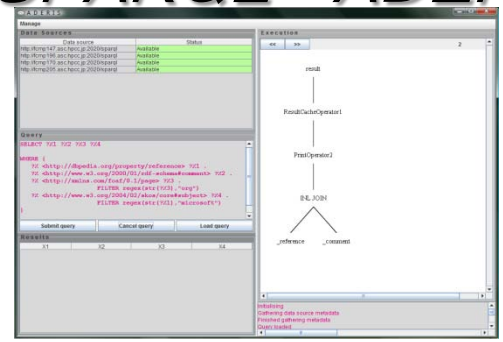
DQP-XML/WebDB (SQL for XML,RDB,WebDB)

SPARQL-ADERIS (SPARQL for RDF)

特徴紹介： 動的な分散最適化機構

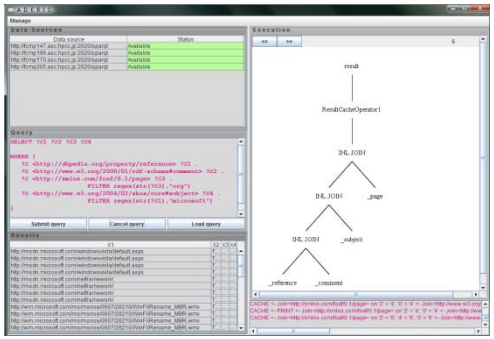
問合せ実行前の分散処理スケジュール

SPARQL－ADERIS GUI



ネットワークの遅延
予想と異なる中間結果
サイトのトラブル、、、

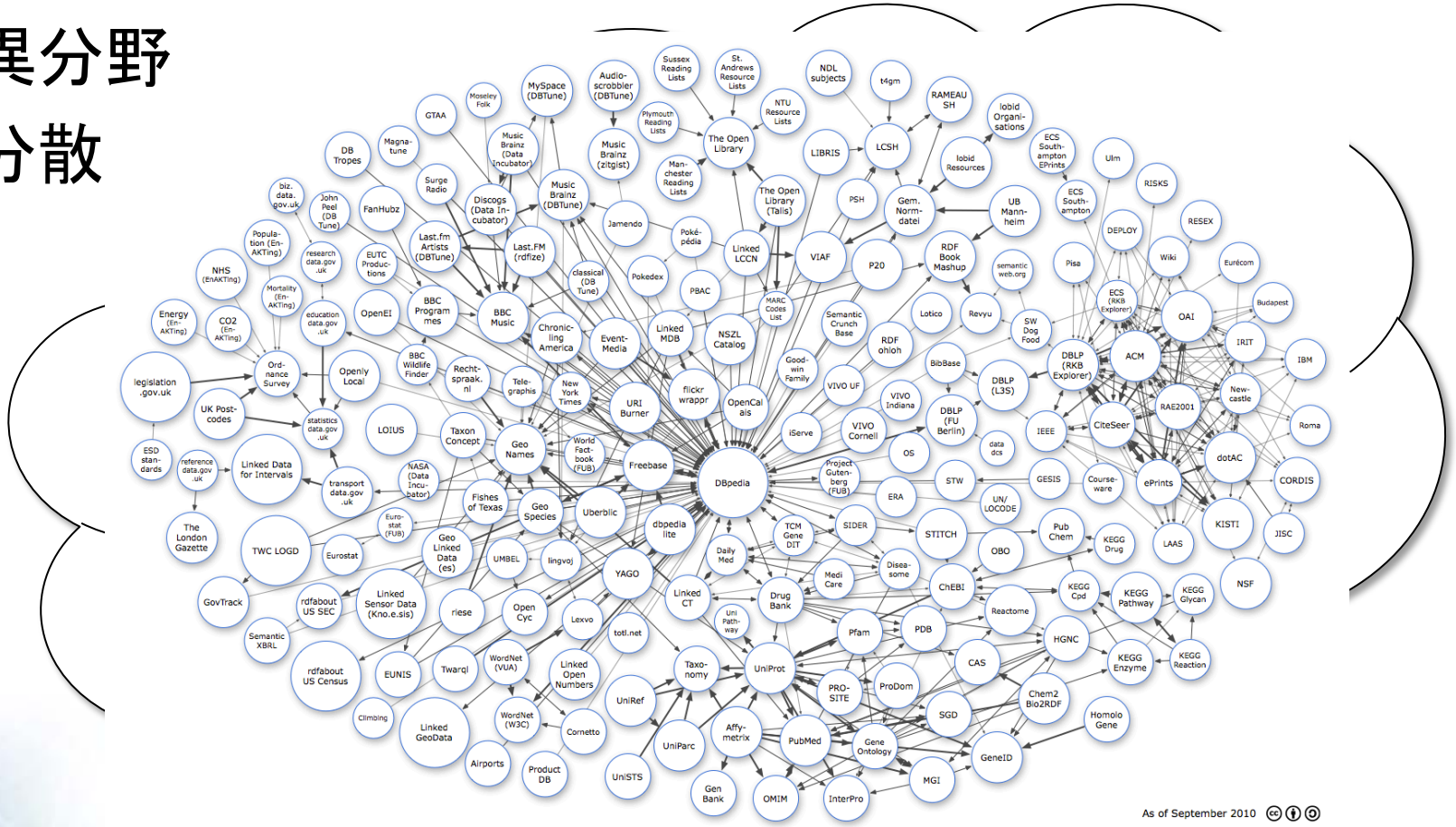
問合せ実行中に動的に変更した
分散処理スケジュール



- Demonstration

Linking Open Data

- RDF Metadata
- 異分野
- 分散



AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

