格子データ諸形式の比較
データモデルの相違など

2013-10-28

数値予報課データ形式勉強会
豊田英司

用語について

- なるべく実用を心がけました
  - 「データ」は通報式公定訳では「資料」だけど、さすがに無理なのでデータで通してます
- でもわからないかも
  - 記録 = record
  - GRIB 報 → GRIB message
  - オクテット = 8 ビット = バイト (ここでは)
  - #2/10 = number 2 out of 10 と読みます

本日のお話

- データ変換をやりたい
  - NuSDaS（たいてい変換元）
  - GrADS binary
  - GRIB Edition 1/2
  - netCDF
- そもそも本質的に何が違うか
  - 構造（データモデル）
  - 語彙（符号）とその管理

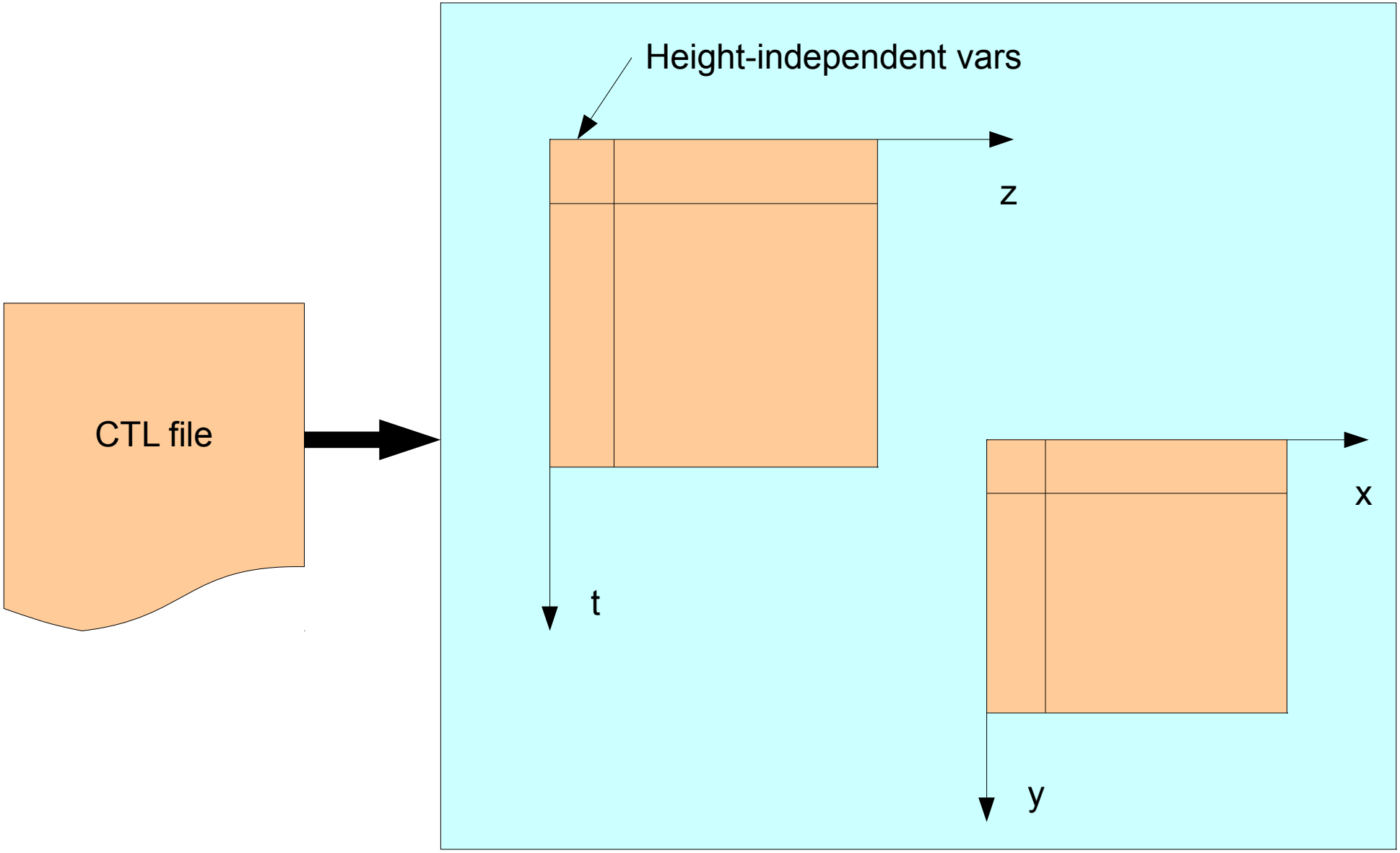
話の進め方

- 配慮
  - 項目ごと比較じゃ頭に入らんだろう
  - 形式ごと説明 ×5 連続じゃくどいだろう
- まず単純と複雑の極致をひととおり説明
- 他の形式の説明を差分・総覧的に

単純の極致： GrADS バイナリ

- 基本的に 1 データセットは単精度実数 5 次元配列
  - 時間 × 変数 × 高度 × 水平 2 次元
  - 一部の変数は高度なしにできる (地表面の量とか)
- 水平 2 次元は原則として経緯度
  - 地図投影面上の矩形格子もできなくはないが稀
- 高度はなんでもいいから数値
- 変数名は 8 文字。自由
- 時間はひとつだけ、等間隔であること

