〔非公開〕

ＴＲ－Ｃ－００１５

文書画像データベース編集プログラム

西村　康　　高橋　友一　　小林　幸雄
YASUSHI NISHIMURA　TOMOICHI TAKAHASHI　YUKIO KOBAYASHI

１９８８．９．　１．

ＡＴＲ通信システム研究所

目　　次

1. 概要

非言語による情報授受の研究の一環として、文書を画像として読み取り、レイアウト構造さらには論理構造を認識・理解する研究を行なっている。

従来の文書画像理解の研究においては、研究施設環境の制約のために大量の文書画像を用いたレイアウトの解析、大量データによる処理手法の検証などは行なわれていなかった。昨今のハードウエア環境の状況は、大量データの容易な取扱を可能としている。

大量データを使用し、リレーショナルデータベースシステムを利用して効率的に解析、実験、検討が行える環境をつくることを目的として、文書画像データベース編集プログラムを作成した。

本システムは、文書画像データをドラムスキャナ装置から入力して、その画像データを補正し、レイアウト構造を抽出して、リレーショナルデータベースに登録するシステムである。

オペレータの介在により、文書画像上で一様な意味を持つ領域を指示し、その領域に属性ラベルを付加することができる。

このシステムでは、画像データの入力をＶＡＸ／ＶＭＳ上で行ない、画像データ補正とレイアウト情報抽出、並びにデータベース登録をＳＵＮ／ＵＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ上で行なう。

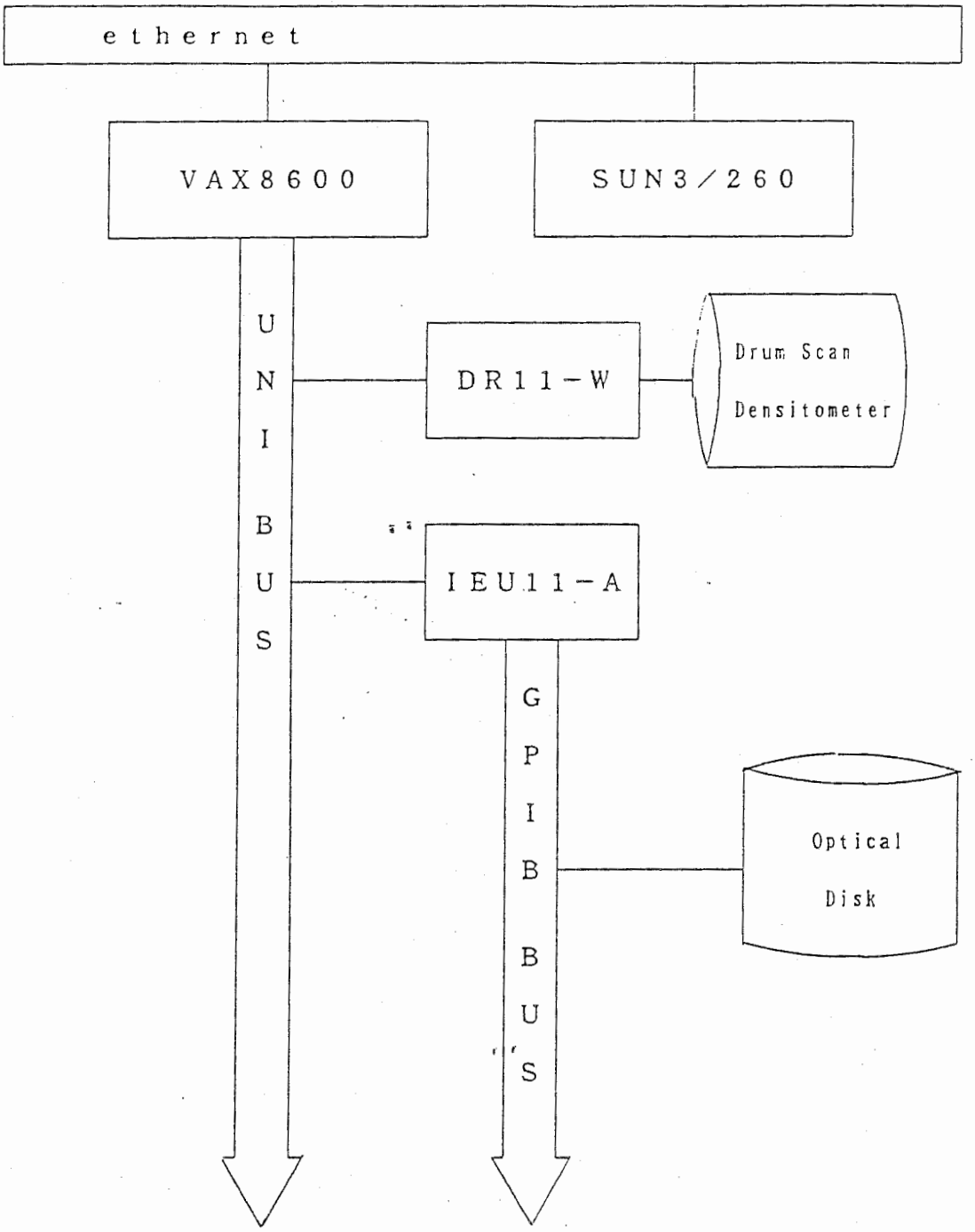
現在、本システムは、科学技術論文タイトルページの画像データベースシステムとして使用しており、以下では論文レイアウトデータベース編集プログラムと呼ぶ。

本報告では、論文レイアウトデータベース編集プログラムの概要及び機能についての記述を行なう。

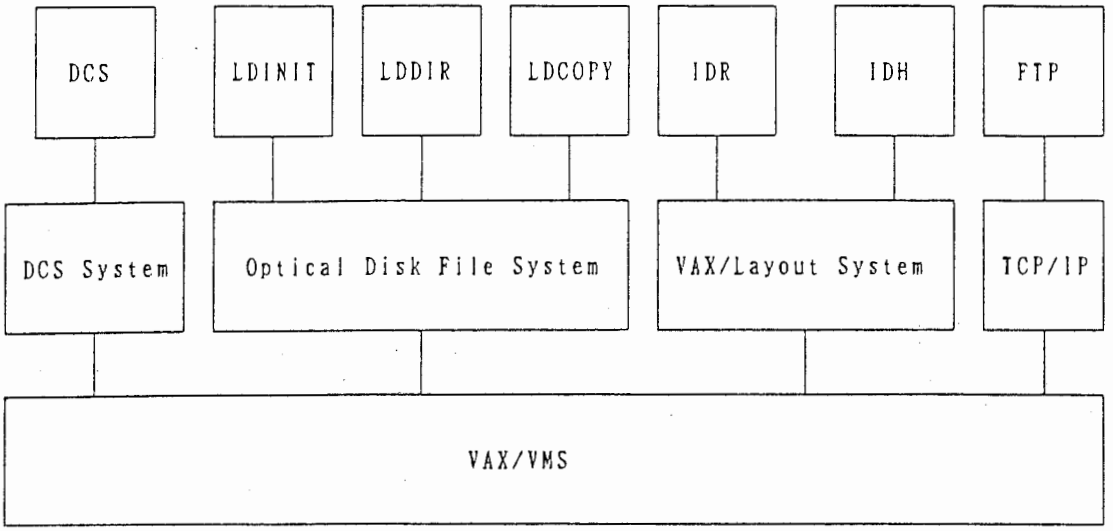
システムのインストール、使用方法の詳細については、以下の関連資料を参照のこと。

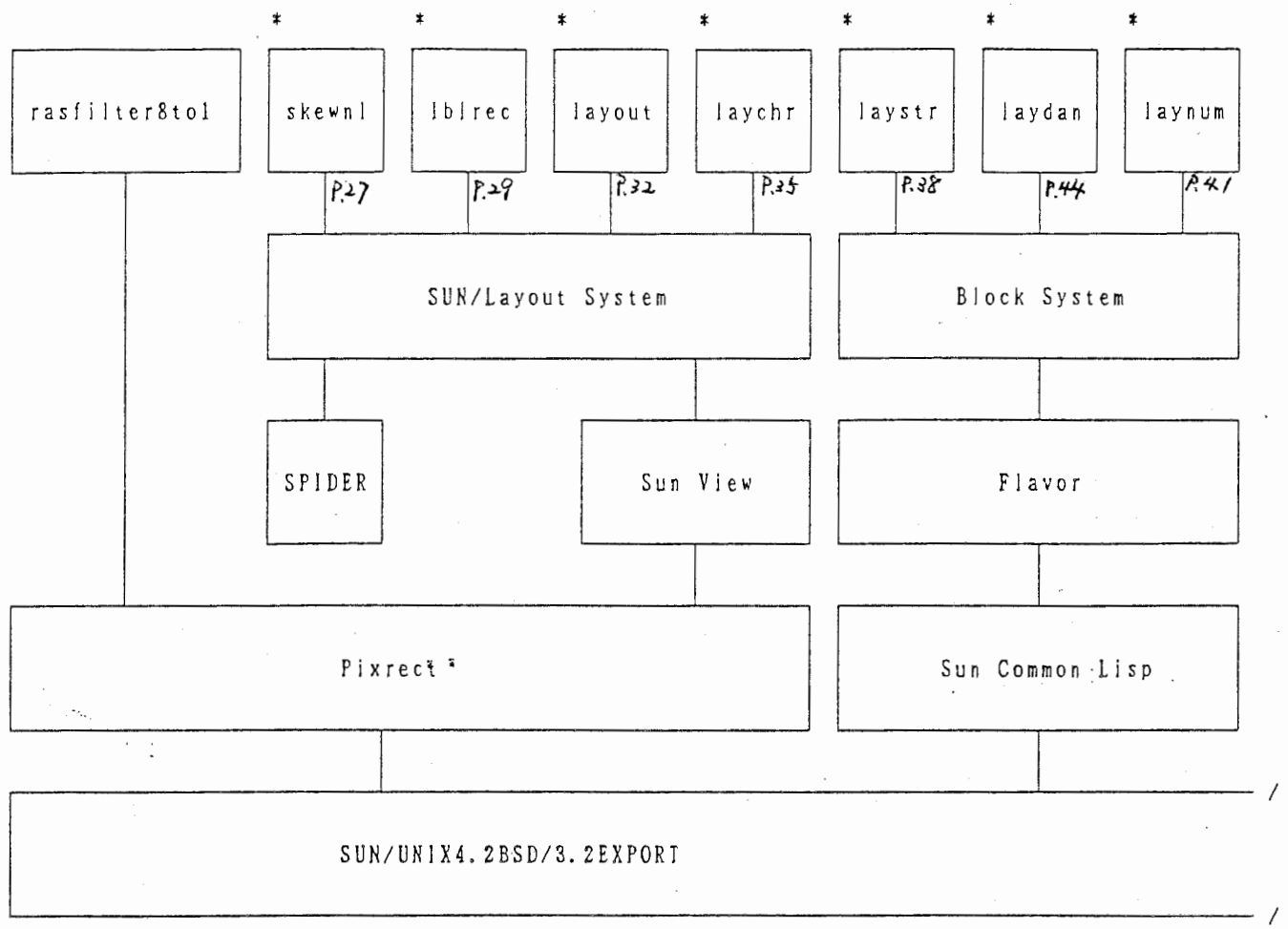
関連資料：VAX/VMSドラムスキャナ画像入力プログラムＤＣＳマニュアル
理経光ディスク装置ファイルシステムVER2.1 OS301-1-0Mマニュアル
論文レイアウト情報抽出用ユーティリティプログラムマニュアル
論文レイアウトデータベース編集プログラムマニュアル

2．ハードウェア構成

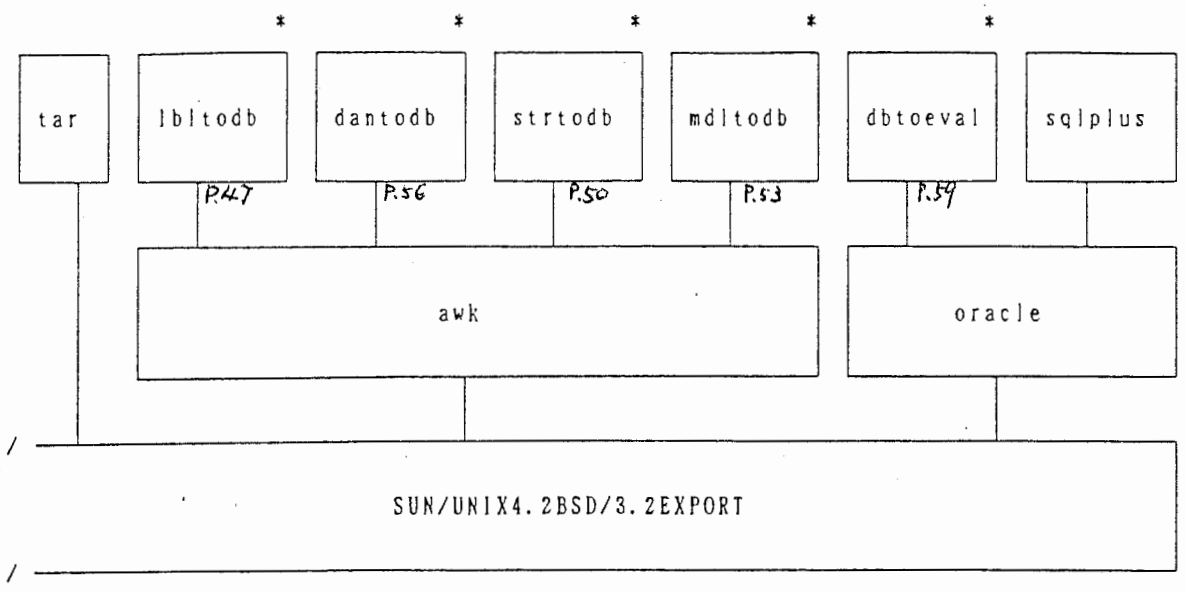

ethernet
VAX8600
SUN3／260
UNIBUS
DR11－W
Drum Scan
Densitometer
IEU11－A
GPIB BUS
Optical
Disk