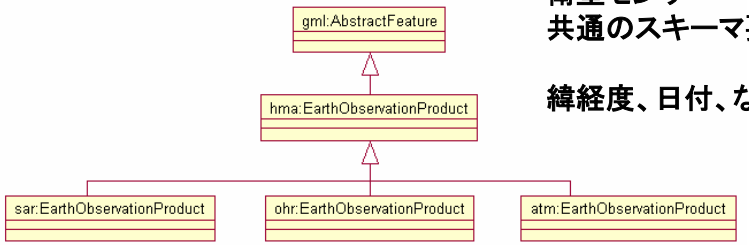
メタデータ統合

メタデータにおける事情

- 標準が多過ぎ
  - 地球観測だけでも、、
    - Dublin Core
    - ISO,JMP
    - ebRim
- それでもさらにフォーマットが必要、、
  - 衛星やセンサ固有の項目を入れてほしい、、
- その割に項目が埋まってないぞ、、

GEO Gridにおける メタデータ設計の考え方

- ebRim(e-bussinessの情報モデル)に基づく
  - Webサービスの(OASIS)標準
  - オブジェクト間の関連で記述(拡張性が高い)
  - OGC標準
- モデルの拡張性(スキーマ間の階層関係)



衛星センサ
共通のスキーマ要素

緯経度、日付、などなど

合成開口レーダー(SAR)
(Synthetic Apature Rador)

