筑波大学
University of Tsukuba

データ科学に対するデータ工学的アプローチについて

天笠俊之，川島英之，北川博之
筑波大学大学院システム情報工学研究科

1

筑波大学
University of Tsukuba

自己紹介

- 名前
  - 天笠俊之
- 所属
  - 筑波大学大学院システム情報工学研究科
    - 北川データ工学研究室
  - 筑波大学計算科学研究センター
- 研究テーマ
  - データ工学
  - データベース，データベースシステム
  - XMLデータ，XMLデータベース
  - 科学分野におけるデータ工学の応用

2

筑波大学
University of Tsukuba

筑波大学計算科学研究センター あらまし

3

筑波大学計算科学研究センター 研究分野

筑波大学
University of Tsukuba

科学分野

- 素粒子宇宙研究部門
  - **素粒子分野**
  - 宇宙分野
- 物質生命研究部門
  - 計算物性科学分野
  - 計算生命科学分野
  - 量子多体分野
- 地球生物環境研究部門
  - **地球環境分野**
  - 生物分野

CS分野

- 超高速計算システム研究分野
  - 計算機アーキテクチャ分野
  - グリッド分野
- 計算情報学研究部門
  - 計算知能分野
  - 計算メディア分野

4

筑波大学 University of Tsukuba

その他のコラボレーション

- 産業技術総合研究所
  - GEOGridプロジェクト
    - 大規模異種衛星センサデータ
- 国土交通省国土技術政策総合研究所
  - 河川測量データ，レーザープロファイラデータ，…
  - 河川シミュレーション
    - 定流計算，不定流計算
    - 河川計画
  - 洪水シミュレーション
    - 都市計画

5

筑波大学 University of Tsukuba

今回の話題

1. 格子QCDメタデータQCDmlの
   ファセット検索インタフェース構築

2. FUSEによる遠隔気象データアクセスミドルウェア

6

筑波大学 University of Tsukuba

格子QCDメタデータQCDmlのファセット検索インタフェース構築

天笠俊之，石井理修，吉江友照，建部修見，佐藤三久

7

筑波大学 University of Tsukuba

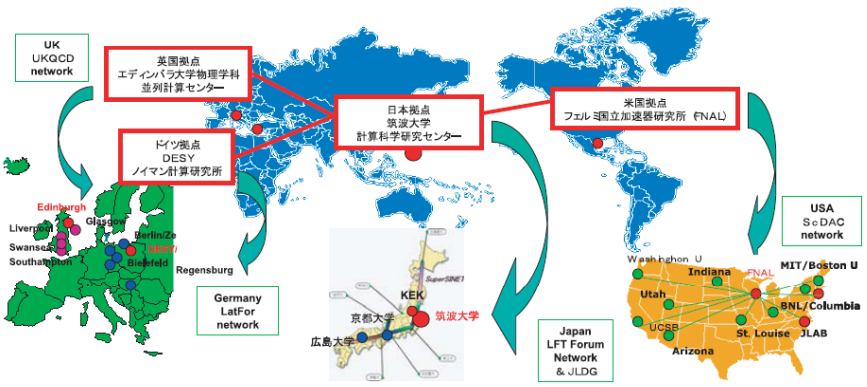
QCDml

- 格子QCD
  - 各子の中のクォークを結びつけている力を記述する力学
    - 量子色力学 (Quantum Chromo Dynamics)
  - QCDを厳密に解く → クォークの質量から陽子や中性子の質量が予言できる
- 格子QCD (Lattice QCD)
  - QCDを解くために，時空を格子化し有限自由度で計算
- ILDG (International Lattice Data Grid)
  - 格子QCD計算の計算結果である配位データを国際的に共有するためのデータグリッド