3．ソフトウェア構成


DCS
LDINIT
LDDIR
LDCOPY
IDR
IDH
FTP
DCS System
Optical Disk File System
VAX/Layout System
TCP/IP
VAX/VMS

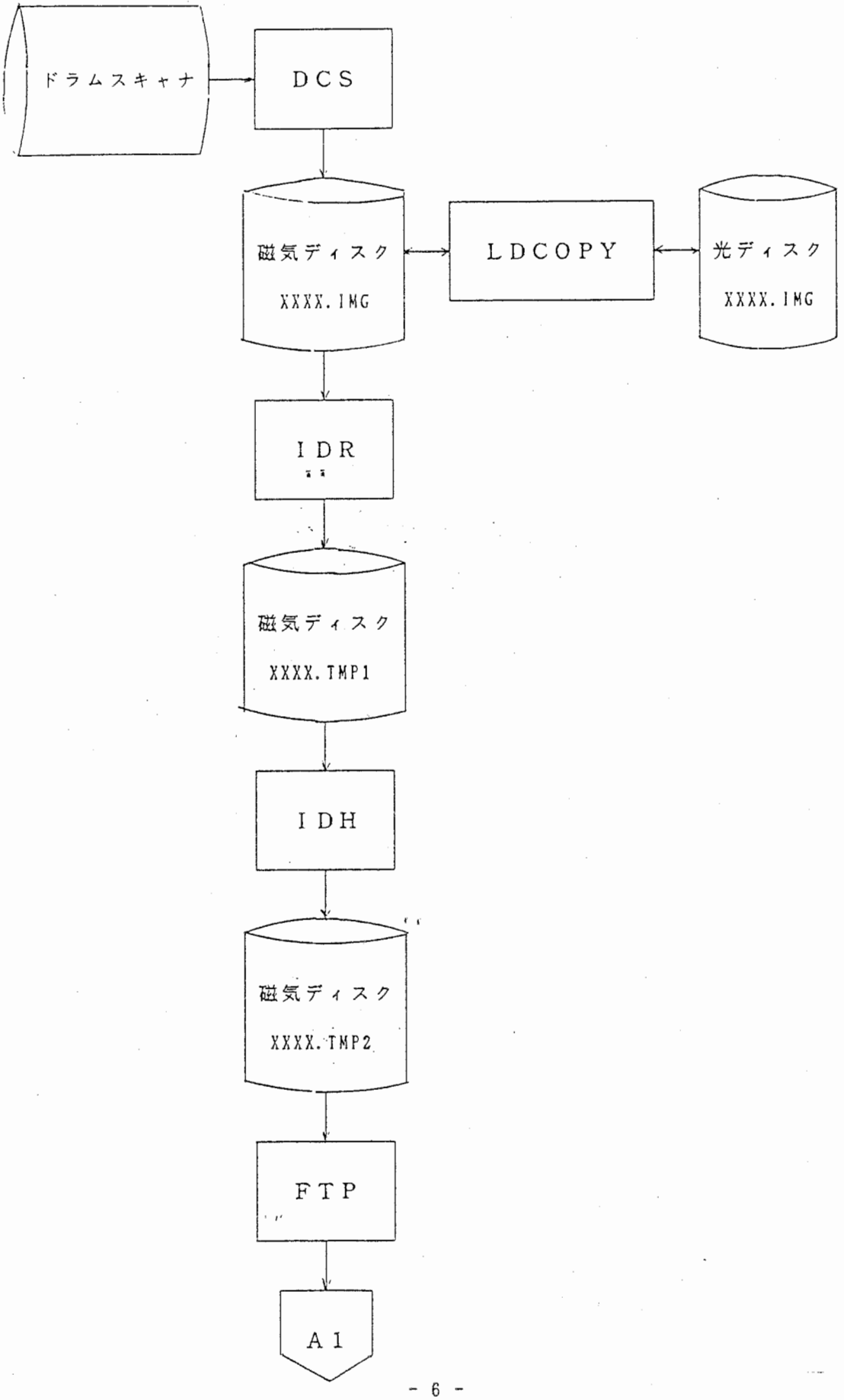

rasfilter8to1
skewn1
lblrec
layout
laychr
laystr
laydan
laynum
P.27
P.29
P.32
P.35
P.38
P.44
P.41
SUN/Layout System
Block System
SPIDER
Sun View
Flavor
Pixrect *
Sun Common Lisp
SUN/UNIX4.2BSD/3.2EXPORT

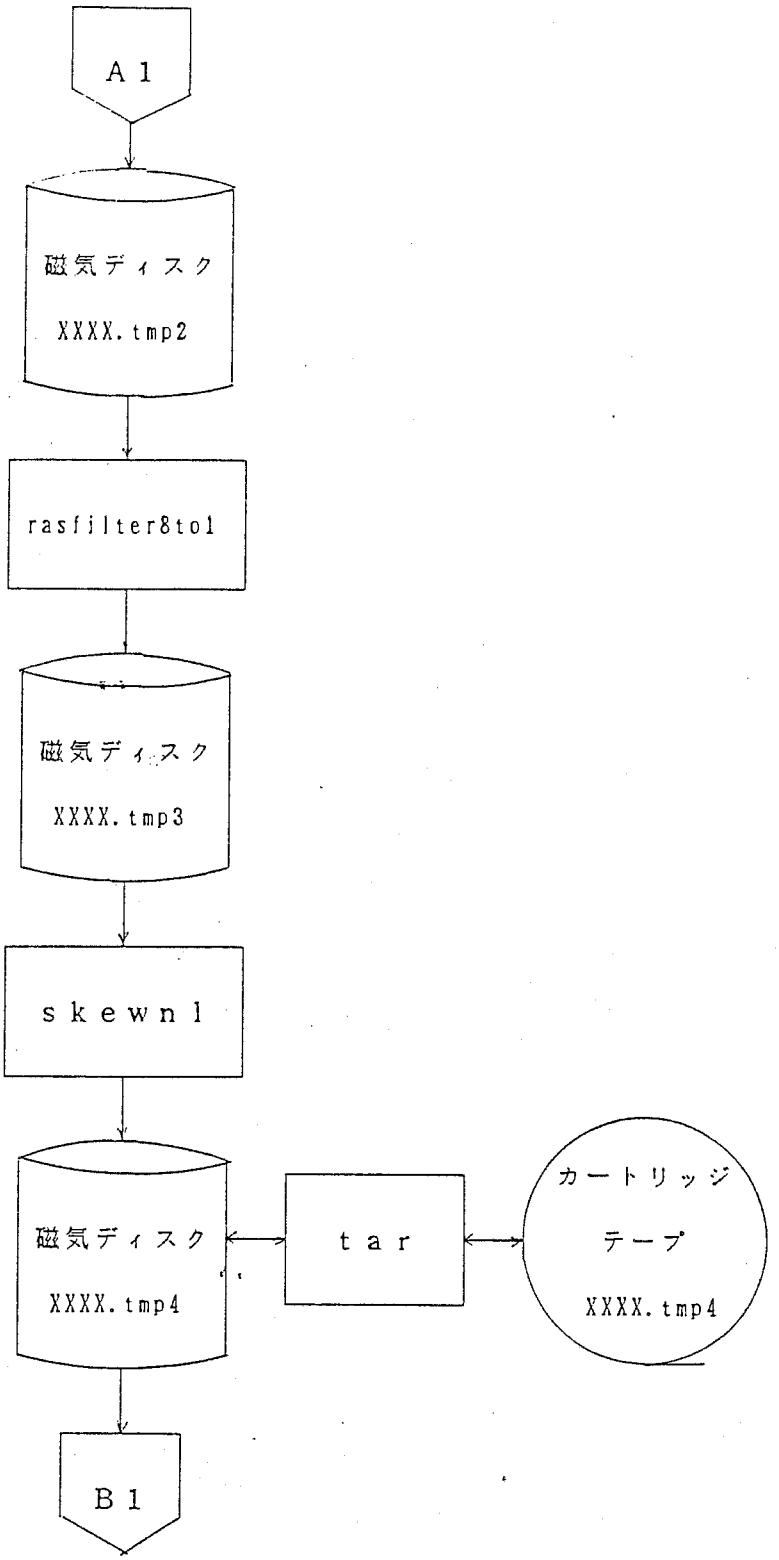
※本仕様書で記述

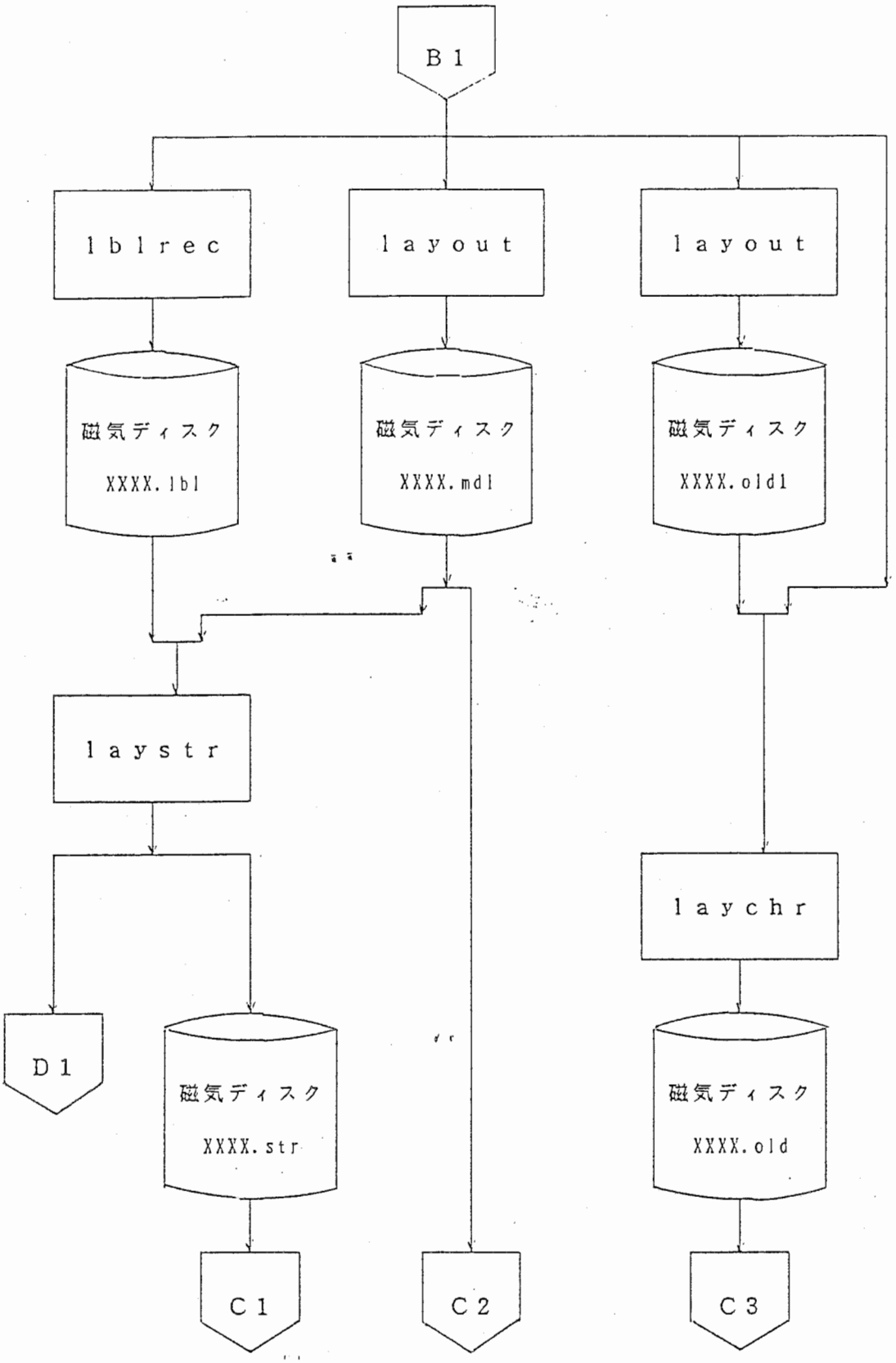

tar
lbltodb
dantodb
strtodb
mdltodb
dbtoeval
sqlplus
P.47
P.56
P.50
P.53
P.59
awk
oracle
SUN/UNIX4.2BSD/3.2EXPORT