GrADS バイナリの構造



世の中 GrADS しかないならば、何も考えることはないんだが

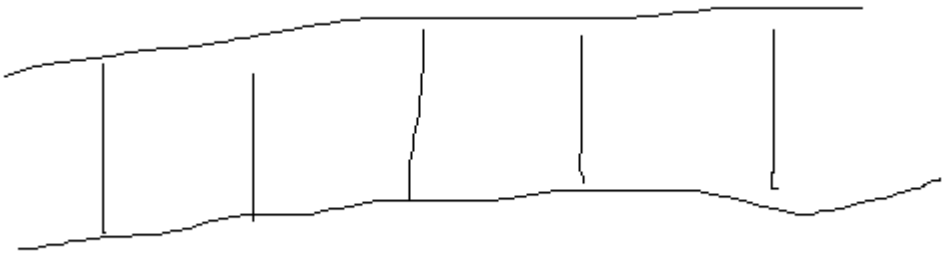
- もちろんそうは問屋が卸さない
- それぞれの必要に応じてそれぞれの複雑さが育ってきたのである
- 同じ気象（あるいは地球科学）だから全部共通というわけでもない

いちばん複雑： NuSDaS

- 数値予報データセットシステム
- Numerical Weather Prediction Standard Dataset System であるらしい
- 主に格子データを扱うもの
- 気象庁内で作られた
- データセットシステムとは、現代的にいえばデータベースなのだろう

データベースとは（俺定義）

- 大きな情報を何らかの方法で分節し
- その一部を読み書きできるようにする
- ソフトウェアまたはサービス
  - あるいは文系的には資料収集事業なども含む

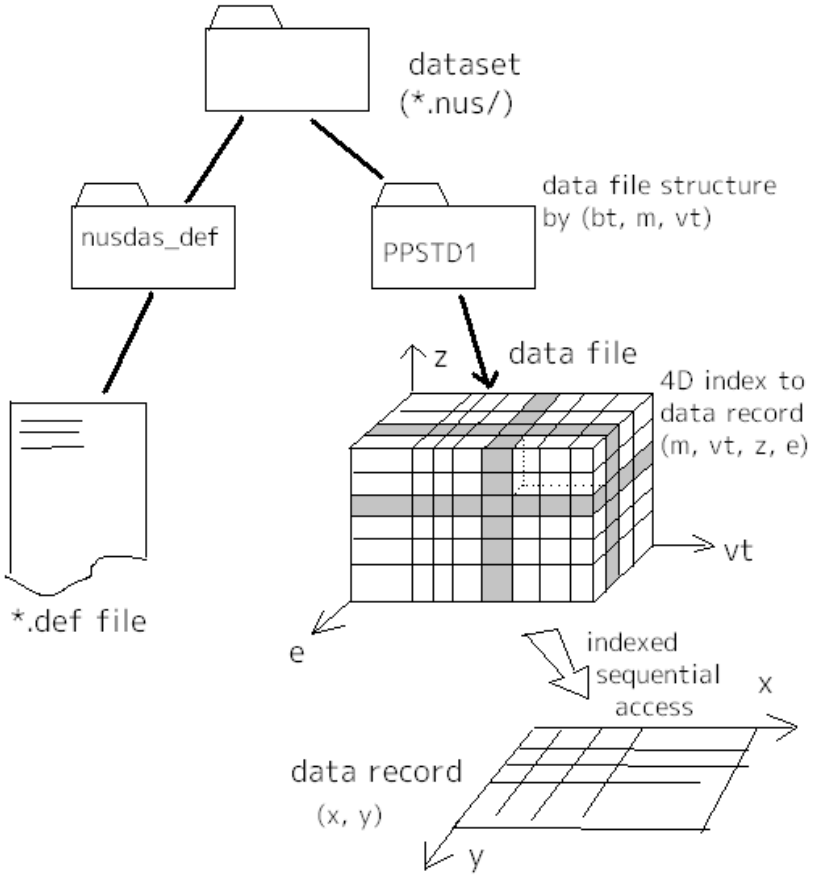


分節（イメージ）

NuSDaS はどうデータを分節するか

- データ記録：読み書きの単位
  - 2 次元配列 (ふつう東西･ 南北)
- データファイル
  - 要素、(鉛直) 面、対象時刻 (valid time) 、(アンサンブル) メンバーが異なるデータレコード群
- データセット
  - 対象時刻、メンバー、基準時刻 (reference time) が異なるデータファイル群
  - 種別名で識別される

NuSDaS データ格納構造



- データセット内ではファイル名でデータファイルを探す
- データファイル内では先頭の４次元配列のインデックスで記録を探す
- データ記録内は添字で探索

配列添字になれないものを強引に添字にするために

- 配列添字は連続した整数でなければならない
  - ちょっとくらいは飛んでいてもいいけど
- ところがファイル中のインデックスは
  - 文字列：メンバー、面、要素
  - すごく飛ぶ整数：対象時刻 (= 基準時刻 + 予報時間)
- 先頭部分の CNTL 記録にリストを置く
  - 文字列または予報時間→順番を得る

NuSDaS 定義ファイルの役割

- データファイル名の規則
  - パス←(基準時刻、メンバ、対象時刻)
  - 後からでも変更可能 (直接やればできる)
- データファイル作成時の初期値
  - リスト (メンバ、予報時間、面、要素)
  - その他情報 (格子配置など)
  - 一度作成されたデータファイルのリストに入っていなかった予報時間・要素などは追加できない

NuSDaS の語彙管理

- 数値イントラウェブサイトで登録
  http://nusdas.npd.naps.kishou.go.jp/
  - 本庁外で使うときはどうするんだろうね？
- 半管理：種別名、要素
  - 数値予報ルーチンでは登録済名前だけ使用可
- 管理なし：メンバー、面
  - あまり激しく変わらない、はず、だから

