大規模地球環境観測データを対象としたデータクオリティコントロールシステムの構築とその有効性の検討

Development of Data Quality Control System for Huge Scale Earth Environmental Observation Data and Study of its Effectivity

生駒 栄司♣ 玉川 勝徳♦
小池 俊雄◆ 喜連川 優♠

Eiji IKOMA Katsunori TAMAGAWA
Toshio KOIKE Masaru KITSUREGAWA

昨今の観測技術の進歩により,取得される地球環境データの量は爆発的に増大しつつある一方,そのデータのクオリティを維持するための作業量も膨大なものとなっている.一般的に地球環境に関する観測においては,その観測技術の特殊性,観測機器の多様性のため,取得されたデータの妥当性の検証は実際に観測を行った研究者にしか満足に行うことが出来ない場合が多い.反面,観測分野の研究者は必ずしもコンピュータの操作に習熟しているとは限らず,汎用的なパソコン用ソフトウェアを用いても膨大な労力を費やして検証を行っているのが現状である.そこで本論文では,実際に世界中で観測されたデータを対象とし,容易な操作で効率的なデータ検証と修正が可能なデータクオリティコントロールシステムの開発を行い,実際の観測者と協力してその有効性の検証を行った.

The amount of earth environmental data has become increased explosively because of the advance of recent observation technique. On the other hand, it is true that the quantity of work to keep the quality of data has also become increased. In general, about the earth observation research field, only the observer can validate the evidence of the adequacy of acquired data because the particularity of observed technique and the variety of observation equipment. However, most researchers belonging to earth observation field are not specialists about computer. Actually, most of them are using general PC software and spend much labor to validate the evidence. In this research, we collaborate deeply with some researchers of the field, and develop a support system for checking the quality of data with easy operation targeting the actual data observed from various site of the world. Moreover, we also validate the evidence of the availability of our system with them.

1. はじめに

昨今の地球環境への関心の高まりとともに地球観測に関する技術も劇的に進歩し,各種地上観測機器から得られるポイントデータ等はその質・量共に大幅に向上しつつある.その結果,従来は取得が困難であったさまざまな次元におけるより詳細かつ有用なデータが容易に取得出来るようになり,当該分野の研究の発展に大きく貢献しつつある.

反面,そのデータ量が膨大であるが故に,取得されたデータの妥当性を十分に検討することが出来ず,品質の悪いデータが流通するといった問題も起こっている.

一般に地球環境に関する観測データの場合,その取得手法の複雑さ故にデータの妥当性の検証は実際にそのデータを取得した観測者しか出来ない場合が多い.すなわち,取得されるデータ量は増加の一途を辿っているが,その妥当性を検証する研究者はさほど増加せず,その検証手法に進歩がなければ結果として十分に検証出来ない事態が起こりうるのである.そこで本研究では,上述のような地球環境に関するデータを観測している研究グループと協力し,実観測データを対象としたデータクオリティコントロールシステムの開発を行った.本稿では,実際のデータを本システムに導入するために開発したデータローディングシステム,観測者が利用するユーザインターフェース,管理者側が利用する進捗管理システム等に関し概説し,この運用を通して顕在化した問題点および今後の方針について述べる.

2. 地球観測データを対象としたデータクオリティコントロールシステムの構築

昨今のIT技術の進歩により,より高速に大容量のデータの処理が可能なコンピュータが安価に普及しつつあるが,これまで述べたような地球環境観測データのクオリティコントロールを行う上での大きな問題点として,実際に妥当性の検証を行う研究者が上記のような計算機の高次利用が容易ではないという点も挙げられる.実際,測定データの検証を行う際,観測を行った多くの当該分野研究者は,簡単に扱うことのできるパーソナルコンピュータ上で市販の表計算ソフトを用いて１つ１つ手作業で図化を行い,目で見て確認しているのが現状である.

そのため,上述のように大量に取得されるデータの検証に膨大な時間が必要とされ,せっかく取得された有用なデータも十分な検証が行われていないために当該分野の研究者からの信用が得られず,結果として使われないケースも想定される.