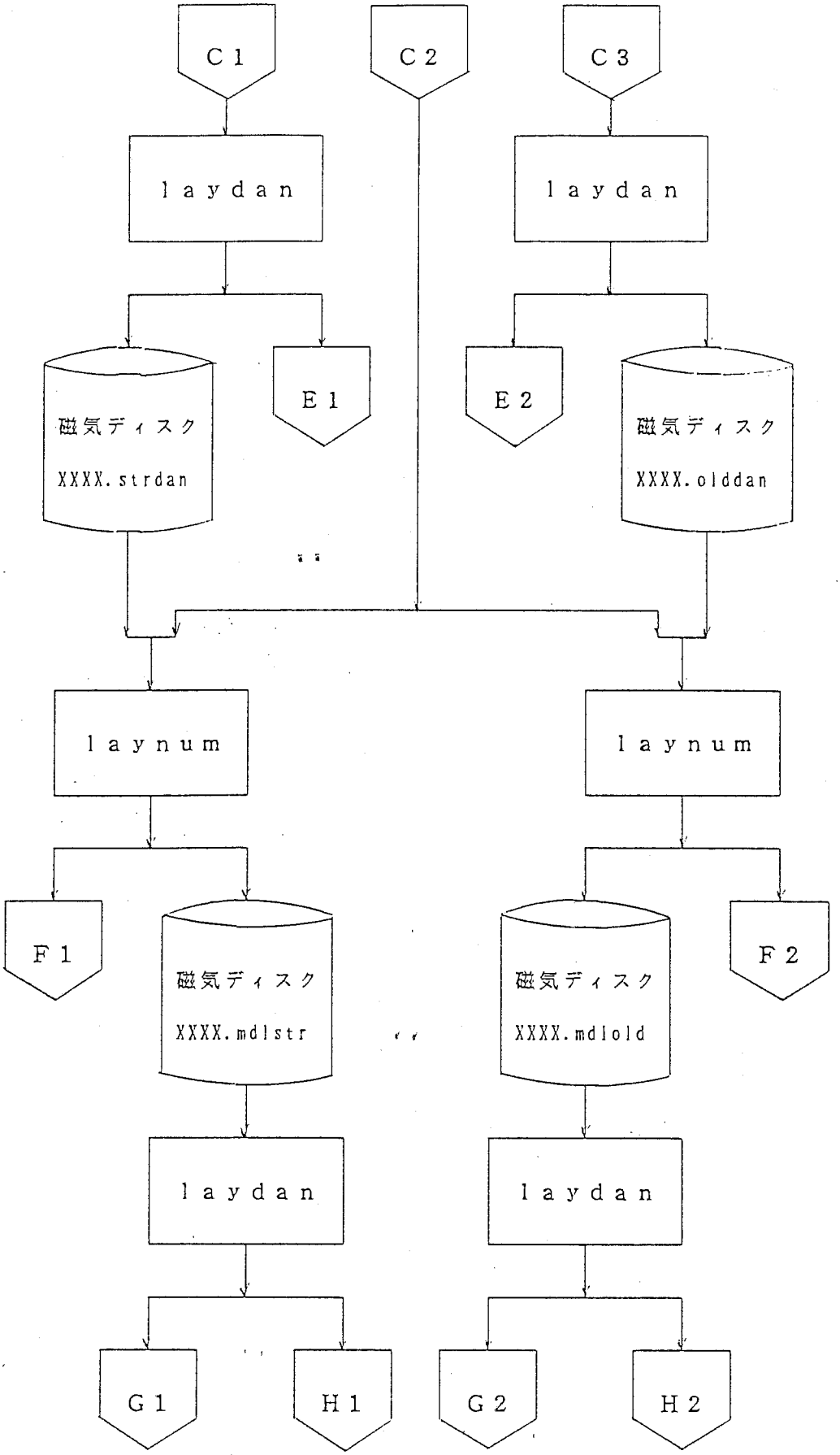
※本仕様書で記述

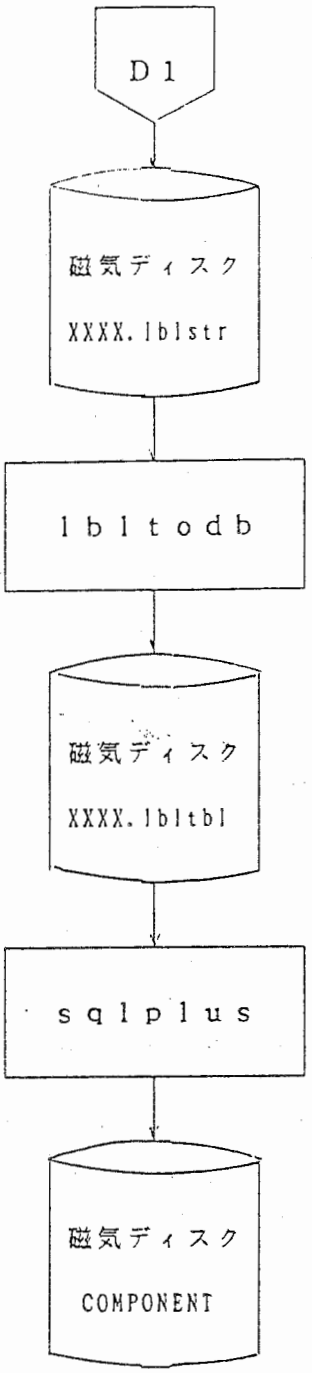
4．データフロー

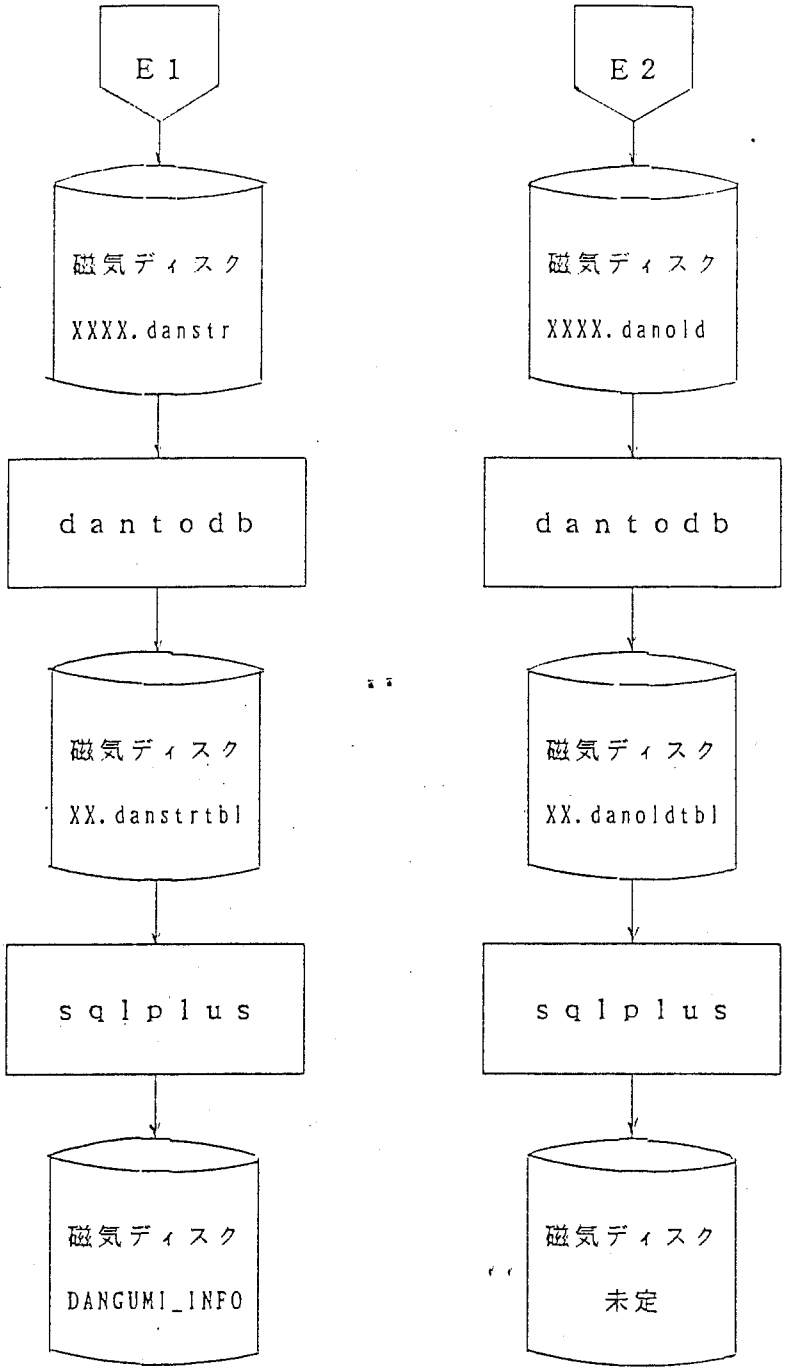

ドラムスキャナ
DCS
磁気ディスク
XXXX.IMG
LDCOPY
光ディスク
XXXX.IMG
IDR
磁気ディスク
XXXX.TMP1
IDH
磁気ディスク
XXXX.TMP2
FTP
A1

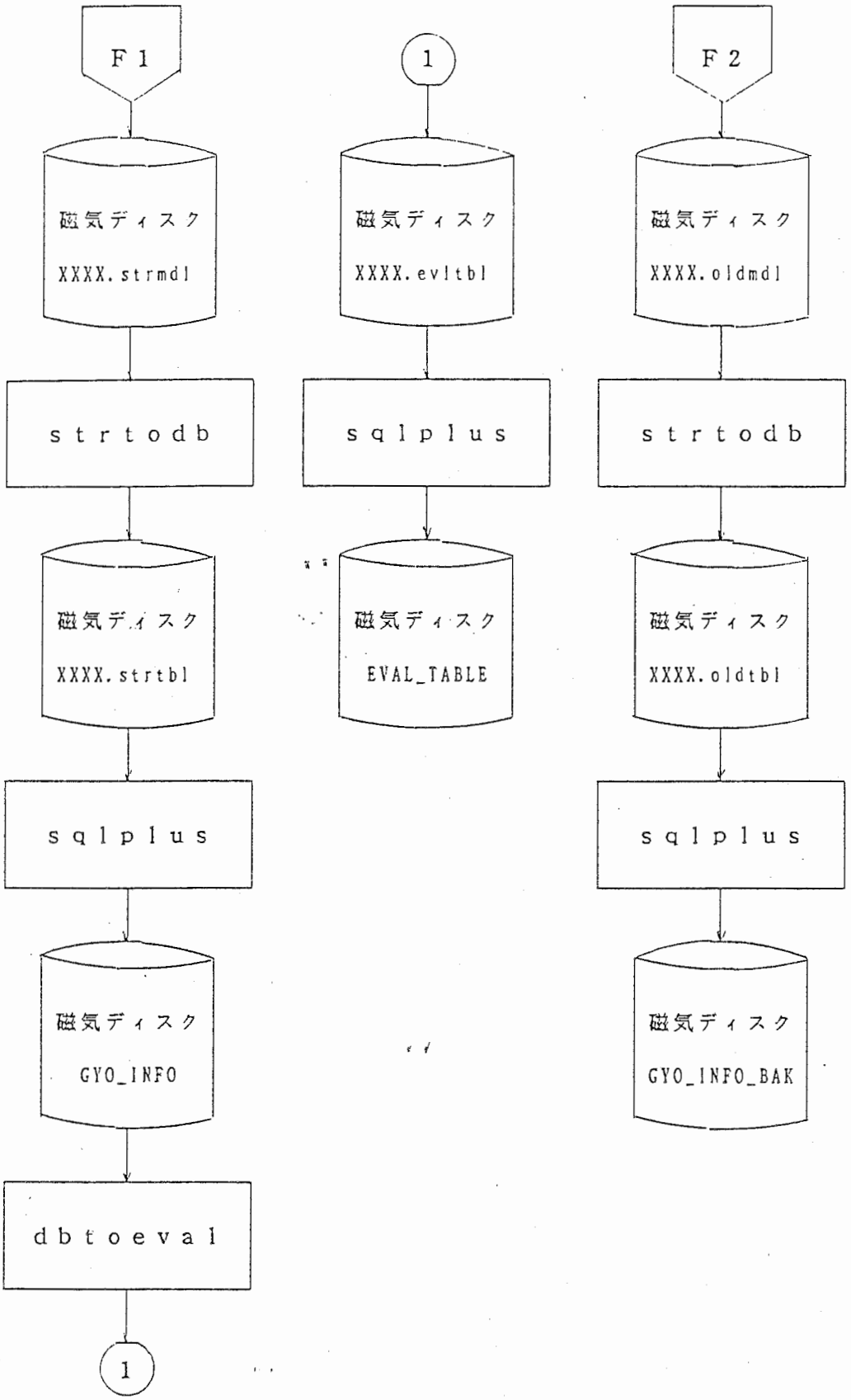

A 1
磁気ディスク
XXXX.tmp2
rasfilter8to1
磁気ディスク
XXXX.tmp3
s k e w n 1
磁気ディスク
XXXX.tmp4
t a r
カートリッジ
テープ
XXXX.tmp4
B 1