8

筑波大学 University of Tsukuba

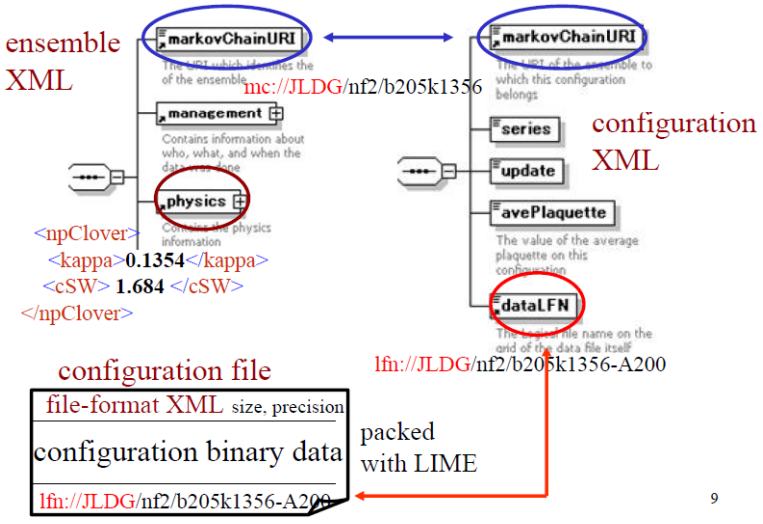
ILDG (Int'l Lattice Data Grid)



9

筑波大学 University of Tsukuba

QCDml



10

筑波大学 University of Tsukuba

アンサンブルXML（抜粋）

```
<markovChain xmlns="…">
<markovChainURI>mc://JLDG/CP-PACS/RCNF2/RC12x24-
B1800K014090C1600</markovChainURI>
<management>
<revisions>1</revisions>
<collaboration>CP-PACS</collaboration>
<projectName>RCNF2 (Nf=2 full QCD with iwasaki RG gauge and
tadpole improved clover quark action)</projectName>
<ensembleLabel>B1800</ensembleLabel>
<reference>Phys.Rev. D65 (2002) 054505 (hep-lat/0105015), Erratum-
ibid. D67 (2003) 059901</reference>
<archiveHistory>
<elem>
<revision>1</revision>
<revisionAction>add</revisionAction>
<participant>
<name>T.Yoshie</name>
<institution>Center fof Computational Sciences, University of
Tsukuba</institution>
```

11

筑波大学 University of Tsukuba

QCDml

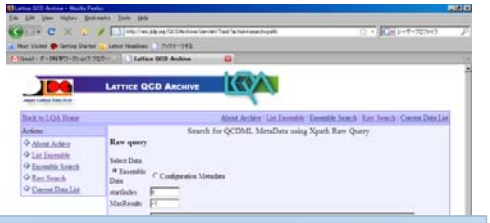
- アンサンブルXML
  - ファイル数：177
  - サイズ：1.1MB
    - 世界6拠点
- コンフィギュレーションXML
  - ファイル数：29,198
  - サイズ：116MB
    - 筑波大学計算科学研究センターのみ

12

筑波大学 University of Tsukuba

現在の検索インタフェース

- Lattice QCD Archive http://www.jldg.org/lqa/
- 検索方法
  - ファイルリスト
  - 問合せ言語
    - XPath
    - XQuery




```
declare default element namespace
       "http://www.lqcd.org/ildg/QCDml/config1.3";
for $i in collection("configurationCon")//gaugeConfiguration
let $lfn := $i/markovStep/dataLFN
where $i//markovChainURI =
       "mc://JLDG/CP-PACS/RCNF2/RC12x24-B1800K014090C1600"
return $lfn
```