種別名の内部構造

- 例： _GSMLLPP.FCSV.ASIA

例	例の意味	正式名称	注記
_GSM	GSM	モデル名	変換で使用
LL	経緯度格子	二次元座標名	変換で使用、一部要 SUBC データファイル作成時に地図投影法パラメタをチェック
PP	気圧座標	三次元座標名	変換で使用、一部要 SUBC
FC	予報値	属性名	変換で使用
SV	瞬間値	時間種類名	変換で使用（積算…）
ASIA	アジア域	種別3	単なる名前

その他の名前

- メンバ名
  - 00: コントロール、 01p, 01m: 第 1 摂動正負
- 面名
  - 地表面 : SURF; 等圧面 :1000, 500 等 ; モデル面 : 1, 2, 3, ...; FL: F290 等 ; エコー頂 : ECTOP; ...
- 要素名
  - T: 気温、 Z: ジオポテンシャル、 P: 気圧、 PSEA: 海面更正気圧、 ...

GSM p 面予報値は、きれいな例

- 四角い 2 次元配列のデータ記録が
- 一見してわかりやすい分類で並び
- データファイル内の記録の種類数は最小
  - NUSD
  - CNTL　メタデータ (名前リスト、投影法情報など)
  - INDY　インデックス
  - DATA × くりかえし
  - END

一部データでは SUBC 記録が必要

- CNTL 記録 (固定形式) には書ききれないメタデータを格納
  - たいていは配列だけど、データ記録とは形が違う
- 鉛直座標情報 : SIGM, ETA, ZHYB
- 水平格子系情報 : RGAU
- 時間関係の情報 : TDIF, DELT
- その他 : SRF

水平矩形格子でないデータ記録

- Reduced Gauss 格子 (RG)
  - x 軸は単に順番だけ、 y 軸は 1 格子しかない
  - SUBC RGAU が各緯度の格子数を与える
- アメダス地点 (ST)
  - 要素 NUM, LON, LAT が地点番号・経緯度
  - 付随情報 (AQC) は種別 _DCD→_AQC
- 帯状平均子午面断面データ (YP)
  - x 軸は緯度、 y 軸は気圧座標
  - 面名 ZONAL

さまよえる時間積算

- NAPS7 当初規則
  - 時間種類名を変える：瞬間値 SV→ 積算値 AV
  - Validtime1, validtime2 の範囲で積算期間を指定
  - データセットを分けるのが面倒なので実効性なし
- 現行マニュアル
  - 積算値なのに FCSV のままでいいや
  - SUBC TDIF に積算範囲を書き込む
  - 要素名に "." を前置して積算を指示 (守ってる？)
- 現実 ...

データ表現についても少しだけ

- Packing　数値列のバイト列での表現
  - R4　　単精度実数列をそのまま格納
  - 2UPC　線形変換で 16 ビット整数列として表現
  - 2UPJ　　〃　　を JPEG2000 で圧縮
  - RLEN　１バイト数値列をランレングス圧縮
- Missing　欠損値の表現
  - NONE　欠損値なし
  - UDFV　１つの値を欠損値と指定
  - MASK　ビットマスク (場所指定)

NuSDaS はこのへんにして、
次は GRIB の話

GRIB とは

- WMO Manual on Codes に定める通報式

  http://www.wmo.int/pages/prog/www/WMOCodes.html

  - GTS での国際交換のために作られたもの
- Edition 1 と Edition 2 がある
  - 似ているけど、互換性はほとんどない
  - 新規プロダクトは原則 GRIB2 でお願いします
  - Edition 3 はまだ何年も後の話