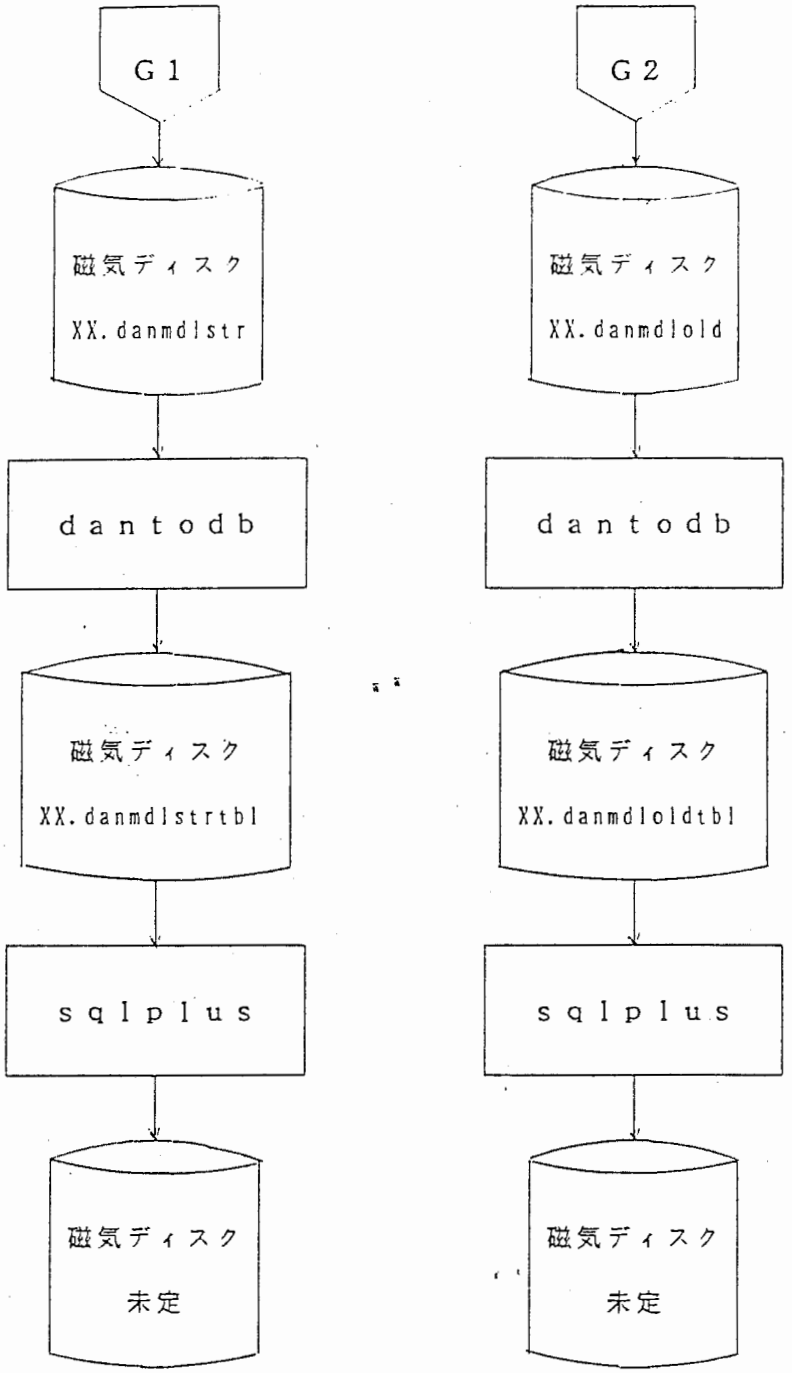

B1
lblrec
layout
layout
磁気ディスク
XXXX.lbl
磁気ディスク
XXXX.mdl
磁気ディスク
XXXX.old1
laystr
laychr
D1
磁気ディスク
XXXX.str
磁気ディスク
XXXX.old
C1
C2
C3

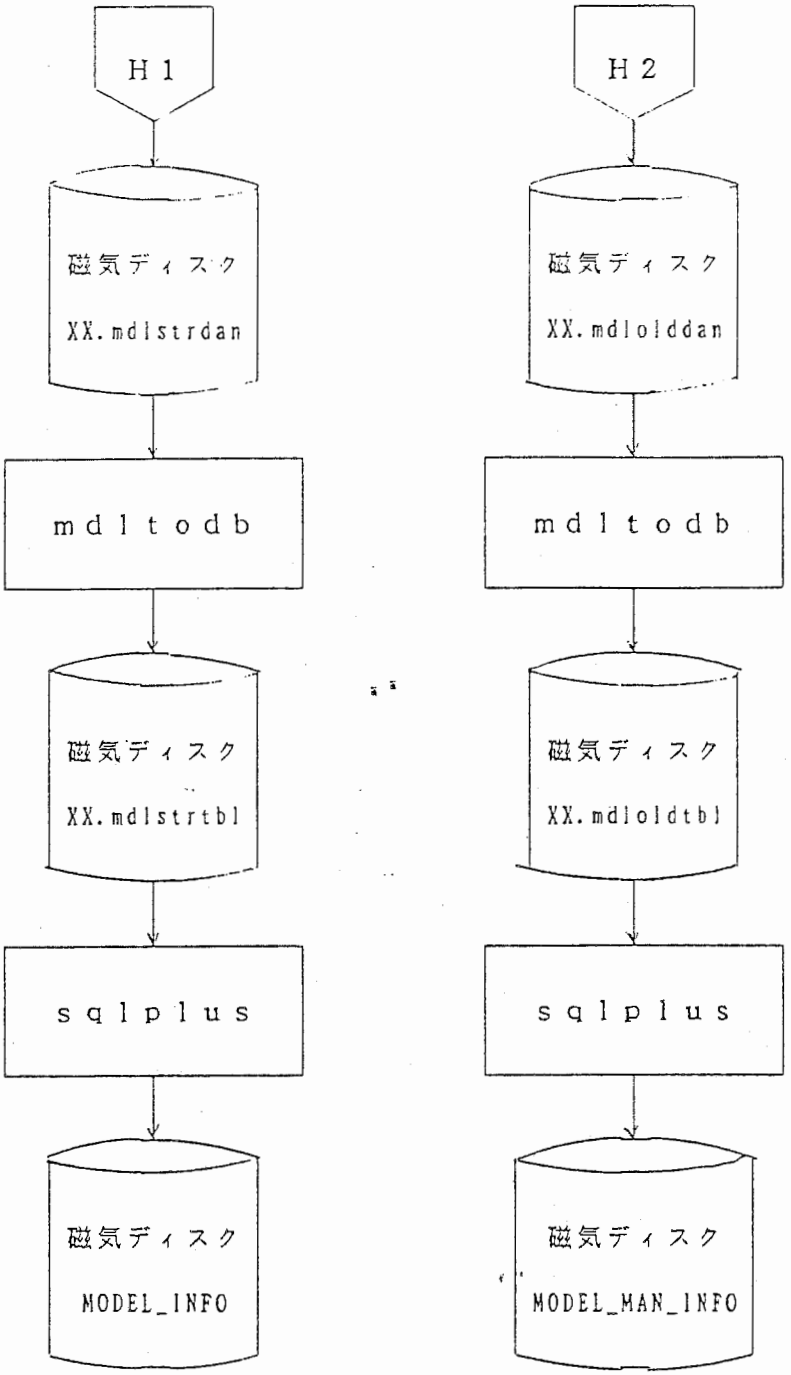

C1
C2
C3
laydan
laydan
磁気ディスク
XXXX.strdan
E1
E2
磁気ディスク
XXXX.olddan
laynum
laynum
F1
磁気ディスク
XXXX.mdlstr
磁気ディスク
XXXX.mdlold
F2
laydan
laydan
G1
H1
G2
H2


D1
磁気ディスク
XXXX.lblstr
lbltodb
磁気ディスク
XXXX.lbltbl
sqlplus
磁気ディスク
COMPONENT


E1
磁気ディスク
XXXX.danstr
dantodb
磁気ディスク
XX.danstrtbl
sqlplus
磁気ディスク
DANGUMI_INFO
E2
磁気ディスク
XXXX.danold
dantodb
磁気ディスク
XX.danoldtbl
sqlplus
磁気ディスク
未定


F1
磁気ディスク
XXXX.strmdl
strtodb
磁気ディスク
XXXX.strtbl
sqlplus
磁気ディスク
GYO_INFO
dbtoeval
1
1
磁気ディスク
XXXX.evltbl
sqlplus
磁気ディスク
EVAL_TABLE
F2
磁気ディスク
XXXX.oldmdl
strtodb
磁気ディスク
XXXX.oldtbl
sqlplus
磁気ディスク
GYO_INFO_BAK


G1
磁気ディスク
XX.danmdlstr
dantodb
磁気ディスク
XX.danmdlstrtbl
sqlplus
磁気ディスク
未定
G2
磁気ディスク
XX.danmdlold
dantodb
磁気ディスク
XX.danmdloldtbl
sqlplus
磁気ディスク
未定


H1
磁気ディスク
XX.mdlstrdan
mdltodb
磁気ディスク
XX.mdlstrtbl
sqlplus
磁気ディスク
MODEL_INFO
H2
磁気ディスク
XX.mdloldtbl
sqlplus
磁気ディスク
MODEL_MAN_INFO

５．ファイル構造

論文レイアウトデータベース編集システムで用いるファイル種別は、大別して以下の５種類になる。

（１）画像データファイル

（２）ラベル矩形ファイル

（３）文字列矩形ファイル

（４）モデル矩形ファイル

（５）抽出データファイル

５－１．画像データファイル

このファイルは、ＶＡＸ上とＳＵＮ上では、異なる形式をしている。（大きな違いは、画像幅などを記録するヘッダ部がＶＡＸ上では分離しているのに対し、ＳＵＮ上では同じファイルの先頭に持つ事である。また、赤、緑、青の各色画像データを、ＶＡＸ上では違うファイルに持つのに対し、ＳＵＮ上では同じファイルに持つ。）

【データ形式】

バイナリデータ

【関連ファイル】

・*.img ― ２０画素／ｍｍの濃淡画像（ドラムスキャナから入力）

参照）ＤＣＳ機能仕様書

・*.tmp1 ― ５画素／ｍｍの濃淡画像（imgから生成）

参照）ＩＤＲ機能仕様書

・*.tmp2 ― ＳＵＮ３用の濃淡画像（tmp1から生成）

参照）ＩＤＨ機能仕様書

・*.tmp3 ― ２値画像（tmp2から生成）

参照）/usr/include/rasterfile.h

・*.tmp4 ― 傾き補正済の２値画像（tmp3から生成）

参照）/usr/include/rasterfile.h

５－２．ラベル矩形ファイル

このファイルは、８連結ラベリングを行った結果の矩形であり、そのほとんどが文字に等しい。従って、文字が集まって出来る文字列へのポインタを持っている。

これらのファイルは、最終的にリレーショナルデータベース（ＯＲＡＣＬＥ）内のテーブルにする。

【データ形式】

アスキーデータ

【関連ファイル】

- *.lbl — ８連結ラベル矩形（tmp4から生成）
- *.lblstr — 文字列番号付８連結ラベル矩形（lblから生成）
- *.lbltbl — データベース用８連結ラベル矩形（lblstrから生成）
- COMPONENT — データベース内ラベル矩形テーブル（lbltblから生成）

５－３．文字列矩形ファイル

このファイルのデータは、文字が集まって出来た矩形データであり、そのほとんどが文字列に等しい。

これらの文字列に等しい矩形データは、文字列が集まって出来る段組矩形データへのポインタを持っている。更に、モデル矩形データへのポインタも持っている。

また、以前は人手で入力したファイルも使っていたが、この旧ファイルも同じ構造を持っている。

これらのファイルは、最終的にリレーショナルデータベース（ＯＲＡＣＬＥ）内のテーブルにする。

【データ形式】

アスキーデータ

【関連ファイル】

- *.str ― 文字列矩形（lblから生成）
- *.strdan ― 段組番号付文字列矩形（strから生成）
- *.strmdl ― ノード番号付文字列矩形（strdanから生成）
- *.strtbl ― データベース用文字列矩形（strmdlから生成）
- GYO_INFO ― データベース内文字列矩形テーブル（strtblから生成）
- *.old ― 旧文字列矩形（tmp4を用いて手入力）
- *.olddan ― 段組番号付旧文字列矩形（oldから生成）
- *.oldmdl ― ノード番号付旧文字列矩形（olddanから生成）
- *.oldtbl ― データベース用旧文字列矩形（oldmdlから生成）
- GYO_INFO_BAK ― データベース内旧文字列矩形テーブル（oldtblから生成）

５－４．モデル矩形ファイル

このファイルは、人手で入力されたモデル矩形である。また、以前は複数のファイルを用いて木構造を表現していたが、今は、木構造の末端だけを使っている。

これらのファイルは、最終的にリレーショナルデータベース（ＯＲＡＣＬＥ）内のテーブルにする。

【データ形式】

アスキーデータ

【関連ファイル】

- *.mdl — モデル矩形（tmp4を用いて手入力）
- *.mdlstr — 文字列矩形のノード番号付モデル矩形（mdlから生成）
- *.mdlstrdan — 段組済モデル矩形（mdlstrから生成）
- *.mdlstrtbl — データベース用モデル矩形（mdlstrdanから生成）
- MODEL_INFO — データベース内モデル矩形テーブル（mdlstrtblから生成）
- *.mdlold — 旧文字列矩形のノード番号付モデル矩形（mdlから生成）
- *.mdlolddan — 段組済モデル矩形（mdloldから生成）
- *.mdloldtbl — データベース用モデル矩形（mdlolddanから生成）
- MODEL_MAN_INFO — データベース内モデル矩形テーブル（mdloldtblから生成）