13

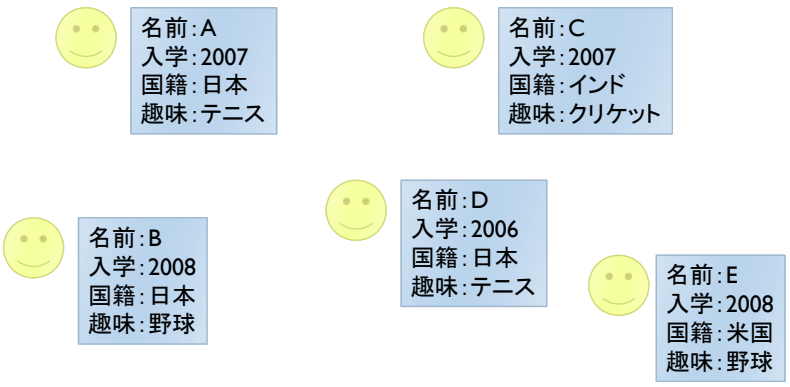
筑波大学 University of Tsukuba

ファセット探索

14

筑波大学 University of Tsukuba

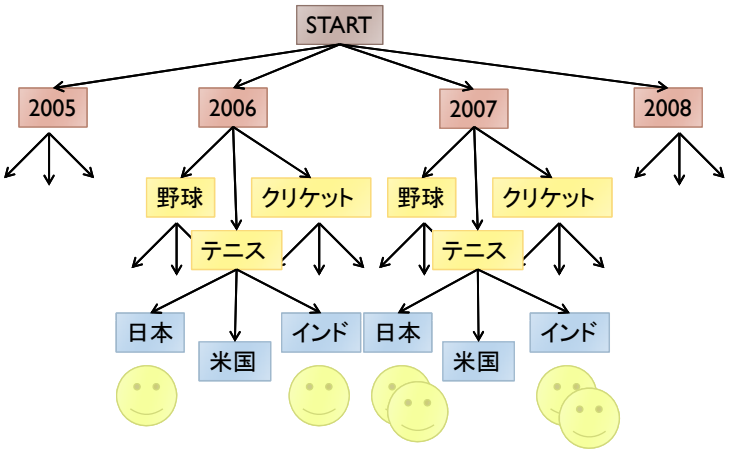
オブジェクト集合



15

筑波大学 University of Tsukuba

階層型分類



17

筑波大学 University of Tsukuba

階層型分類手法の問題点

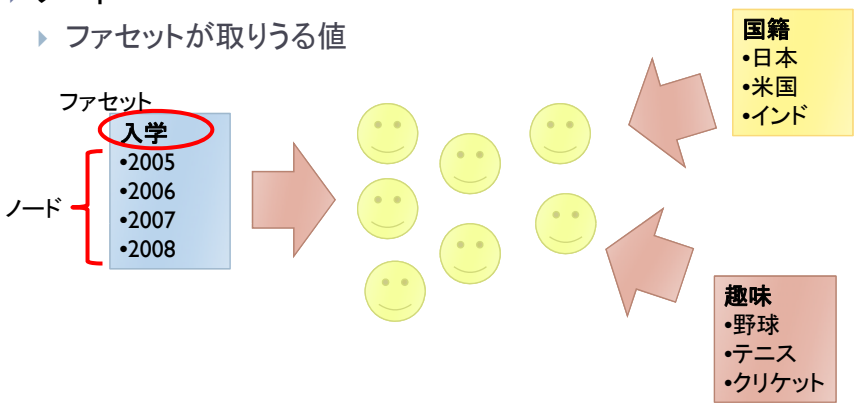
- 分類方法の柔軟性に欠ける
  - 構造があらかじめ決められている
    - 年→趣味→国籍
  - 異なる分類構造 → 作り直し
- 冗長性
  - 階層の深いところで，大量の繰り返し構造が存在
  - 限られた空間で提示できる情報量に限界