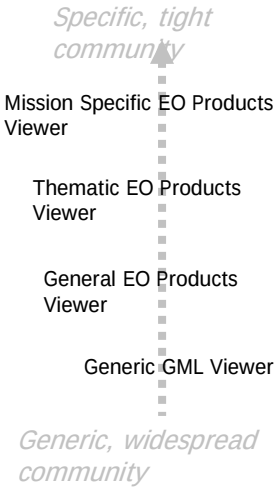
PolarisationModeなど

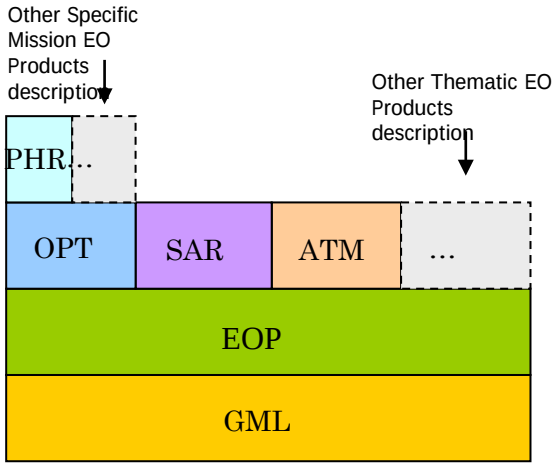
光学(OHR)
(Optical High Resolution)

雲量・雪など

大気(ATM)
(Atmospheric)

DatalayerSpeciesなど
(NO2,O3,,,,)





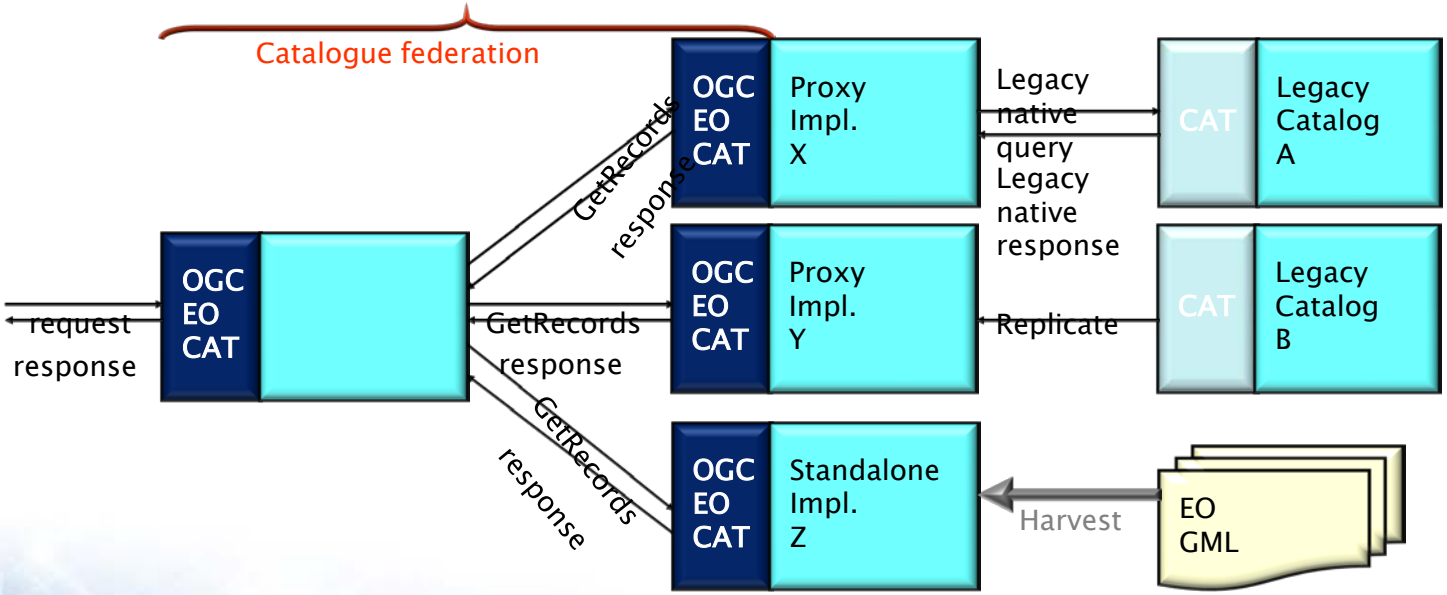
<rim:ExtrinsicObject id="urn:uuid:bce71bb1-d71b-40a9-ae91-201cbfdc61e7" objectType="urn:x-ogc:specification:csw-ebrim-cim:ObjectType:DataMetadata">
<rim:Slot name="modified" slotType="dateTime">
<rim:ValueList> <rim:Value>2006-06-15T15:00:00Z</rim:Value>
</rim:ValueList>
</rim:Slot>
<rim:Slot name="envelope" slotType="geometry">
<wrs:AnyValue xmlns:gml="http://www.opengis.net/gml"> <gml:Polygon srsName="EPSG:4326">
<gml:outerBoundaryIs>
<gml:LinearRing>
<gml:coordinates>139.4282,35.4882 140.2614,35.4882 140.2614,36.1379 139.4282,36.1379 139.4282,35.4882</gml:coordinates>
</gml:LinearRing>
</gml:outerBoundaryIs>
</gml:Polygon>
</wrs:AnyValue>
</wrs:ValueList>
</rim:Slot>
<rim:Slot name="title" slotType="string">
<rim:ValueList> <rim:Value>URI: {
"uri": "http://maps.geogrid.org/mapserv/ms_aster.pl?",
"option": {
"LAYERS": "ASTL1A_0606161247510612129002.dat",
"SERVICE": "WMS",
"VERSION": "1.1.1"
}
<rim:Value>THUMBNAIL-URI: {
"uri": "http://www.geogrid.org/cgi-bin/thumb.pl?",
"option": {
"res": "small",
"type": "jpeg",
"filename": "ASTL1A_0606161247510612129002.dat"
}
}</rim:Value>