５－５．抽出データファイル

段組矩形ファイルは、文字列矩形ファイル、及びモデル矩形ファイルを用いて生成する。

評価用ファイルは、データベース内文字列矩形テーブルを用いて生成する。

これらのファイルは、最終的にリレーショナルデータベース（ＯＲＡＣＬＥ）内のテーブルにする。

【データ形式】

アスキーデータ

【関連ファイル】

- *.danstr — 段組矩形（strから生成）
- *.danstrtbl — データベース用段組矩形（danstrから生成）
- DANGUMI_INFO — データベース内段組矩形テーブル（danstrtblから生成）
- *.danold — 段組矩形（oldから生成）
- *.danoldtbl — データベース用段組矩形（danoldから生成）
- *.danmdlstr — 段組矩形（mdlstrから生成）
- *.danmdlstrtbl — データベース用段組矩形（danmdlstrから生成）
- *.danmdlold — 段組矩形（mdloldから生成）
- *.danmdloldtbl — データベース用段組矩形（danmdloldから生成）
- *.evltbl — 評価用データ
- EVAL_TABLE — データベース内評価用データテーブル（evltblから生成）

論文レイアウト情報抽出システムは、以下の機能を持つ。

また、各機能は，単体で個別に動作する。

（1）スキャナ装置通信システム（ＶＡＸ／ＤＣＳ）

ドラムスキャナから文書画像を入力する。

これは、スキャナ装置通信システムを用いる。

（2）光ディスク装置ファイルシステム（ＬＤＣＯＰＹ）

入力した文書画像を、光ディスクへコピーする。

これは、光ディスク装置ファイルシステムのファイルコピー機能を用いる。

（3）画像データ縮小機能（ＩＤＲ）

磁気ディスク上にある文書画像を、１／４に縮小する。

これは、ＶＡＸ側レイアウト情報抽出システムの画像データ縮小機能を用いる。

（4）画像フォーマット変換機能（ＩＤＨ）

縮小した画像を、ＳＵＮ３用にフォーマット変換する。

これは、ＶＡＸ側レイアウト情報抽出システムのフォーマット変換機能を用いる。

（5）ファイル転送（ＦＴＰ）

変換した画像を、ＳＵＮ３／２６０へファイル転送する。

これは、ＩＰ／ＴＣＰシステムのファイル転送を用いる。

（６）画像２値化コマンド（rasfilter8to1）

転送した画像を、２値画像に変換する

これは、ＳＵＮ３／２６０の画像２値化コマンドを用いる。

（７）テープ記録コマンド（ｔａｒ）

補正した画像を、カートリッジテープに記録する。

これは、ＳＵＮ３／２６０のテープ記録コマンドを用いる。

（８）リレーショナルデータベース操作言語（ｓｑｌｐｌｕｓ）

各データベース登録用データを、データベースに登録する。

これは、ＳＵＮ３／２６０のリレーショナルデータベース操作言語を用いる。

注：（１）から（８）までの機能は、外部システムが持つ機能である。これらについての詳細は、各システムの機能仕様書を参照する事。

（９）横書き文書画像傾き補正機能（ｓｋｅｗｎｌ）

変換した画像の傾きを補正する。

（１０）ラベル矩形機能（ｌｂｌｒｅｃ）

補正した画像に、連結成分のラベル付けをして、ラベル矩形を得る。

（１１）レイアウト矩形編集機能（ｌａｙｏｕｔ）

補正した画像を表示し、モデル矩形と旧文字列矩形の生成、並びに編集を行う。

（１２）文字情報付与機能（ｌａｙｃｈｒ）

一時的にラベル矩形を生成し、文字列矩形に文字情報を付与する。

（１３）文字列矩形生成機能（ｌａｙｓｔｒ）

ラベル矩形から、文字列矩形を生成する。（文字情報付与機能も含んでいる）

（１４）ノード番号付与機能（ｌａｙｎｕｍ）

モデル矩形と（旧）文字列矩形に、ノード番号を付ける。

（１５）段組矩形生成機能（ｌａｙｄａｎ）

モデル矩形と（旧）文字列矩形から、段組矩形を生成する。

（１６）ラベル矩形フォーマット変換機能（ｌｂｌｔｏｄｂ）

文字列矩形生成済のラベル矩形を、データベース登録用に変換する。

（１７）文字列矩形フォーマット変換機能（ｓｔｒｔｏｄｂ）

ノード番号付の（旧）文字列矩形を、データベース登録用に変換する。

（１８）モデル矩形フォーマット変換機能（ｍｄｌｔｏｄｂ）

ノード番号付のモデル矩形を、データベース登録用に変換する。

（１９）段組矩形フォーマット変換機能（ｄａｎｔｏｄｂ）

各段組矩形を、データベース登録用に変換する。

（２０）評価データ算出機能（ｄｂｔｏｅｖａｌ）

データベース内文字列矩形テーブルから、評価データを算出する。

６－１．横書き文書画像傾き補正機能

横書き文書画像傾き補正機能は、標準入力から画像データを入力し、標準出力に画像データ出力するフィルタである。

入力と出力のファイルの形式は、ＳＵＮ３標準のカラーマップなし２値画像ファイルでなければならない。

傾きを補正するには、以下の方法を使う。

（１）８連結ラベル付けを行う

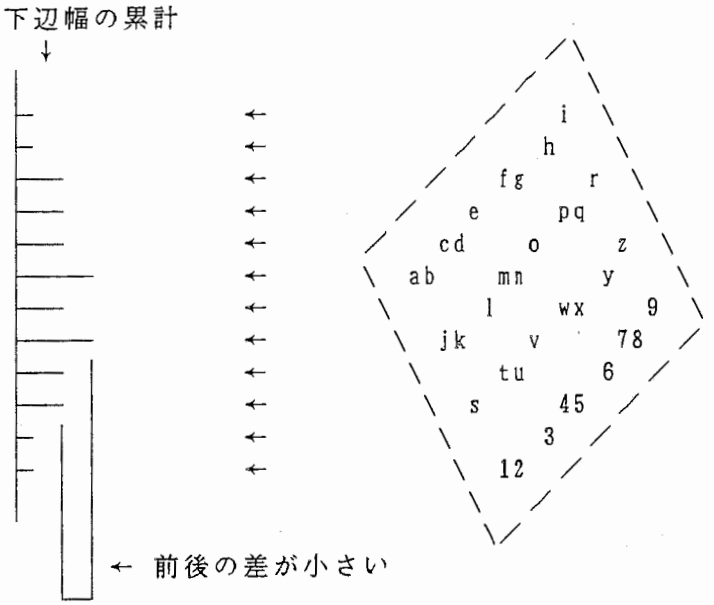
（２）８連結ラベル矩形の下辺をＹ軸方向に角度を変えて投射する

（３）最も尖鋭になる角度を見付ける

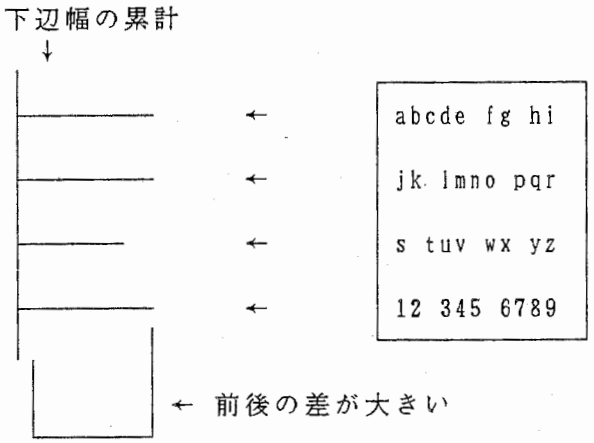
（４）最も尖鋭になる角度にアフィン変換をする

最も尖鋭になるとは、以下の状態を指す。

※ 尖鋭ではない



※ 尖鋭である



【プログラム名】

skewn1 ― 横書き文書画像の傾きを補正する（傾き補正プログラム）

【ハードウェア】

・ＳＵＮ３／２６０Ｃ

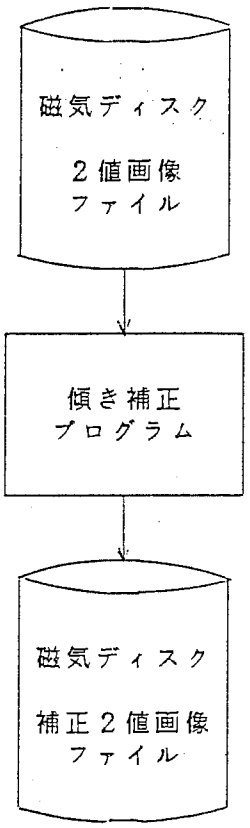
【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・Ｃ言語

・ＦＯＲＴＲＡＮ（ＳＰＩＤＥＲ－ＩのＡＦＩＮ３とＴＲＭＴを使用）

【データフロー】



【構文】

skewnl

【詳細】

ｓｋｅｗｎｌは、標準入力から横書き文書画像データを読みこんで、横書き文書画像データの傾きを補正する。そして、補正済横書き文書画像データを標準出力に出力する。

【制限】

ファイルの形式は、ＳＵＮ３標準のカラーマップなし２値画像ファイルでなければならない。（参照： /usr/include/rasterfile.h）

【関連ファイル】

- *.tmp3 ― 文書２値画像ファイル
- *.tmp4 ― 補正済文書２値画像ファイル

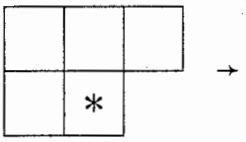
【参照】

- rasfilter8to1 ― ８ビット画像ファイルを１ビットに変換（ＳＵＮ３マニュアル）
- lblrec ― ラベル矩形プログラム（論文レイアウト抽出システム）
- layout ― 矩形データ作成支援プログラム（論文レイアウト抽出システム）

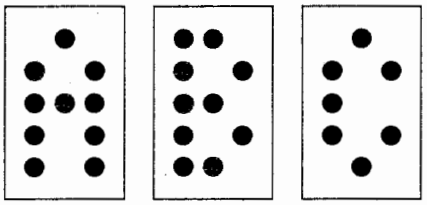
6－2．ラベル矩形機能

ラベル矩形機能は、標準入力から横書き文書画像データを入力し、標準出力に画像データ連結成分のラベル矩形データを出力するフィルタである。

ラベル付けの方法は、下記の＊を現在位置として、四角で囲まれた部分にある黒画素を同じラベルで連結していく。これを矢印の方向に進める。（もし、画像の右端にたどりついたら、現在位置を１画素下の画像の左端に移す。）



ラベル矩形とは、同じラベルで連結された黒画素を囲む最小の矩形を指す。



【プログラム名】

lblrec ― ２値画像にラベル付けをしてラベル矩形を得る（ラベル矩形プログラム）

【ハードウェア】

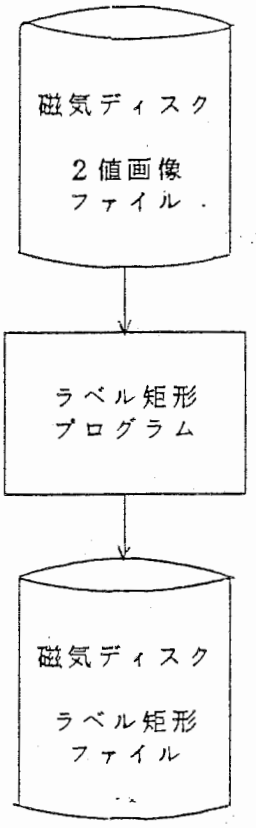
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・Ｃ言語

【データフロー】



【構文】

lblrec

【詳細】

ｌｂｌｒｅｃは、標準入力から画像データを読みこんで、画像データの連結成分のラベル付けをする。そして、得られたラベル矩形を標準出力に出力する。

【制限】

ファイルの形式は、ＳＵＮ３標準のカラーマップなし２値画像ファイルでなければならない。（参照： /usr/include/rasterfile.h）

【関連ファイル】

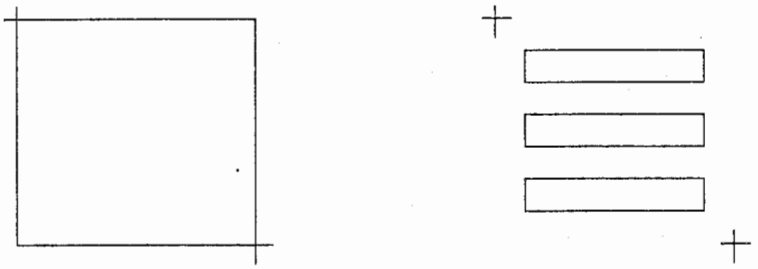
- *.tmp4 — 補正済文書２値画像ファイル
- *.lbl — ８連結ラベル矩形ファイル

【参照】

- rasfilter8to1 — ８ビット画像ファイルを１ビットに変換（ＳＵＮ３マニュアル）
- laystr — 文字列矩形生成プログラム（論文レイアウト抽出システム）

画像ファイルの画像をＳＵＮＴＯＯＬ上のウィンドウに表示する。この上でマウスを用いて矩形の座標の指示を行い、矩形データを配列上に記憶する。なお、この時に指示どおり一つの矩形を得るモードと、自動的に分割し複数の矩形を得るモードを用意する。