18

筑波大学 University of Tsukuba

ファセット

- 独立したカテゴリ
  - 階層あり／なし
- ノード
  - ファセットが取りうる値



19

筑波大学 University of Tsukuba

ファセット探索

1. ファセットを選び，一つ（複数の）ノードを選択
   - オブジェクトの絞り込み
2. 絞り込み条件に応じて，オブジェクトのリストを更新
3. （繰り返し）

- 利点
  - どのファセットを選択するかは，利用者がコントロール可
  - データ管理コストの低減化
    - オブジェクトの追加が大域的な変化を与えない
  - 大量のオブジェクトを効率的に分類
    - Busch's Law
      - 10,000オブジェクトの分類には，10ノードからなる四つのファセットで十分

20

筑波大学 University of Tsukuba

ポイント

- ファセット・代表的な値の一覧を表示
- 現在選択されているオブジェクトの総数を動的に計算

- XMLを扱う際の問題点
  - 検索対象の粒度がまちまち
    - XMLは本質的に木構造
    - どの部分XMLデータを検索したいのか
  - データ構造の規則性
    - 硬い／ゆるいスキーマ
  - ファセット値（ノード）の抽出
    - 部分木／属性

24

筑波大学 University of Tsukuba

ファセットの決定

- 検索対象要素からの相対パス（問合せ）で指定
- QCDmlの場合
  - markovChainURI配下の情報が候補
    - コラボレーション
    - プロジェクト名
    - 実験パラメータ
      - 格子サイズ
      - Gluonアクション
      - Fermionアクション
    - 更新日時

25

筑波大学 University of Tsukuba

QCDmlのファセット リテラルを持つ要素

- 値をそのまま用いる
  - コラボレーション (collaboration)
  - プロジェクト名 (projectName)
- 値の加工が必要
  - 登録日 (date)
    - 年
    - 年-月
    - 年-月-日

CP-PACS
CP-PACS+JLQCD
CSSM
LHPC
MILC
RBC-UKQCD
UKQCD
dik
etmc
gral
qcdsf
sesam
theta
txl
…

2+1 DWF
2+1 Dynamical AsqTAD
Baryon Resonances
Dynamical FLIC Studies
Electromagnetic Form Factors
FLIC Overlap Studies
Flux Tube Test
Gluon Propagator
Long_aqstad_run
Pentaquark Volume Dependence
…

2000
2005
2006
2007
2008

26

QCDmlのファセット
子要素を持つ要素

筑波大学 University of Tsukuba

- 例：格子サイズ
  - どのように見せるかは応用依存
- 典型的なパターン
  - テキストのみを連結
    - X10Y10Z10T32
  - 特定のテキストを列挙
    - 10 10 10 32
    - 10 / 10 / 10 / 32

```
<physics>
  <size>
    <elem>
      <name>X</name>
      <length>12</length>
    </elem>
    <elem>
      <name>Y</name>
      <length>12</length>
    </elem>
    <elem>
      <name>Z</name>
      <length>12</length>
    </elem>
    <elem>
      <name>T</name>
      <length>24</length>
    </elem>
    …
```

27

QCDmlのファセット
要素名自身がファセット値

筑波大学 University of Tsukuba

- gluonAction / fermionAction

```
<action>
  <gluon>
    <iwasakiRGGluonAction>
      <glossary> http://www.jldg.org/JLDG/...
```

```
<action>
  <gluon>
    <DBW2GluonAction>
      <glossary> www.iqcd.org/ildg/pla...
```