GRIB2 のデータモデル

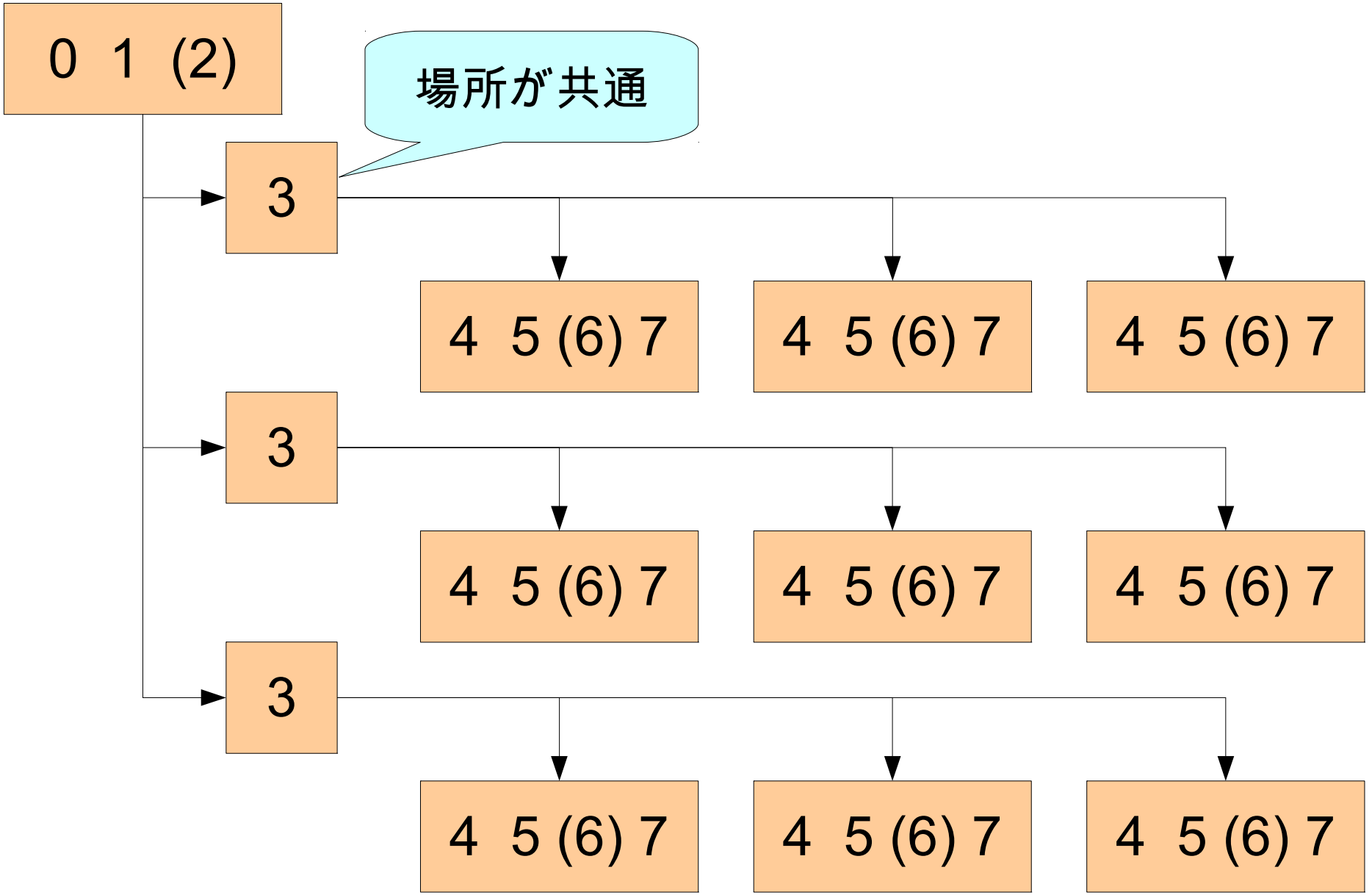
- 1 つの GRIB 報が任意個の 2 次元配列の集まり
- NuSDaS データファイルにちょっと似ている
  - 話は逆で、 NuSDaS が GRIB2 に倣って作られた
- でもね、ちょっと違うんですよ
  - 構造から話したほうがいいでしょう

GRIB2 の構造

- 第 0 節：マジックナンバー等
- 第 1 節：作成中枢、参照時刻など
- （第 2 節：作成中枢独自定義）
- 第 3 節：格子系定義節　　水平 2 次元格子配置
- 第 4 節：プロダクト定義節　　予報時間、要素、統計、高度等
- 第 5 節：データ表現節　　packing/missing 方式指定
- (第 6 節：ビットマップ節–欠損を表現する場合)
- 第 7 節：データ節

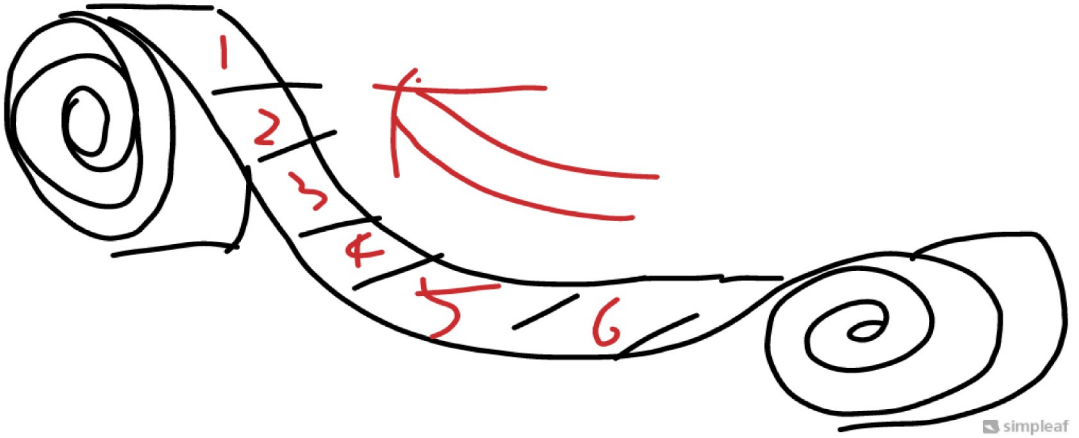
第 2~7, 3~7 又は 4~7 節を繰り返せる

GRIB2 の節の繰り返し



GRIB2 の節構造は一種のシーケンシャルファイル

- 第 1~7 節の先頭
  - オクテット 1~4: 節の長さ
  - オクテット 5: 節の番号 (第 1 節なら 1)
- 不定長 = 先頭から読み進んでいくほかない
  - 欲しい 500 hPa 面気温が 24 個目の第 4 節に書いてあることを知っていても、何バイト目にあるかは未知



データベース構造：
３つのうち２つをお選びください

構造	書くとき	ファイル長	読むとき	今回の例
Sequential	先着順：**単純**	**コンパクト**	順番に読む：**遅い**	GRIB2
Direct	記録番号から定まる位置に出力：**単純**	**無駄**に大きくなる	記録番号から定まる位置から直接：**速い**	GrADS-binary netCDF
Indexed Sequential Access Method	先着順、ただし位置をインデックスに保存：**複雑**	**まあコンパクト**：インデックスが過大でなければ	インデックスから得られる位置から直接：**まあ速い**	NuSDaS