※一つの矩形を得る　　※複数の矩形を得る



（＋は、ユーザの指示）

また、記憶した矩形をマウスを用いて指示し、その座標値の表示と、付随データの入力ならびに、その矩形の消去を出来るようにする。

最後に配列上の矩形データを矩形データファイルに書き込んで、後で作業を継続できるようにする。

【プログラム名】

layout ― 論文レイアウトデータ作成を支援する（矩形データ作成支援プログラム）

【ハードウェア】

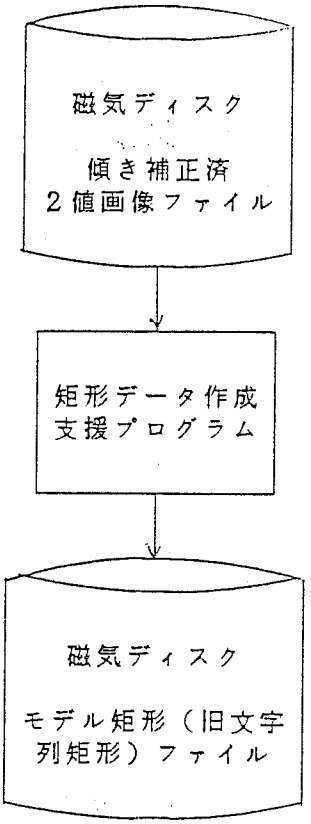
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・Ｃ言語

【データフロー】



【構文】

layout [-b] [-R] [-a〈いき値〉] [-W〈ウィンドウ操作引数〉…]
〈SUN標準画像ファイル〉 [〈矩形データファイル〉]

【引数】

-b … ブロックデータの編集（矩形データファイル名の後ろに0から3までの数値を持つファイルに対して編集を行う）

-R … 矩形データファイルの書き込みを不可能にする

-a … 自動矩形入力時のいき値（0から255、省略値は128）

-W … フレームウィンドウに対する一般的な引数（参照： suntools(1)）

【詳細】

SUN標準画像ファイルの画像をSUNTOOL上のウィンドウに表示する。この上でマウスを用いて矩形の座標の指示を行い、矩形データを配列上に記憶する。

なお、この時に指示どおり一つの矩形を得るモードと、自動的に分割し複数の矩形を得るモードを用意する。（自動的に分割し複数の矩形を得るモードの時には与えたいき値を用いる。）

また、記憶した矩形をマウスを用いて指示し、その座標値の表示と、付随データの入力ならびに、その矩形の消去を出来るようにする。

最後に配列上の矩形データを矩形データファイルに書き込んで、後で作業を継続できるようにする。（すなわち次回はこの矩形データファイルを読んでから始める。）

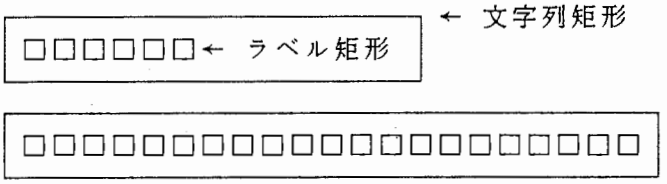
【関連ファイル】

- *.tmp4 — 傾き補正済2値画像ファイル
- *.mdl — モデル矩形ファイル
- *.old — 旧文字列矩形ファイル

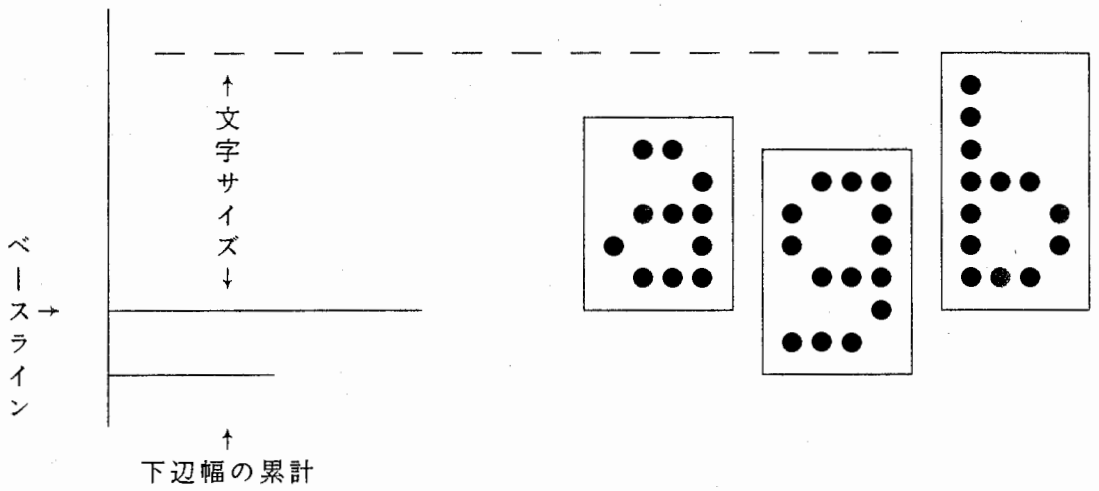
【参照】

- skewnl — 傾き補正プログラム（論文レイアウト抽出システム）
- laynum — 番号付与プログラム（論文レイアウト抽出システム）
- laydan — 段組生成プログラム（論文レイアウト抽出システム）

横書き文書画像データの黒画素を連結し、ラベル矩形を得る。（参照：ラベル矩形機能）但し、効率化の為に各文字列矩形内だけを走査する。



こうして得られたラベル矩形から、ベースラインと文字サイズを得る。ベースラインは、ラベル矩形の下辺幅の累計の一番大きい座標である。また文字サイズは、ベースラインとベースラインに下端が接しているラベル矩形の上端座標の最小値との差である。



【プログラム名】

laychr ― 文字列矩形に文字情報を付与する（文字情報付与プログラム）

【ハードウェア】

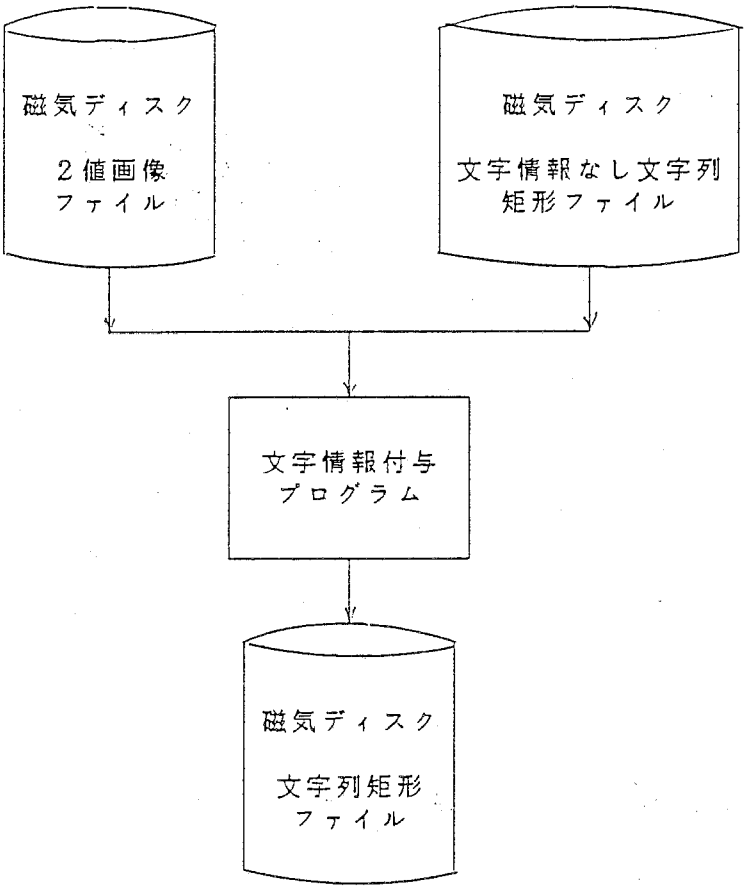
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・Ｃ言語

【データフロー】



【構文】

laychr 〈横書き文書画像ファイル〉

〈文字情報なし文字列矩形ファイル〉 〈文字列矩形ファイル〉

【詳細】

ｌａｙｃｈｒは、一時的にラベル矩形データを生成し、文字列矩形ファイルに文字情報を付与する。

この為に、画像データの連結成分のラベル付けをする。（参照：ラベル矩形機能）そして、得られたラベル矩形からベースラインと文字サイズを得る。

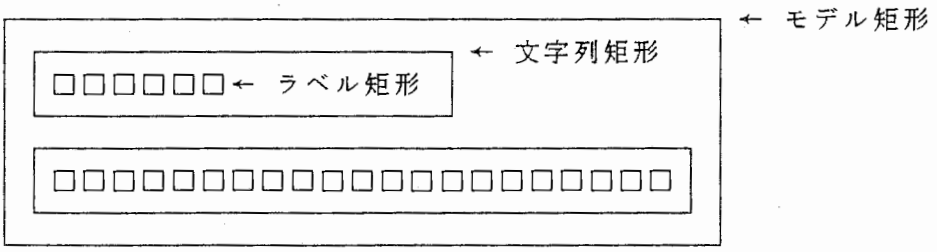
【関連ファイル】

- *.tmp4 — 補正済文書２値画像ファイル
- *.old1 — 文字情報なし旧文字列矩形ファイル
- *.old — 旧文字列矩形ファイル

【参照】

- layout — 矩形データ作成支援プログラム（論文レイアウト抽出システム）
- laydan — 段組生成プログラム（論文レイアウト抽出システム）

ラベル矩形ファイルをモデル矩形ファイルで区切られている範囲まで、横に連結する。横に連結されたものは、文字列としてまとめる。この時、この文字列の番号を連結したラベル矩形に付ける。



【プログラム名】

laystr ― ラベル矩形から文字列矩形を生成する（文字列矩形生成プログラム）

【ハードウェア】

・ＳＵＮ３／２６０Ｃ

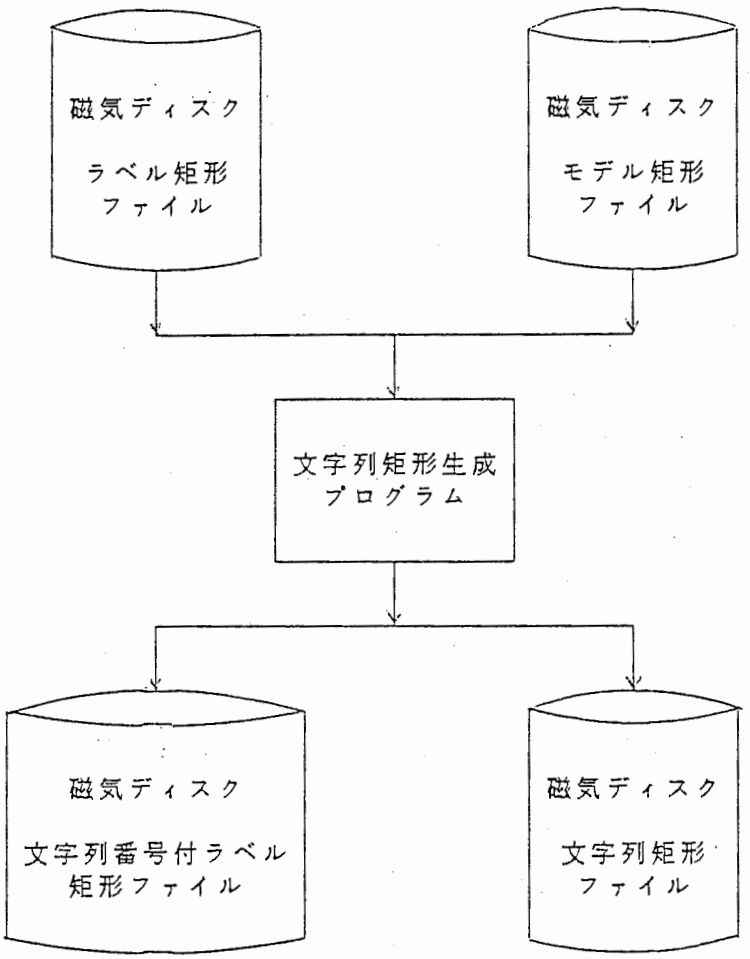
【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・シェループログラム

・Ｓｕｎ Ｃｏｍｍｏｎ Ｌｉｓｐ（Ｖｅｒｓｉｏｎ ２．１．１）

【データフロー】



【構文】

laystr 〈ラベル矩形ファイル〉 〈モデル矩形ファイル〉

〈文字列番号付ラベル矩形ファイル〉 〈文字列矩形ファイル〉

【詳細】

ｌａｙｓｔｒは、ラベル矩形データをモデル矩形データで区切られている範囲内で横に連結する。

横に連結されたものは、文字列としてまとめて、文字列矩形ファイルとして出力する。さらに、ラベル矩形データには、この文字列の番号を付与する。

【関連ファイル】

- *.lbl — ８連結ラベル矩形ファイル
- *.mdl — モデル矩形ファイル
- *.lblstr — 文字列番号付８連結ラベル矩形ファイル
- *.str — 文字列矩形ファイル

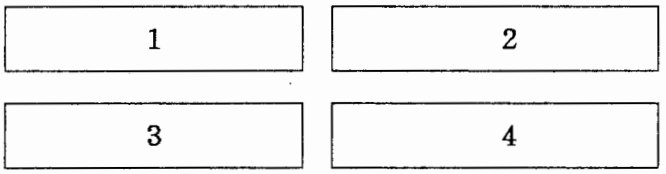
【参照】

- lblrec — ラベル矩形プログラム（論文レイアウト抽出システム）
- layout — 矩形データ作成支援プログラム（論文レイアウト抽出システム）
- lbltodb — ラベル矩形ファイル変換プログラム（論文レイアウト抽出システム）
- laydan — 段組生成プログラム（論文レイアウト抽出システム）