そこで本研究では,地球環境工学分野の研究者と密接に協力し,多量の観測データを対象としたデータクオリティコントロールシステムを開発し,実際の検証作業において容易に利用が可能なWebベースのユーザインターフェースを実装したシステムの構築を行った.

2.1 システム概要

本システムは図１に示す流れで運用される.

まず,観測者は自身で取得した未検証のデータをデータクオリティコントロールシステム管理者に送付する.システム管理者は,後述するデータローディングツールを用いて観測地データベースシステムにデータを導入する.観測者はロード

♣ 正会員 東京大学空間情報科学研究センター
eikoma@csis.u-tokyo.ac.jp
♦ 非会員 東京大学大学院工学系研究科 {tamagawa, tkoike}@hydra.t.u-tokyo.ac.jp
♠ 正会員 東京大学生産技術研究所
kitsure@tkl.iis.u-tokyo.ac.jp

されたデータに対してクオリティコントロールインターフェースを通じてアクセスを行い,データ検索ツールで検索されたデータを対象にデータ視覚化ツールによって視覚化された検証データおよび比較データを確認し,必要があればインターフェース上からデータアップデートとフラグ付与要求を出す.データアップデートツールはその要求に従い動的にSQLを生成し,観測値データベースを更新すると同時に,再度データ視覚化ツールが更新後のデータを視覚化しインターフェース上のグラフを更新する.

以下にここで述べた各ツールおよびインターフェースの具体的な内容を示す.

2.2 データローディングツール

観測者から送付されたデータをデータクオリティコントロールシステム管理者がデータベースにデータをローディングするためのツールである.

データをデータベースにロードする以外に本ツールが有する機能は以下の通りである.

・観測データの一次チェック機能
・パターン指定による異常データ訂正機能
・データごとの利用権限に基づいたユーザ管理テーブル生成
・インデックス生成

まず,データロード時には予め観測責任者や一般的な研究者によって作成された,当該データの上限・下限値データベースにアクセスし,その両値の範囲内に含まれていないデータを異常データとし,フラグ「M」を自動で付与する.

また,欠損値等,あらかじめ判明しているパターンを有する異常値に関しては,本システムで定義する欠損表記方法に変換を行う.この情報は,データ提出者が提供する場合と,管理者側が作成する場合がある.

本システムでは,まだ未公開のデータを対象としているため,各観測データごとに利用権限の設定を行っている.データごとに利用可能なユーザを定義したテーブルを参照し,それに基づいてユーザごとに利用可能なデータ一覧を生成し,後述するインターフェース上の検索部分で利用する.

また,本観測データのように多くの観測者からデータを集めて実装を行う場合は,一般的に同時にデータが到着しないケースが多いため,データを追加ロードする度に自動でインデックスを再度生成する機能を実装した.

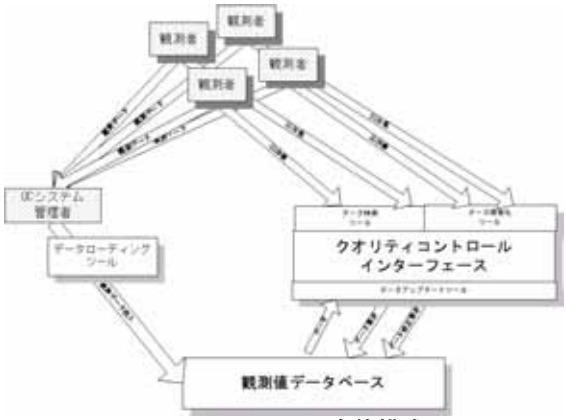

図1 QCシステム全体構成
Fig.1 Structure of QC System

2.3 ユーザインターフェース

各ユーザはログイン後,ユーザインターフェースは図2に示すように,4つのフレームを持つページが表示される.