ebRim Object

ないものは書いてない。
ISOと統合して検索したい、
Polygon,
Date
etc

Catalog Service Web

- GEO Grid でサポートする一連のOGC規格の一つ
  - **REST(HTTP GET/PUT)** & SOAP
    - OpenSearchの上位互換（次のCSW3.0)
  - 分散カタログの検索を支援→**複数組織・横断検索に有利**



基盤における考え方

- 全文検索に基づく
  - どんなデータ表現でも検索できる
    - システムへの取り込みは間口を広く
      - 形式が違ってても全文検索ならOK
- タグの検索を支援
  - 構造に基づいた検索「も」できる
    - 構造を意識したい人にはそれなりに

全文検索エンジンに基づく実装

DBMS(PostGISとか）を使ってない

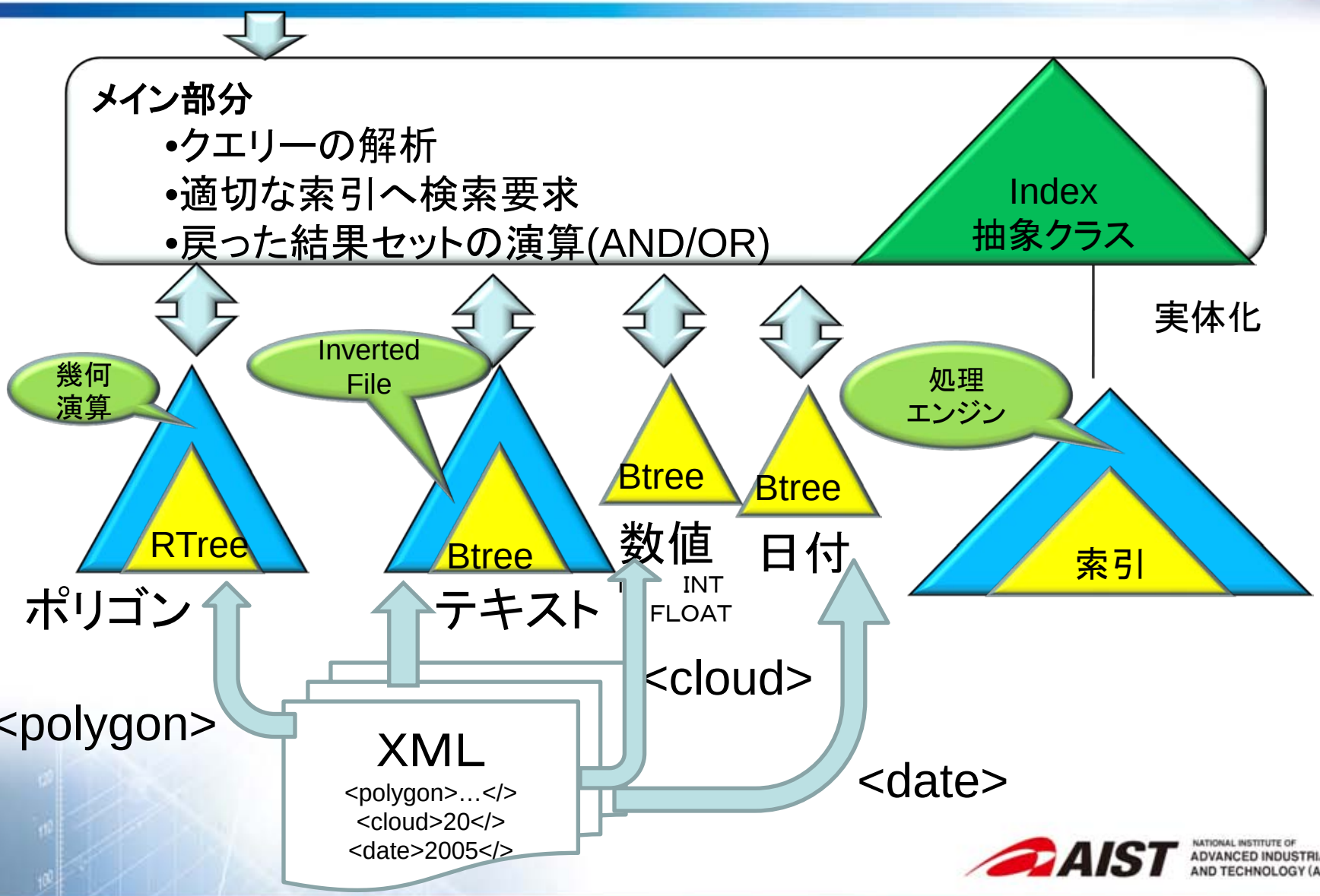
遠目には、、ジオメトリ検索(overlap、cover等）ができる全文検索と
(類似システムあり）

- 検索：特定のフィールドに対する検索条件のAND・OR：
  - SQLのような記述力は不要。
- 様々なXML形式の格納：
  - スキーマレス
- 全文検索：検索結果の数に対する応答性能が高い。
  - ページングなど。
- 地球観測メタデータ
  - 更新がほとんどなく追加のみ。複雑な索引構造でもOK.
- 全文検索
  - やっぱり便利

- 概念的には
  - ebRIM XML Schema
    - スキーマの階層関係を意識
  - ISOなどとの対応関係を設定・保持して統合
- 物理表現的には
  - CSV＆JSONで処理
    - 中間処理が速い
  - 標準(ebRim)のXML表現を全文検索インデクスに保持
    - 結果表示が速い
    - その他の表現はXML間のデータ変換で
- アクセスプロトコル
  - 標準（OGC-CSW）に従う→将来的にはもう少し簡便な標準に
    - OpenArchiveなどへの対応なども含む
  - メタデータの自動取得が可能(メタデータ作成は労力大）