６－６．ノード番号付与機能

文字列矩形ファイルとモデル矩形ファイルを入力して木構造を生成する。そして、この木構造の各ノードに対するノード番号を付ける。

同じレベルでのノード番号は、以下のようになる。（つまり、左から右、上から下の順になる。）



【プログラム名】

laynum ― ブロック木構造に識別番号を付ける（番号付与プログラム）

【ハードウェア】

・ＳＵＮ３／２６０Ｃ

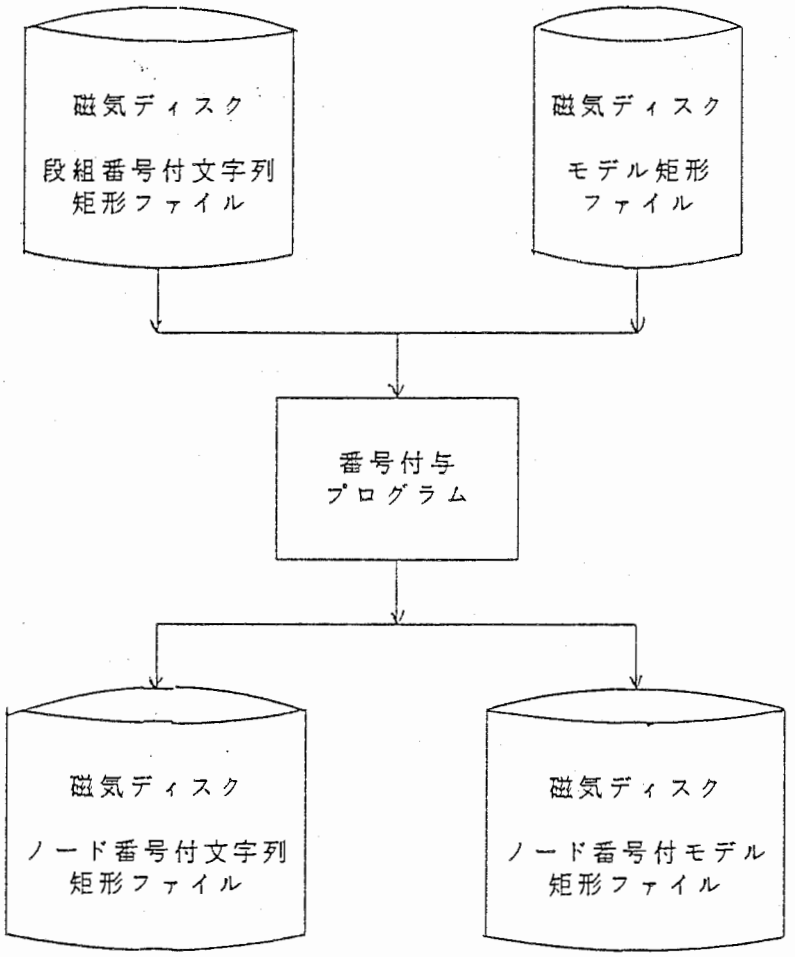
【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・シェループログラム

・Ｓｕｎ Ｃｏｍｍｏｎ Ｌｉｓｐ（Ｖｅｒｓｉｏｎ ２．１．１）

【データフロー】



【構文】

laynum 〈段組番号付文字列矩形ファイル〉 〈モデル矩形ファイル〉

〈ノード番号付文字列矩形ファイル〉 〈ノード番号付モデル矩形ファイル〉

【詳細】

段組番号付文字列矩形ファイルから文字列矩形データを読み、モデル矩形ファイルからモデル矩形データを読みこんで、本来のブロック木構造に戻す。

そして、このブロック木構造の各ノードに番号付けを施し、ノード番号付文字列矩形ファイルとノード番号付モデル矩形ファイルを作る。

【関連ファイル】

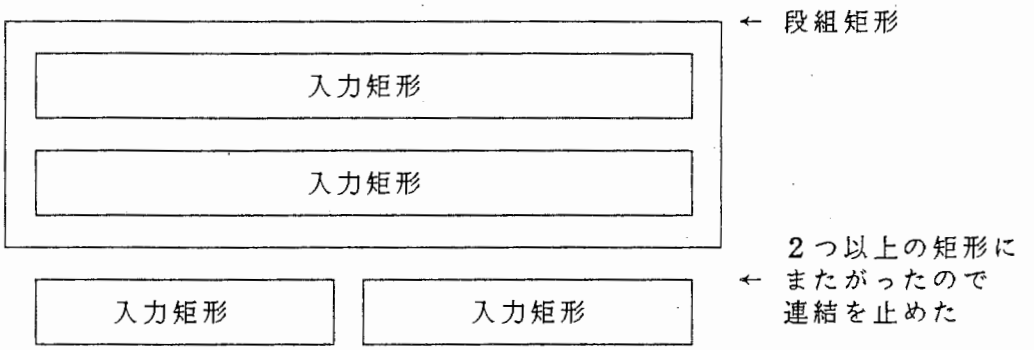
- *.strdan — 段組番号付文字列矩形ファイル
- *.olddan — 段組番号付旧文字列矩形ファイル
- *.mdl — モデル矩形ファイル
- *.strmdl — ノード番号付文字列矩形ファイル
- *.oldmdl — ノード番号付旧文字列矩形ファイル
- *.mdlstr — 文字列矩形のノード番号付モデル矩形ファイル

【参照】

- layout — 矩形データ作成支援プログラム（論文レイアウト抽出システム）
- laydan — 段組生成プログラム（論文レイアウト抽出システム）
- strtodb — 文字列矩形ファイル変換プログラム（論文レイアウト抽出システム）

6－7．段組矩形生成機能

矩形ファイルを入力し、縦方向に連結する。この時、２つ以上の矩形にまたがる場合は、そこで連結を止める。そして、全ての入力矩形を段組矩形で囲む。



【プログラム名】

laydan ― 各種矩形ファイルから段組矩形ファイルを生成（段組生成プログラム）

【ハードウェア】

・ＳＵＮ３／２６０Ｃ

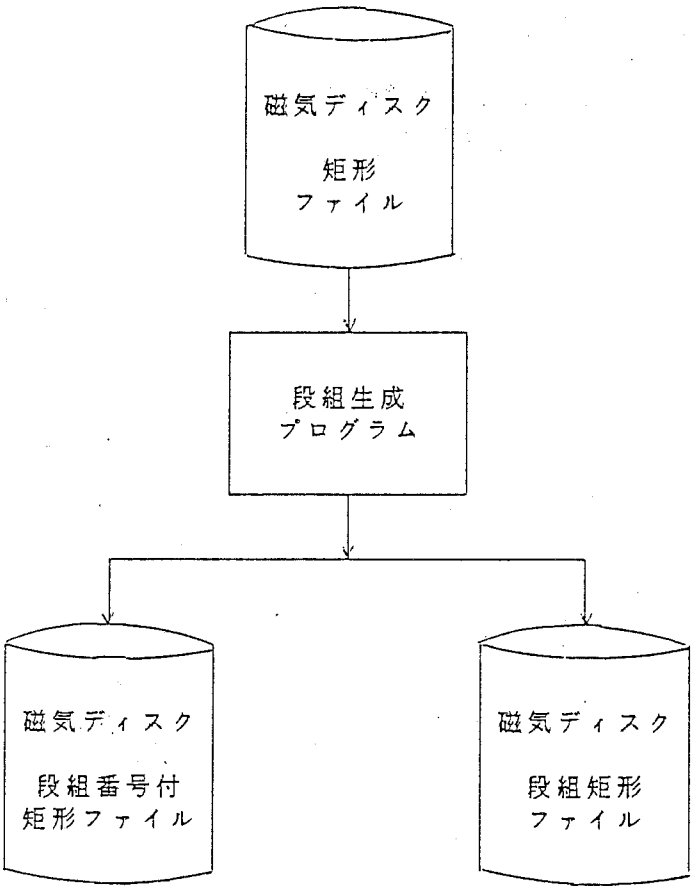
【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・シェループログラム

・Ｓｕｎ Ｃｏｍｍｏｎ Ｌｉｓｐ（Ｖｅｒｓｉｏｎ ２．１．１）

【データフロー】



【構文】

laydan 〈矩形ファイル〉 〈段組矩形ファイル〉 〈段組番号付矩形ファイル〉

【詳細】

ｌａｙｄａｎは、矩形ファイルを入力して、各矩形を縦方向に連結する。この時、２つ以上の矩形にまたがる場合は、そこで連結を止める。

連結した矩形は、段組矩形として出力する。

【関連ファイル】

- *.str — 文字列矩形ファイル
- *.strdan — 段組番号付文字列矩形ファイル
- *.danstr — 段組矩形ファイル
- *.old — 旧文字列矩形ファイル
- *.olddan — 段組番号付旧文字列矩形ファイル
- *.danold — 段組矩形ファイル
- *.mdlstr — 文字列矩形のノード番号付モデル矩形ファイル
- *.mdlstrdan — 段組済モデル矩形ファイル
- *.danmdlstr — 段組矩形ファイル
- *.mdlold — 旧文字列矩形のノード番号付モデル矩形ファイル
- *.mdlolddan — 段組済モデル矩形ファイル
- *.danmdlold — 段組矩形ファイル

【参照】

- laystr — 文字列矩形生成プログラム（論文レイアウト抽出システム）
- layout — 矩形データ作成支援プログラム（論文レイアウト抽出システム）
- laynum — 番号付与プログラム（論文レイアウト抽出システム）

６－８．ラベル矩形フォーマット変換機能

いくつかのラベル矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

変換形式は以下の通り。

※入力ファイル

〈ラベル矩形ファイル〉 ：：＝ {〈行〉〈改行文字〉}*

〈行〉 ：：＝ 〈開括弧〉〈左端座標〉〈上端座標〉
〈右端座標〉〈下端座標〉〈閉括弧〉
〈矩形番号〉〈文字列番号〉

〈改行文字〉 ：：＝ {
}

※出力ファイル

バイト	1	13	23	33	43	53	/
桁数	12	10	10	10	10	10	/
意味	ファイル名	左端座標	上端座標	右端座標	下端座標	文字列番号	/

/

63
1
改行文字

/

【プログラム名】

lbltodb ― ラベル矩形ファイルをＤＢ登録用に変換する（ラベル矩形ファイル変換プログラム）

【ハードウェア】

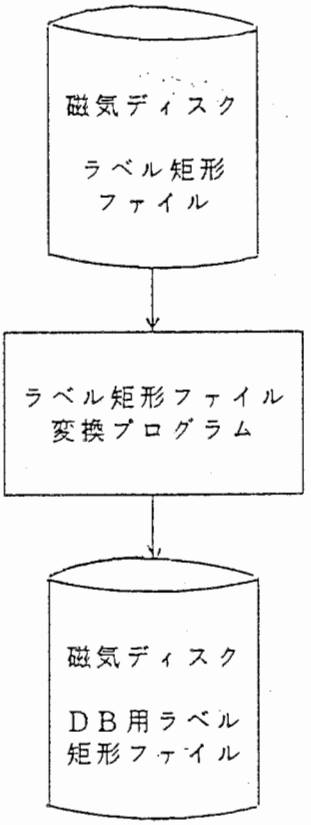
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・シェループログラム

【データフロー】



【構文】

lbltodb 〈ラベル矩形ファイル〉…

【詳細】

ｌｂｌｔｏｄｂは、いくつかのラベル矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

【関連ファイル】

・*.lblstr ― 文字列番号付８連結ラベル矩形ファイル

・*.lbltbl ― データベース用８連結ラベル矩形ファイル

【参照】

・laystr ― 文字列矩形生成プログラム（論文レイアウト抽出システム）

・sqlplus ― リレーショナルデータベース操作言語（ＯＲＡＣＬＥマニュアル）

・awk ― パターン検索処理言語（ＳＵＮ３マニュアル）

6－9．文字列矩形フォーマット変換機能

いくつかの文字列矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

変換形式は以下の通り。

※入力ファイル

〈文字列矩形ファイル〉 ：：＝ {〈行〉〈改行文字〉}*

〈行〉 ：：＝ 〈開括弧〉〈左端座標〉〈上端座標〉
〈右端座標〉〈下端座標〉〈閉括弧〉
〈縦棒〉〈文字コード列〉〈縦棒〉
〈文字サイズ〉〈ベースライン座標〉〈文字列番号〉
〈開括弧〉〈ノードレベル〉〈ノード番号〉〈閉括弧〉
〈段組矩形番号〉