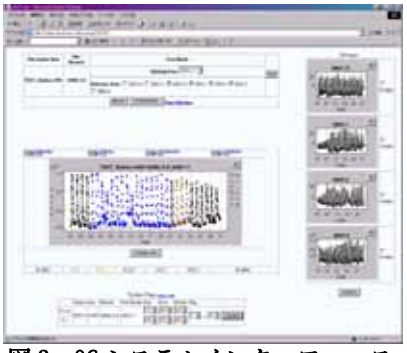

図2 QCシステムインターフェース
Fig.2 Interface of QC System

(1)データ検索フレーム

QCを行いたいデータを検索するフレームである.本フレームではログイン時,観測ステーションおよびアイテム,観測エレメント（データ名）,年・月の各項目に関し選択メニューが表示されている.ここでいずれか1つの項目に関し選択を行うと,その選択された条件を元に動的にSQLを生成し検索を行い,本システムに収蔵されているデータから可能性のある項目のみが表示される.

すなわち,例えば観測エレメントから選択する場合,観測エレメントAを選択すると,Aを観測したデータが1つでもあるステーションおよびアイテム,年月が候補としてそれぞれ観測ステーション・アイテム欄,年・月欄に候補として表示される.

続いて,同様に残り2つのうちの任意の項目に関し指定を行うと,最後の1つの項目に関しては図3に示すように,Updating Data という欄とReference Dataという2欄が表示される.このUpdate Data欄はプルダウンメニューになっており,実際にQCを行いたいデータを1つ選ぶことが出来,Reference Data欄はUpdate Data欄で選んだデータと比較対照を行いたいデータを複数個選ぶことが出来る.

ここで選択したデータは,後述の検索データ視覚化フレームおよび参照データ視覚化フレームに表示される.

図3 2項目選択時の検索ページ
Fig.3 Search Page after chooseing 2 items

(2)検索データ視覚化フレーム

本フレームでは,上記検索フレーム上でUpdate Data欄で選んだデータが表示される.ここで,各データはフラグごとに色分けされたグラフで表示される.そのグラフはクライアント側にロードされたJava Appletであり,以下の表示操作機能を有する.

・マウス左ボタンドラッグによる選択範囲拡大機能
・マウス右ボタンドラッグによる拡大率を維持したままの移動
・マウス両クリックドラッグによる選択範囲縮小機能

例えば,図4のように,1月分のデータが表示されたグラフ上で,上記(1)のようにマウス左ボタンドラッグで関心のある領域を選択すると,拡大されたグラフが表示される.

図4　初期表示のグラフ（1ヶ月分）
Fig.4　Default Graph View(monthly)

また,グラフ下に表示された各フラグの頭文字および数字は,当該フラグの付与されている値の数を示しており,クリックすることで実際のそのフラグが付与された値の一覧が別フレームで表示される.

(3)データアップデートフレーム

本フレームでは,ある期間のデータをまとめてフラグ付与することが可能である.付与したいデータ群の開始日時(ここでは2002年10月7日00時00分),終了日時(2002年10月16日23時00分),元のフラグ(U),付与後のフラグ(G)を図5に示すフレーム上で指定し,Updateボタンを押すことでデータの更新およびグラフの再描画を行う.

図5　データアップデートフレーム
Fig.5　Data Update Frame

(4)参照データ視覚化フレーム

検索フレーム上でReference Dataとして指定されたデータ群が縦に並んで一覧グラフ表示される.各グラフは縮小表示されているためにフラグの色分けは行っていないが,上述の検索データ視覚化フレームと同様に拡大・縮小・移動が可能である.また,Overlay欄にチェックを入れてOverlayボタンを押すことで,選択された全参照データを１つのグラフに重畳して表示することが可能であり,こちらは多くのデータを重ね合わせて表示する場合が多いため,実装されている拡大機能が非常に有効である.

2.4 進捗状況管理

本システムにおいてQCを行った進捗状況を示すページが図6である.これは管理者が各観測者のデータQC進捗状況を確認するためのページであり,各サイト,ステーション,アイテムごとに各フラグの値の数,合計データ数,QC完了数,完了割合%が表示され,進捗度合がグラフで表示される.

図6　QCシステム進捗管理ページ
Fig.6　Status Page of QC System

3. 検討

本システムは前バージョンとしてCEOP-QC System Ver.1.7があり,別の期間のデータに対し一部重複した作業者によってQC作業が行われた経緯がある.