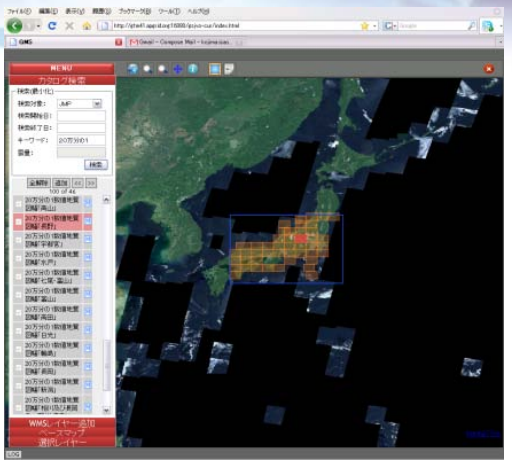
問合せ
Form形式

OGC　Filter条件式：
XMLの各フィールドに対する比較条件の
AND＆OR
（SQLのような結合はない）

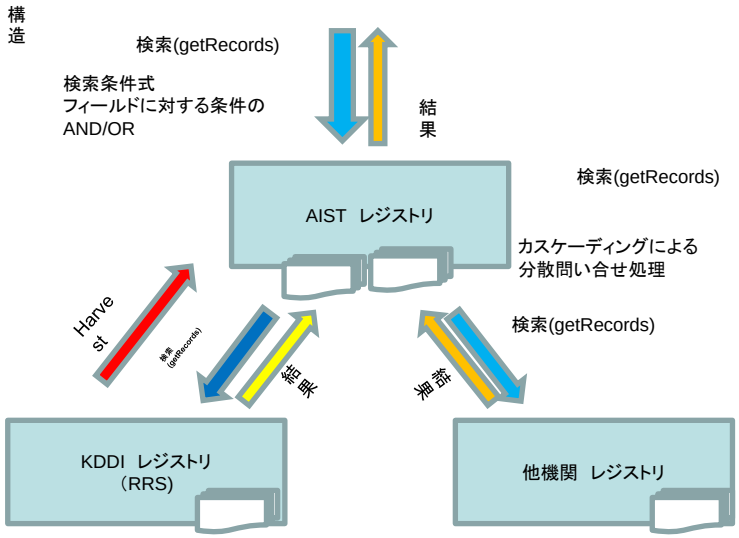


応用

- ASTER衛星シーンカタログ
  - 約180万？レコード
  - BBOXなどのジオメトリ検索



- 計算資源管理レジストリ
  - ネットワーク資源
  - ストレージ資源
  - 計算機資源



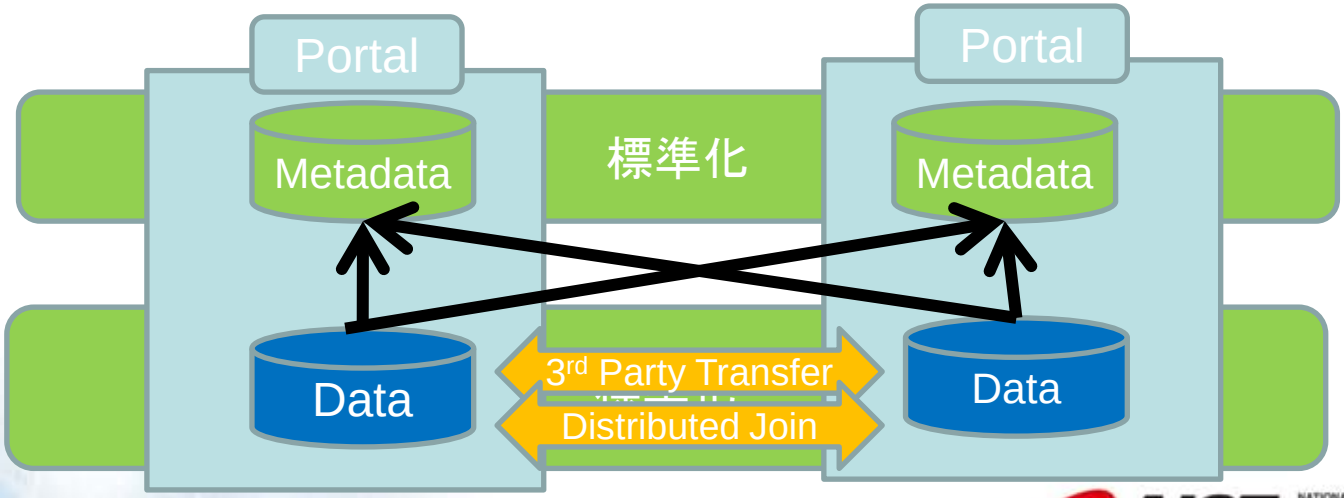
  - Q“何TB以上のディスクと何CPU以上、OSが何を何個、、、”とか

- Demonstration

連携・協力をしませんか

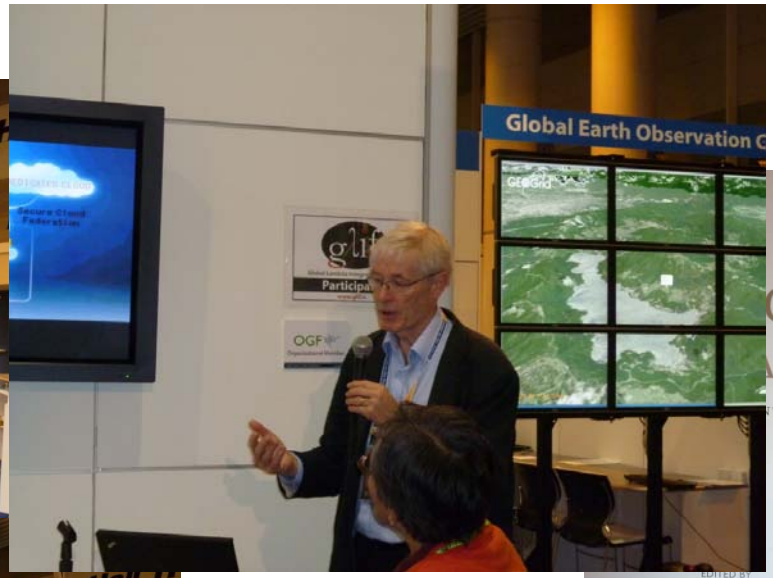
GEO Gridにおける？
現在のSystem of Systems の1方向

- データは各組織が固有に持つ
  - セキュリティポリシーを保持