- ファセット値の抽出の際，
  - テキスト値（属性値）
  - 要素名（属性名）

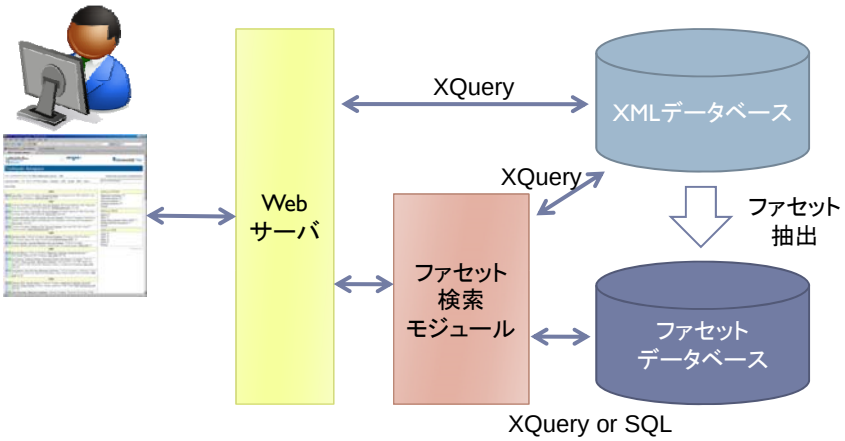
  をケア

28

筑波大学 University of Tsukuba

システム構成

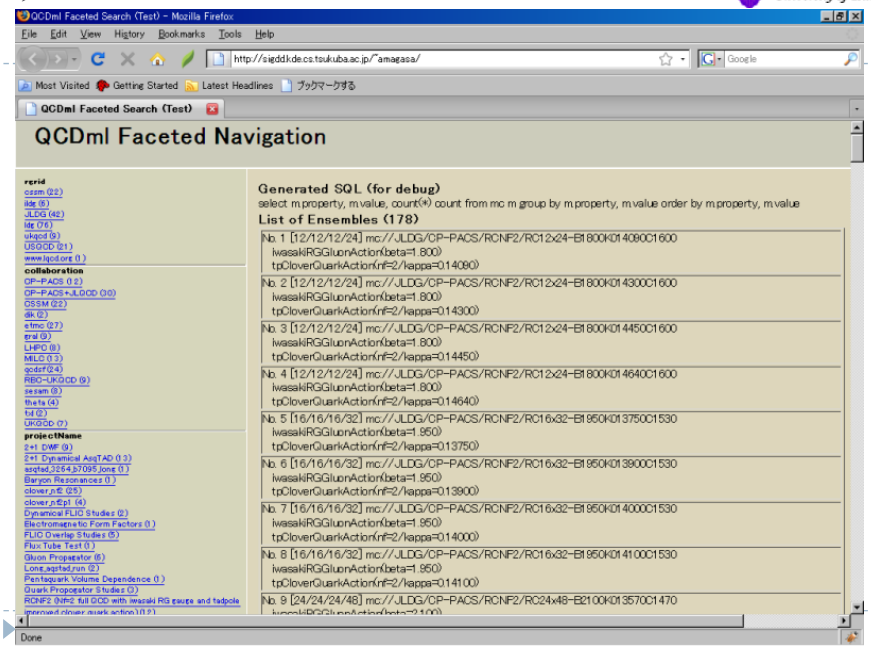
- ファセットはあらかじめ抽出
- RDBMSで集約計算



29

デモ

筑波大学 University of Tsukuba



筑波大学 University of Tsukuba

まとめと今後の課題

- XMLメタデータの探索インタフェース
- 素粒子の専門家には大変好評

- 今後について
  - 一般のXMLデータ上にファセット検索インタフェースを構築するためのフレームワーク作成
  - ほぼ完成
  - 他分野のXMLデータへの適用