本章ではその際と今回のVer.2.0上でのQC作業の実行状況の変化に関し検討を行った.

3.1 前バージョンとの相違

前バージョンから追加された機能としては,主に以下のようなものがある.

・データアップデート時にグラフ上でマウスクリックによるデータ指定(以前はプルダウンメニューを用いて日時を指定し,当該データの値とフラグが表示されて後,変更を行っていた.)

・グラフ上で拡大・縮小した状態での上下左右スクロール機能(以前は拡大した状態で移動が出来ず,毎回元のスケールに戻ってから拡大縮小を繰り返していた.)

・一括フラグ付与時,従来は期間指定した間のデータを全て同一にフラグ付与する手法であったが,今回からは期間および指定したフラグのデータを一括でフラグ付与する手法に変更(以前は飛び飛びの期間をまたいでの一括変換が不可能であったため,細かな連続期間ごとにフラグ付与作業を行っていたが,今回は飛び飛びの期間であっても,その期間内のあるフラグのみを変換することが出来るようになった.)

・進捗管理ページの追加(以前は進捗管理をしておらず,各作業者の自主性に任せていたが,このバージョンでは視覚的に進捗度合いが把握出来るようになった)

3.2 1データあたりの作業時間の変化

検索フレームにおいて１つのデータを検索し,視覚化を行った後,データのフラグ付与を行い,表示されているデータすべてにフラグ付与を行ったケースにおいて,この作業に要した時間を計測してまとめたのが表1である.

表１　１データあたりの処理時間
Table.1　Processing time for 1 data

User	Ver1.7 データ数	Ver1.7 処理時間 (秒)	Ver2.0 データ数	Ver2.0 処理時間 (秒)	比
a	834	288			
b	455	153	44	41	26%
c	855	54			
d	471	203	587	57	28%
e	1703	127			
f	13	9			
g	352	39	1836	17	44%
h	177	256	291	70	27%
i	40	278	105	88	31%
j			21	40	
k			621	46	
l			66	105	
m			416	30	
n			167	125	
		156.3		61.9	39%

ここで,検討対象としたデータは

・1セッションあたり3データ以上を連続して作業

・1データの処理作業が1度に（連続した時間で）行われた場合

のものであり,10データ以上を処理したユーザのみを取り上げた.その結果,同一ユーザによる作業時間の比較では,要した時間はそれぞれ26%,28%,43%,27%,31%と大幅に短縮されていることが分かる.全ユーザの平均でも39%になっており,2度目の作業でその操作に慣れたと思われる要素を差し引いても本バージョンで追加された機能の効果は大きかったと思われる.

3.3 1セッションあたりの作業時間の変化

次に,1セッションあたりの作業時間の変化をまとめたものが表2である.

表2　1セッションあたりの処理時間
Table.1　Processing time for 1 session

User	Ver1.7 セッション数	Ver1.7 処理時間(秒)	Ver2.0 セッション数	Ver2.0 処理時間(秒)	比
a	83	2900			
b	22	3179	5	3885	122%
c	41	1127			
d	39	2453	16	3239	132%
e	69	3137			
f	4	29			
g	40	345	35	920	266%
h	32	1416	22	1857	131%
i	8	1232	7	2224	180%
j			5	2170	
k			13	2234	
l			8	869	
m			10	1273	
n			8	2621	
		1757		2129	121%

ここで,1セッションというのは,ユーザがログインを行い,データ検索や視覚化,修正等QCに関する一連の作業を行い,(ログアウトという機能は定義していないため)30分以上サーバ側との通信を行わず,さらにその後同一ユーザが再度ログイン処理から行った場合の,その再度ログイン直前までの一連の作業を指す.

セッション数の比較に関しては,諸条件の違いも考慮すると一概に評価は出来ないが,全体として前バージョンに比較して減少傾向が見える.反面,1セッションあたりの処理時間に関しては,同一ユーザでも22%,132%,266%,131%,180%,平均で121%と明らかに増加している.

これは,1度のログインにつき作業を行う時間が延びたことを意味しており,従来に比べユーザがQC作業に「飽きにくく」なった可能性もあると思われる.