個々の記録は何により特定されるか

- ダイレクトアクセス
  - (ft=24h step 6h→#5) & ("Z" → #1/6)
    & (500hPa → #5/8)
    → 記録 #197→ 記録長をかけて位置を得る
- インデックス付シーケンシャル
  - (ft=24h, "Z", 500hPa) →9057434 オクテット
- 単なるシーケンシャル
  - (予報時間 , 要素 , 高度) だけで記録が特定されるかどうかは全部読んでみないとわからない
  - メンバー、格子系、統計処理などが違うかも：自由
  - 普通はなんらかのデータ依存の仮定をもって読む

自由はこわい

（京大総人酒井敏教授の金言）

気をとりなおして、第１節の構造

- 6-7 作成中枢
- 8-9 作成副中枢
- 10 マスター表版番号
- 11 ローカル表版番号
- 12 参照時刻意味
- 13-14 年（参照時刻）
- 15~19 月日時分秒
- 20 運用・試験
- 21 データの種類

観察：

- 固定長
- 文字型は原則排除→符号表を参照
  - 表参照通報式というゆえん
  - 自己記述的でないとよく言われる

自由にプロダクトを作るには、東京 (34) の副中枢として登録を

- 気候情報課再解析班はやってくれました
- ローカルのパラメタ番号が合法的に作れます
- WMO マニュアル改正（年２回）には最短で半年かかります
- ご相談は情報通信課国際班へ
  （本来はあたしじゃないよ！）

いちいち通報式改正依頼するの面倒

- まあ、面倒なのはわかりますが
- 形式・内容についてあらかじめ合意するために不可欠
  - 計画的な運営の考え方がなくならない限り、どっかで調整はしなければならない
- 通報式の管理をしていると、
  全世界の測器・モデル開発の情報が集まる

テンプレート構造

- 第３節（格子系定義節）、
  第４節（プロダクト定義節）、
  第５節（データ表現節）には多様な種類がある
- 最初のほうにあるテンプレート番号で後のほうの構造が変わる
- 紙の帳票書式を選択して埋めていくイメージ

第３節（格子系）はわりと簡単

- 0 経緯度
- 3 斜軸伸長経緯度
- 10 メルカトル
- 20 ポーラーステレオ
- 30 ランベルト正角円錐
- 40 ガウス格子
- 50 球面調和係数
- 1000 断面

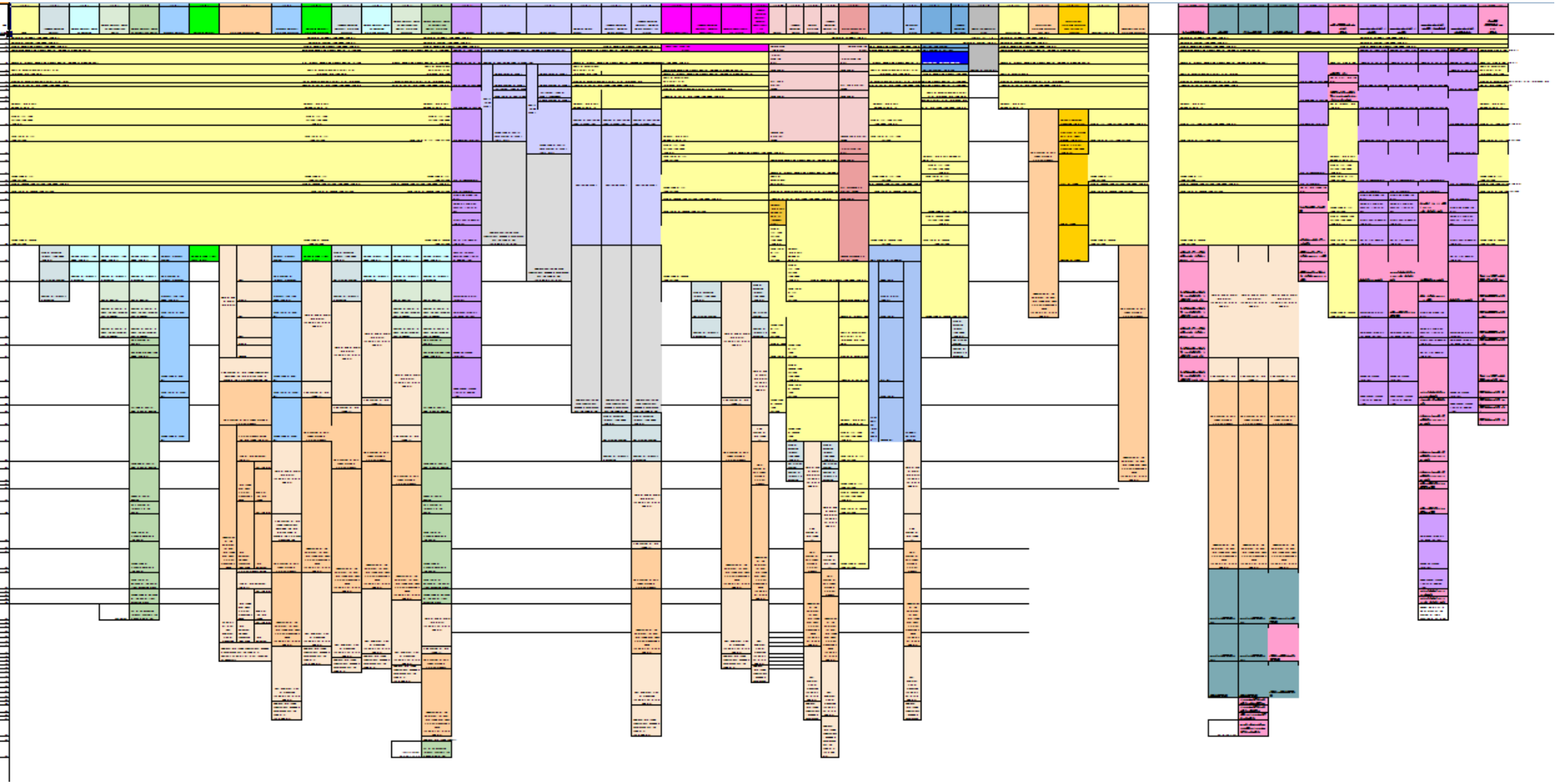
- テンプレートの後に thinned grid 用の情報（各行の格子数）を置くことができる
- 経緯度以外で thinned はあまりやらないが、美しい設計ではある