31

筑波大学 University of Tsukuba

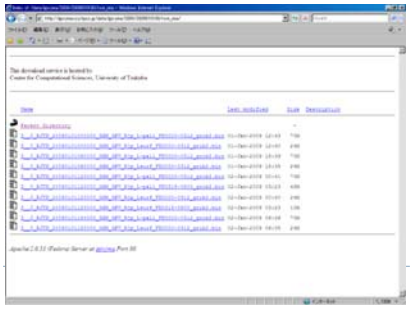
FUSEによる遠隔気象データアクセスミドルウェア

32

筑波大学 University of Tsukuba

気象分野の研究業務

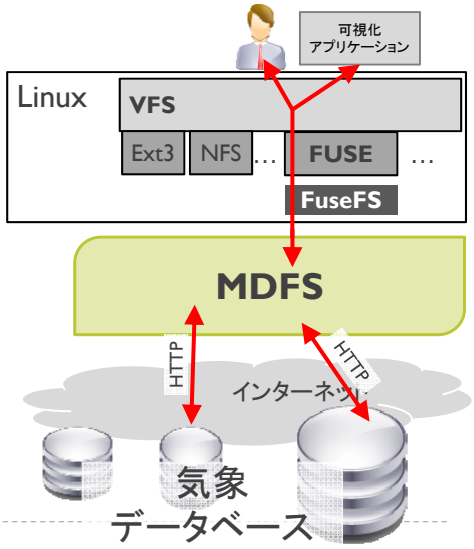
- データの検索
  - Webブラウザ＋フォーム
- データの取得
  - 個別のファイルをダウンロード
- データの加工・レンダリング
  - ローカルファイルに対するプログラムの実行



筑波大学 University of Tsukuba

アイデア

- 計算機に詳しくない気象分野の研究者の気象データアクセス支援
  - FUSE (Filesystem in Userspace) を利用
  - Webサーバ上のファイルをローカルファイルシステムにマッピング
  - 既存のプログラムを直接実行可能



34

筑波大学 University of Tsukuba

既存のアプローチに対する位置付け

	既存	提案
データの検索	ブラウザ＋フォーム	UNIXコマンド (ls, find, …)
データの取得	ブラウザ, wget, …	不要
データ処理・レンダリング	ローカルファイルに対してプログラムを実行	リモートファイルに対して，直接実行（キャッシュ有）

- ワークフロー
  - スクリプト(sh, Perl, Ruby, …)で記述可能

36

筑波大学 University of Tsukuba

関連アプローチとの比較

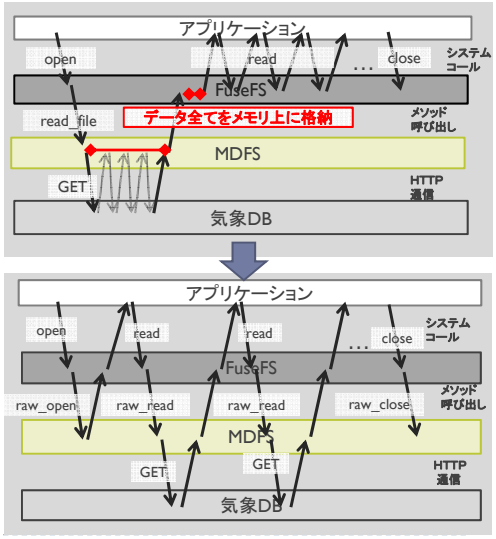
- グリッドファイルシステム・広域分散ファイルシステム
  - Chord, Gfarm-FUSE, …
  - サーバ・クライアント双方に専用ソフトウェアのインストールが必要
- OPeNDAP
  - ネットワークデータアクセスプロトコル
  - ローカルデータアクセスプログラムをネットワーク透過に
  - 専用サーバ＋クライアントライブラリ
- FUSEによるアプローチ
  - サーバ: Webサーバ
  - クライアント: 既存クライアント

37

筑波大学 University of Tsukuba

実装上の工夫 部分データアクセス

- データアクセスの局所性
  - ファイルを仮想的なブロックに分割
  - アクセスのあったブロック単位にデータを取得
  - 取得したデータはキャッシュに保存
    - 2度目以降のアクセスを高速に
  - 非同期アクセスによるブロック取得



38

筑波大学 University of Tsukuba

今後の予定

- アプリケーション固有のアクセスパタンを利用したアクセスの効率化
  - アクセスログ
  - シーケンスマイニングを利用した，アクセスパタンの抽出
  - 先読み・キャッシュ置換アルゴリズムへの組み込み
- Gfdnaviとの連携
  - 気象データに特化したクローリング
  - インターネット上の気象データポータルの半自動構築

39