- メタデータは相互コピー/CrawlingでOK or Cascading
  - みんなポータルを作りたがる→作って可。
  - RDFへ移行
- 標準化プロトコル（OGC）によるアクセス
  - より簡潔なアクセスインターフェイスへ
- サイト間のデータ転送＆分散結合
  - ＋データ変換サービス

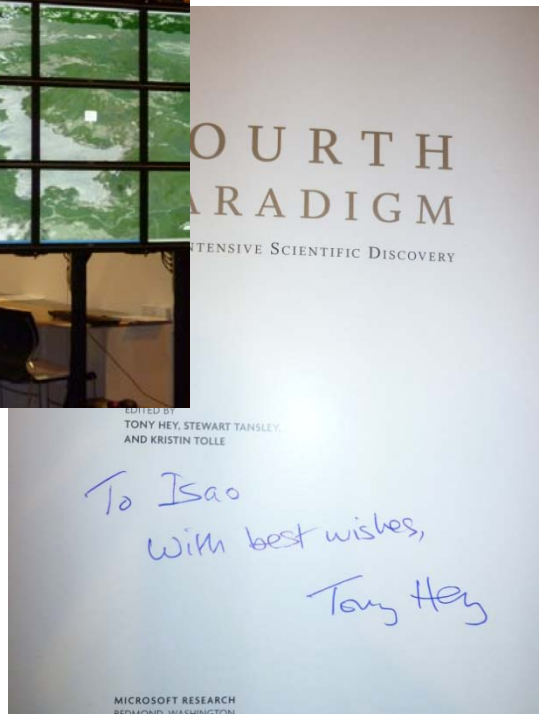


AIST Science Cloud Talks @SC10

Open Science Data Cloud
Rob Grossman@OCC
Malcolm Atkinson@Nesc



AzureMODIS
Tony Hey@Microsoft



Mapreduce&FutureGrid
Geoffrey Fox@Indiana-U

AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

関連イベント

データ工学研究会(Eサイエンス＆一般)

- 12月6日
- 産総研秋葉原(ダイビル)