第 4 節 (プロダクト) は混沌

- 0 標準
- 1 アンサンブル単予報
- 2 アンサンブル統計量
- 8 標準 + 時間統計
- 20 レーダー
- 30, 31 衛星 (旧新)
- 32 擬似衛星
- 40 標準 + 化学物質番号
- 1000-1002 断面 (標準、時間統計、空間統計)

- 性質の違うメタデータの組み合わせが次々と登録されている
  - 統計 (アンサンブル、時間、空間)
  - 測器情報
  - 化学物質のパラメタ番号不足を補う情報
- テンプレートの後に鉛直座標パラメタを置ける

プロダクト定義テンプレート全図



時間積算の蹉跌

- ちょっと難解だけど壮麗な体系
  - 時間統計を含むテンプレートでは
    任意個の統計処理演算子を繰りかえして指定できる
  - 日最大の月平均の平年値、とか書ける
- パラメタ名は時間統計や鉛直位置情報を含んではいけないことにした
  - 後で規則化
  - 降水量など積分量のパラメタが廃止予定に
- しかし、バグがあった
  - 時間積算（統計処理コード１）を使った時にパラメタ表の単位に秒をかけるという規定を忘れていた！

降水強度など微分量を使え

1 時間降水量の表現法

大混乱！！

- 積算値派（ GRIB1 互換）
  - 0-1-8（降水量 kg/m2 ）、統計処理テンプレート不使用、格納データは初期値からの積算値
- 積分量派（ GSM 予報値）
  - 0-1-8 、統計処理テンプレートで 3600 秒積算を明示、データは 1 時間の積算値（ kg/m2 ）
- 微分量派（ ECMWF, MSM 予報値）
  - 0-1-52（降水強度 kg/m2/s ）、統計処理テンプレートで 3600 秒積算を明示、データは 1 時間の積算値 (kg/m2)
- DWD 解釈
  - データの単位は規則どおり kg/m2/s でなければならないので、これに 3600 をかけたものが 1 時間降水量

けっきょくどう収めたか

- まず「解釈不定は通報式の存在理由に反する」、「通報式の文書がどんなに見苦しくなっても、既存のプロダクトを合法化する」に合意させる
- 本則：統計処理コード 1（積算）を用いた場合は、符号表 4.2（パラメタ）の単位に秒をかける
  - 微分派 ECMWF 解釈の一応勝利
- 例外：ただし、符号表 4.2 に注記した一部パラメタについては、この処理をしない
  - 積分派プロダクトの合法化

マニュアル改正
2012 年 11 月
発効

NuSDaS 要素と GRIB2 パラメタ

- 1 対 1 対応しているわけではない
  - 単位が違うこともある
  - 風速の方向
  - 雲水など分類の仕方が違うこともある
- 時間統計や高度を含むパラメタは今でも追加できない
  - さっきの話：同じ NuSDaS から複数の GRIB2 表現
  - RAIN, RR10, RR60 等→同じ GRIB2 に

最後に GRIB1 について

- 節の反復なし

  → 自分で DB 構造を考える必要がある

  支援センターはシーケンシャル、 MARS は IBM DB2

- 格子系テンプレートは似た仕組みがある
- プロダクト定義はテンプレートなし

  → アンサンブル等表現不可
  時間情報は 2 パラメタしかない

- パラメタ番号が 1 バイト

  → すぐ枯渇、ローカル表の乱立

- なぜかファンが多いですが、もう使わないでください

  (そもそも君、センター番号持ってるの？)

そろそろ netCDF にいきましょうか

netCDF とは

- UCAR Unidata で開発されたデータ形式
  - NASA ゴダードの CDF を真似た (互換性なし)
- 標準実装がある (C, Fortran, Java)
  - 標準ダンプツールもついてくる
- バージョン
  - 気象界では簡明な Ver 3 が普及
  - HDF ベースで多機能な Ver 4 が登場 (API 互換)

netCDF v3 のデータモデル

- データセットは任意個の変数からなる
- **変数**は任意個の次元をもつ配列
  - 型 : char, byte, short, int, float, double
- **次元**は固定長
  - 一度作ったら長さを伸ばすことはできない
  - ただし 1 つだけは可変長にできる
- データセットまたは変数には**属性**をつけられる
  - 上記型いずれかの 1 次元配列
  - たいていは char 配列を文字列として使う

なぜ可変長次元は１つだけ？

- NetCDF v3 はダイレクトアクセスファイル
  - ほぼ固定長部分
    - 次元名、変数名の表など
    - 属性名と属性値のペアの羅列
    - 固定長の**変数値**
  - 可変長次元の添字１つひとつについて
    - 可変長次元をもつ**変数値**の当該添字に対応する部分
- 変数値出力開始後にできないこと
  - 次元･変数の加除
  - 属性値のバイト数を増やす書換え

netCDF : 配列データ
= XML : 木構造データ

- NetCDF というだけではほとんど何も決まらない
  - 変数 (配列データ) と、属性 (付随する情報) の表現法
  - 規約 (何の情報をどういう変数名、属性名で格納するかの約束) が必要
- XML というだけではほとんど何も決まらない
  - エクセル、 HTML 、 RSS も防災 XML もみな XML
  - 要素 (木構造のノード) と属性の表現法
  - スキーマ (何の情報をどういう要素名、属性名で格納するかの約束) が必要