3.4 各観測者が自分の担当データの QC を終えるまでの時間

最後に,表 3 は,ユーザごとの QC 開始から完了までの日数を示している.

表3　ユーザごとの QC に要した時間
Table.3　Processing time for each user

User	Ver1.7 日数(日)	Ver2.0 日数(日)
a	19	
b	29	11
c	15	17
d	22	
e	22	
f	1	
g	38	25
h	12	19
i	33	3
j		8
k		5
l		9
m		12
n		8
平均	19.6	11.7

ユーザごとに担当するデータ数が若干違うため,異なったユーザ間の比較は必ずしも有意ではないが,参考までに全てのユーザのQCに要した日数の集計を行った.この日数は,最初にQC作業を開始し,担当データすべてにフラグ付与がされるまでに日数を計測したものである.作業者の多忙度など時期的な要素もあるので一概に比較は出来ないが,実作業者の感想としても作業完了までの日数が大幅に短縮されたという評価があった.

これは,前述の進捗状況管理ページによる効果も考えられるが,実際に携わった作業者,最終的に取りまとめた管理者共に同様の意見を述べていたため,一定の効果があったと考えられる.

4. おわりに

本稿では,筆者らが開発した地球環境観測データを対象とした品質管理システムの背景,対象データ,機能,利用方法等を概説した.

また,その利用ログに関し,基本的な検討を行った.

今後はそのQC結果に加え,ログの解析をより詳細に行い,その結果を元に各ユーザごとの作業内容の解析を進めた上で,より効率的なQCが実現出来るインターフェースの検討を行う.

[謝辞]
本研究は文部科学省 科学技術振興調整費「水循環インフォマティクスの確立」（代表：東京大学大学院工学系研究科 小池俊雄）の成果の一部である.ここに記して謝意を表します.

[文献]

[1] Wallce, J.M. and Gutzler D.S., “Teleconections in the geopotential height field during the Northern hemisphere winter”, Mon. Wea. Rev., vol.109,pp.785-812, 1981.
[2] University Corporation for Atmospheric Research (UCAR), “CEOP Data Flag Definitions”, http://www.joss.ucar.edu/ghp/ceopdm/refdata_report/data_flag/definition.html
[3] Coordinated Enhanced Observing Period(CEOP), “CEOP Data Archive Center”, http://www.ceop.net/
[4] CEOS Year Book, “CEOS satellite data exchange principles for global change data”, http://monsoon.t.u-tokyo.ac.jp/ceop/policy1.html

生駒 栄司　Eiji IKOMA
東京大学空間情報科学研究センター助手．2000 年東京大学大学院工学系研究科博士課程了．博士（工学）．大規模データベースシステム,ユーザインターフェース,ビジュアリゼーションに関する研究・開発に従事.電子情報通信学会会員．日本データベース学会会員.

玉川 勝徳　Katsunori TAMAGAWA
東京大学大学院工学系研究科特任研究員．1995 年長岡技術科学大学大学院工学研究科修士課程修了．修士（工学）．人工衛星によるマイクロ波リモートセンシング，地上観測データアーカイブシステムに関する研究に従事.

小池 俊雄　Toshio KOIKE
東京大学大学院工学系研究科教授. 1985 年東京大学大学院工学系研究科博士課程修了，工学博士．水循環，衛星リモートセンシング研究に従事．統合地球水循環強化観測期間プロジェクト（CEOP)リードサイエンティスト．著書に「地球環境論」（岩波書店，共著）など.

喜連川 優　Masaru KITSUREGAWA
東京大学生産技術研究所教授，同所戦略情報融合国際研究センター長. 1983 年東京大学大学院工学系研究科情報工学専攻博士課程修了，工学博士．データベース工学，並列処理，Web マイニングに関する研究に従事．本会理事，情報処理学会フェロー，SNIA-Japan 顧問，ACM SIGMOD Japan Chapter Chair (H11-H14)，電子情報通信学会データ工学研究専門委員会委員長(H9, 10)．VLDB Trustee, IEEE TKDE Assoc. Editor, IEEE ICDE, PAKDD, WAIM Steering Comm. Member.