規約の決定版： CF Conventions

http://cf-pcmdi.llnl.gov/

古い翻訳
https://www.gfd-dennou.org/arch/netcdf/cf-conventions-ja/cf-ja.html

- CF = Climate and Forecast
  - あまりにも成功したので事実上唯一の規約化
  - Open Geospatial Consortium 標準にもなった
- 成功の鍵は公開管理
  - 誰でも改正を発議可能（ ML, Trac ）
  - 議論経緯は完全公開

netCDF/CF の思想

- 自己記述性
  - 外部の表の参照は最小限に止める
- テキスト主義
  - GRIB で符号表になる情報は、だいたい文字列の属性で表現される
- 次元の平等
  - 統計処理など、どの次元か (時間、鉛直、水平次元など) で特別な扱いをしない
  - 一応、時空間 4 次元を特定する方法は標準化

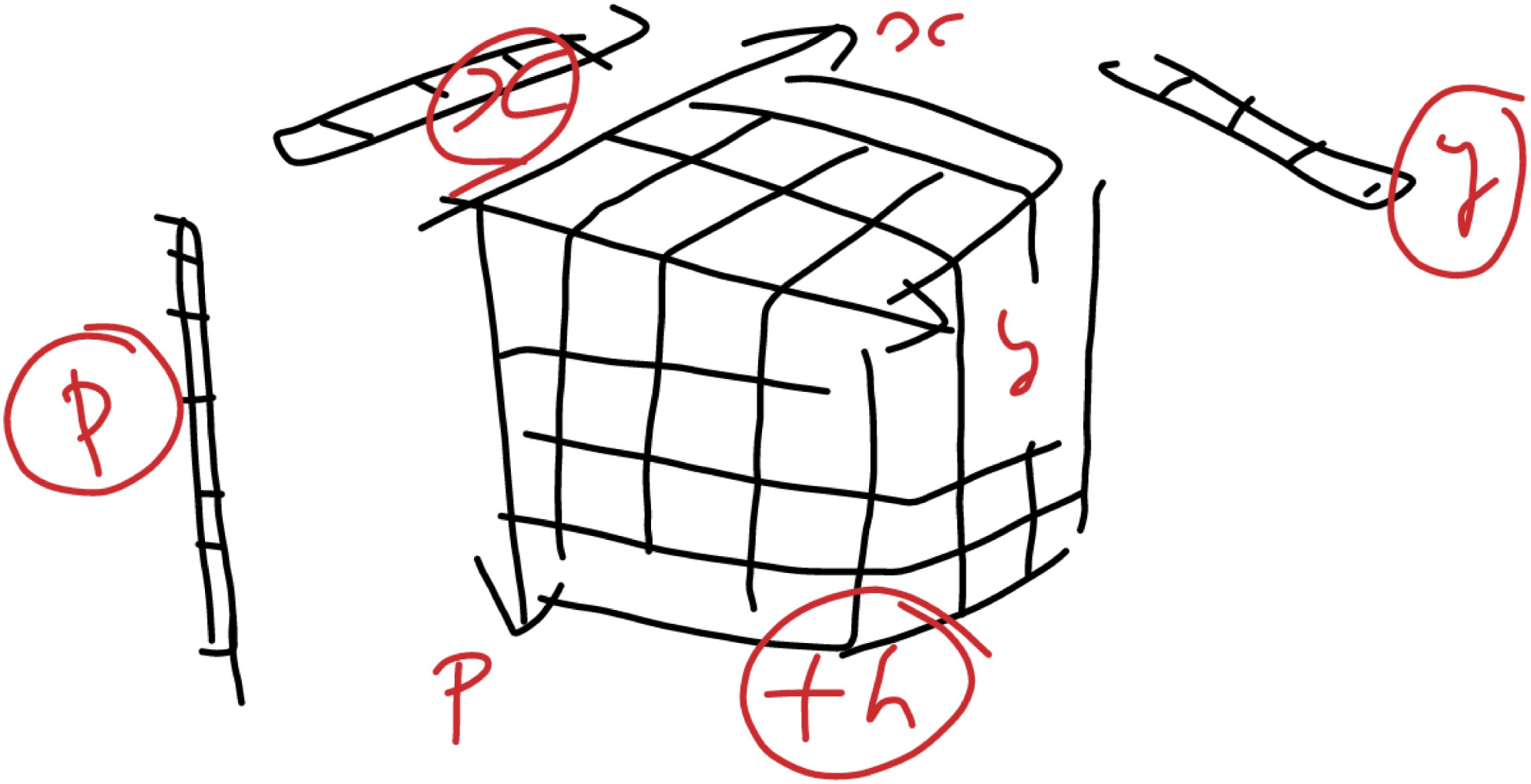
そうは言っても外部表はある

- 属性 standard_name
  - 物理量の表
  - 頻繁に追加が議論・実施されている
- 属性 units – udunits.txt
  - 追加要望が時折あるが、本来 Unidata 管理というと誰も要望しないので反映されたためしがない
- 海域・大陸名 : NASA GCMD 由来
  - ほとんど注目されていない
  - "sea_of_japan" あり

比較的単純な例： p 面予報値

```
dimensions:
  x = 143;  y = 72;  p = 8;  vt = 7;
variables:
  float th(vt, p, y, x);
    th:long_name = "potential temperature" ;
    th:units = "K" ;
    th:standard_name = "air_potential_temperature";
  float p(p);
    p:long_name = "pressure" ;
    p:standard_name = "air_pressure" ;
    p:units = "hPa" ;
  float vt(vt);
    vt:long_name = "valid time" ;
    vt:standard_name = "time" ;
    vt:units = "hours since 2013-10-28T12:00:00Z" ;
```

データも、座標値も、おなじ変数
(次元と同じ名前の変数が座標値)



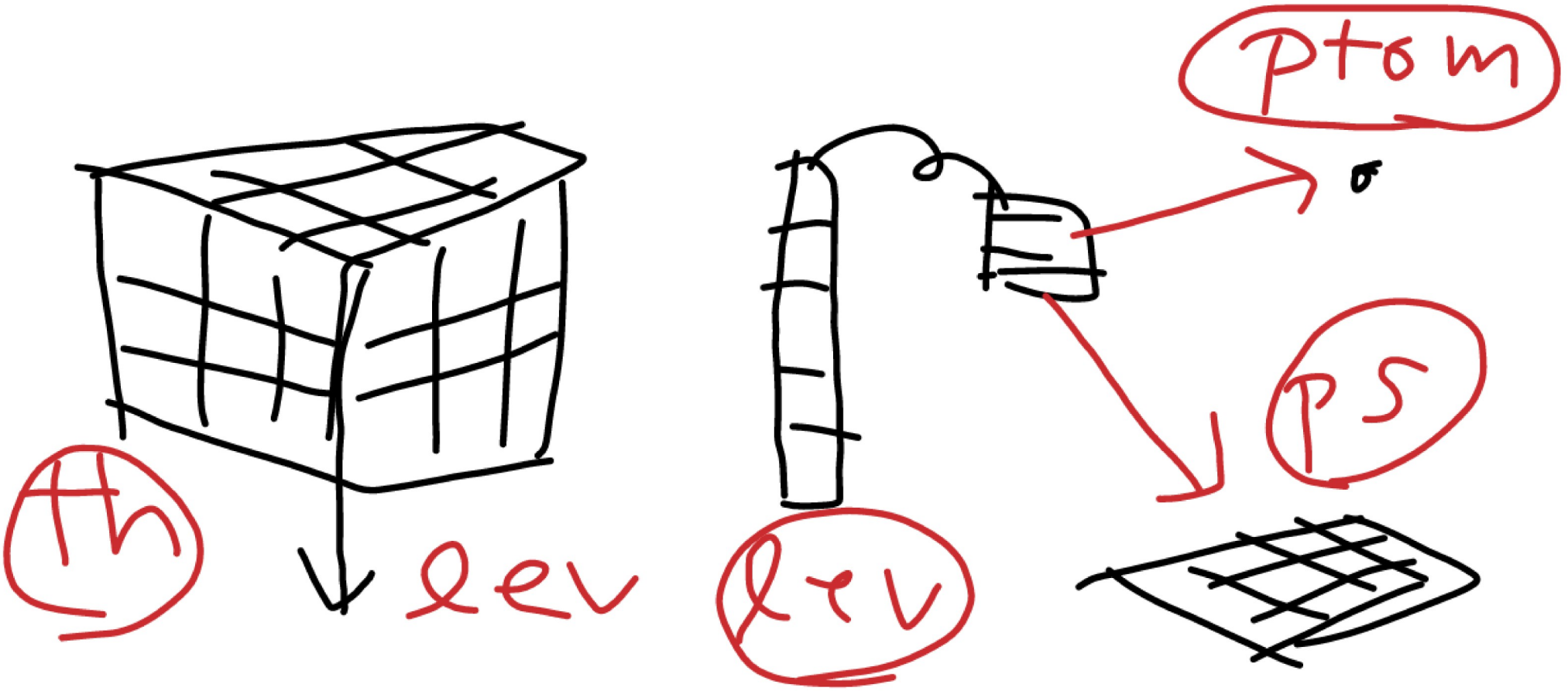
軸の付随情報： シグマ座標

```
dimensions:
  sigma = 16;
variables:
  float th(vt, lev, y, x);
    ...
  float lev(lev);
    lev:long_name = "sigma at layer midpoints" ;
    lev:positive = "down";
    lev:standard_name = "atmosphere_sigma_coordinate";
    lev:formula_terms = "sigma: lev ps: ps ptop: ptom";

  float ptom;
    ptom:long_name = "model top pressure";
    ptom:units = "hPa";
    ptom:standard_name = "air_pressure";

  float ps(vt, y, x);
```

シグマ座標の付随情報は別の変数に



CF-netCDF v3 で

- 便利
  - 新しい物理量を登録する
- 面倒
  - 地図投影面上格子に lat/lon データをつける
- できない
  - １６ビットパックで実用的精度を出す
  - ランレングス圧縮など
- できるけど NuSDaS に変換しにくい
  - （時空間じゃない）特別な座標をもつ配列
  - ずれた格子（ Arakawa C Grid ）
  - 未知のモデル、要素、単位など自由文での入電

まとめ

- GrADS バイナリは単純
- NuSDaS はわりと複雑
- GRIB2 は帳票。番号割り当て待ちが面倒
- GRIB1 は忘れてほしいんだけど
- CF-NetCDF は言語。配列の組み合わせで表現

  理屈は美しいが、ソフトウェアはちょっと複雑