〈改行文字〉 ：：＝ {
}

※出力ファイル

バイト	1	13	23	33	43	53
桁数	12	10	10	10	10	10
値	ファイル名	文字列番号	左端座標	上端座標	右端座標	下端座標

63	73	83	93	103
10	10	10	10	10
文字サイズ	ベースライン座標	ノードレベル	ノード番号	段組矩形番号

113	313
200	1
文字コード列	改行文字

【プログラム名】

strtodb ― 文字列矩形ファイルをＤＢ登録用に変換する（文字列矩形ファイル変換プログラム）

【ハードウェア】

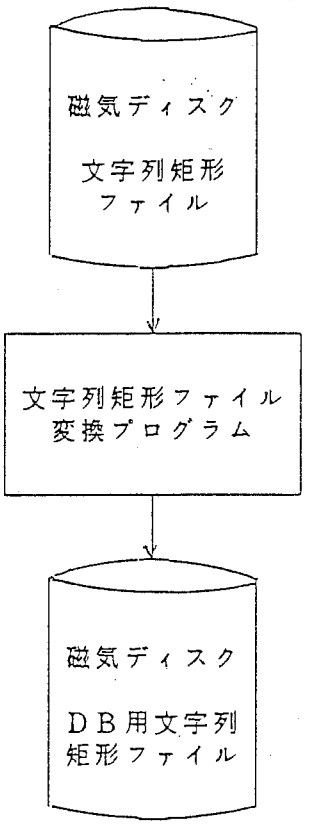
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・シェループログラム

【データフロー】



【構文】

strtodb 〈文字列矩形ファイル〉…

【詳細】

ｓｔｒｔｏｄｂは、いくつかの文字列矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

【関連ファイル】

・*.strmdl — ノード番号付文字列矩形ファイル

・*.strtbl — データベース用文字列矩形ファイル

・*.oldmdl — ノード番号付旧文字列矩形ファイル

・*.oldtbl — データベース用旧文字列矩形ファイル

【参照】

・laynum — 番号付与プログラム（論文レイアウト抽出システム）

・sqlplus — リレーショナルデータベース操作言語（ＯＲＡＣＬＥマニュアル）

・awk — パターン検索処理言語（ＳＵＮ３マニュアル）

６－１０．モデル矩形フォーマット変換機能

いくつかのモデル矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

変換形式は以下の通り。

※入力ファイル

〈モデル矩形ファイル〉 ：：＝ {〈行〉〈改行文字〉}*

〈行〉 ：：＝ 〈開括弧〉〈左端座標〉〈上端座標〉
〈右端座標〉〈下端座標〉〈閉括弧〉
〈ブロックタイプ名〉〈開括弧〉〈属性〉〈閉括弧〉
〈開括弧〉〈自番号〉〈閉括弧〉〈開括弧〉〈親番号〉〈閉括弧〉
〈平均文字サイズ〉〈平均行ピッチ〉
〈開括弧〉〈全体インデント〉〈閉括弧〉
〈開括弧〉〈行インデント〉〈閉括弧〉

〈改行文字〉 ：：＝ {
}

※出力ファイル

バイト	1	13	23	33	43
桁数	12	10	10	10	10
値	ファイル名	左端座標	上端座標	右端座標	下端座標

53	63	73	83
10	10	10	10
自ノードレベル	自ノード番号	親ノードレベル	親ノード番号

93	103	113	123
16	10	10	1
ブロックタイプ名	平均文字サイズ	平均行ピッチ	改行文字

【プログラム名】

mdltodb ― モデル矩形ファイルをＤＢ登録用に変換する（モデル矩形ファイル変換プログラム）

【ハードウェア】

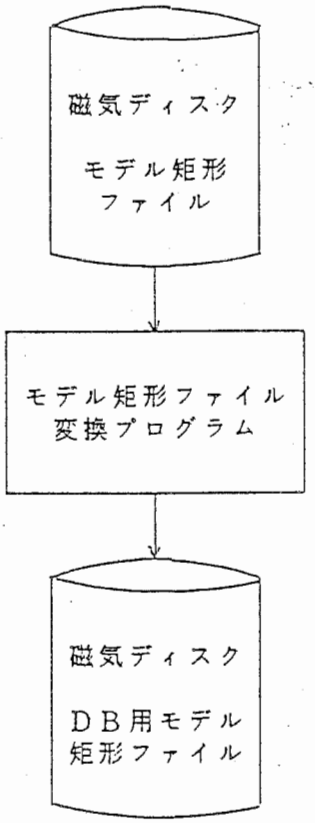
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・シェループログラム

【データフロー】



【構文】

mdltodb 〈モデル矩形ファイル〉…

【詳細】

ｍｄｌｔｏｄｂは、いくつかのモデル矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

【関連ファイル】

・*.mdlstrdan — 段組済モデル矩形ファイル

・*.mdlstrtbl — データベース用モデル矩形ファイル

・*.mdlolddan — 段組済モデル矩形ファイル

・*.mdloldtbl — データベース用モデル矩形ファイル

【参照】

・laydan — 段組生成プログラム（論文レイアウト抽出システム）

・sqlplus — リレーショナルデータベース操作言語（ＯＲＡＣＬＥマニュアル）

・awk — パターン検索処理言語（ＳＵＮ３マニュアル）

６－１１．段組矩形フォーマット変換機能

いくつかの段組矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

変換形式は以下の通り。

※入力ファイル

〈段組矩形ファイル〉 ：：＝ {〈行〉〈改行文字〉}*

〈行〉 ：：＝ 〈開括弧〉〈左端座標〉〈上端座標〉
〈右端座標〉〈下端座標〉〈閉括弧〉〈段組番号〉

〈改行文字〉 ：：＝ {
}

※出力ファイル

バイト	1	13	23	33	43	53	
桁数	12	10	10	10	10	10	
値	ファイル名	段組番号	左端座標	上端座標	右端座標	下端座標	

63
1
改行文字

【プログラム名】

dantodb ― 段組矩形ファイルをＤＢ登録用に変換する（段組矩形ファイル変換プログラム）

【ハードウェア】

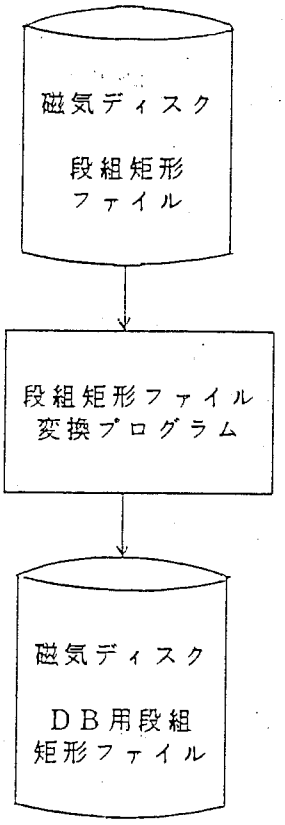
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・シェループログラム

【データフロー】



【構文】

dantodb 〈段組矩形ファイル〉…

【詳細】

ｄａｎｔｏｄｂは、いくつかの段組矩形ファイルをまとめて固定長に変換し、標準出力に出力する。

【関連ファイル】

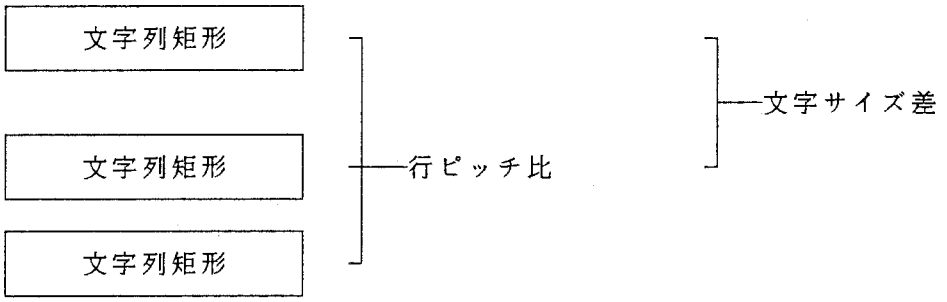
- *.danstr — 段組矩形ファイル
- *.danstrtbl — データベース用段組矩形ファイル
- *.danold — 段組矩形ファイル
- *.danoldtbl — データベース用段組矩形ファイル
- *.danmdlstr — 段組矩形ファイル
- *.danmdlstrtbl — データベース用段組矩形ファイル
- *.mdlold — 段組矩形ファイル
- *.danmdloldtbl — データベース用段組矩形ファイル

【参照】

- laydan — 段組生成プログラム（論文レイアウト抽出システム）
- sqlplus — リレーショナルデータベース操作言語（ＯＲＡＣＬＥマニュアル）
- awk — パターン検索処理言語（ＳＵＮ３マニュアル）

6－12．評価データ算出機能

データベースの文字列矩形テーブルを取り出し、上下行を比較して文字サイズ差や行ピッチ比を求める。



【プログラム名】

dbtoeval ― 文字列矩形テーブルの評価データを算出（評価データ算出プログラム）

【ハードウェア】

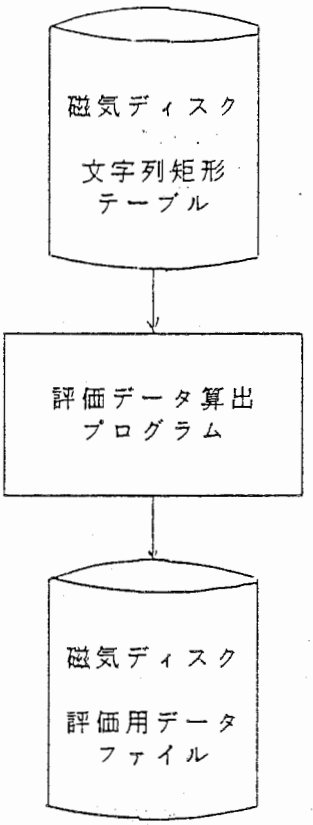
・ＳＵＮ３／２６０Ｃ

【ソフトウェア】

・Ｓｕｎ／ＪＮＩＸ４．２ＢＳＤ／３．２ＥＸＰＯＲＴ

・Ｐｒｏ＊Ｃ

【データフロー】



【構文】

dbtoeval

【詳細】

ｄｂｔｏｅｖａｌは、データベースの文字列矩形テーブルを取り出し、上下行を比較して文字サイズ差や行ピッチを求める。

【関連ファイル】

・GYO_INFO — データベース内文字列矩形テーブル

・*.evltbl — 評価用データファイル

【参照】

・sqlplus — リレーショナルデータベース操作言語（ＯＲＡＣＬＥマニュアル）

・pcc — Ｐｒｏ＊Ｃコンパイラ（ＯＲＡＣＬＥマニュアル）

「付録」

データベース情報の概要

［１］テーブル一覧

JOUHOU: 紙面の書誌情報（ex.タイトル、著者名）を記述した論理情報記述。検索情報抽出システムが出力すべき情報である。

MODEL_INFO:ユーザが指示入力した紙面モデル。紙面上の論理情報を囲む矩形をマウスにより指示した座標をモデル情報構造化プログラムが処理した結果。

MODEL_ATT: モデルの各矩形に対して、その矩形領域に含まれる論理情報の属性の記述。

IJOU_DATA: モデル矩形と文字列矩形の対応が取れない異常データの表。

GYO_INFO: 紙面の各文字列を囲む最小矩形座標。MODEL_INFOテーブル中の対応する論理情報領域（論理ブロックと呼ぶ）へのポインタを持つ。さらに、後述の段組情報テーブルDANGUMI_INFO中の対応する段へのポインタも持つ。

COMPONENT:紙面上の黒画素８連結矩形を囲む最小矩形の座標。GYO_INFO中の対応する文字列へのポインタを持つ。

DANGUMI_INFO:GYO_INFO中の各文字列から段組抽出プログラムにより求めた、各段を囲む最小矩形座標。

EVAL_TABLE:DANGUMI_INFOの各段組に含まれる文字列に対して、上の文字との文字サイズ差、上下の文字列との行間隔の比を求めたもの。

[2]テーブルのcolumn一覧

データベースの詳細および使用例については、別途テクニカルレポート「文書画像データベース説明書」にて記載する。

'JOUHOU'

CNAME	COLTYP	WIDTH
FILENAME	CHAR	12
TITLE1	CHAR	80
TITLE2	CHAR	80
TITLE3	CHAR	44
ETITLE1	CHAR	80
ETITLE2	CHAR	80
ETITLE3	CHAR	80
ETITLE4	CHAR	60
AUTHOR1	CHAR	40
EAUTHOR1	CHAR	30
SHOZOKU1	CHAR	100

CNAME	COLTYP	WIDTH
ESHOZOKU1	CHAR	200
AUTHOR2	CHAR	40
EAUTHOR2	CHAR	30
SHOZOKU2	CHAR	100
ESHOZOKU2	CHAR	200
AUTHOR3	CHAR	40
EAUTHOR3	CHAR	30
SHOZOKU3	CHAR	100
ESHOZOKU3	CHAR	200
AUTHOR4	CHAR	40
EAUTHOR4	CHAR	30

CNAME	COLTYP	WIDTH
SHOZOKU4	CHAR	100
ESHOZOKU4	CHAR	200
AUTHOR5	CHAR	40
EAUTHOR5	CHAR	30
SHOZOKU5	CHAR	100
ESHOZOKU5	CHAR	200
AUTHOR6	CHAR	40
EAUTHOR6	CHAR	30
SHOZOKU6	CHAR	100
ESHOZOKU6	CHAR	200
AUTHOR7	CHAR	40

CNAME	COLTYP	WIDTH
EAUTHOR7	CHAR	30
SHOZOKU7	CHAR	100
ESHOZOKU7	CHAR	200
AUTHOR8	CHAR	40
EAUTHOR8	CHAR	30
SHOZOKU8	CHAR	100
ESHOZOKU8	CHAR	200
AUTHOR9	CHAR	40
EAUTHOR9	CHAR	30
SHOZOKU9	CHAR	100
ESHOZOKU9	CHAR	200

CNAME	COLTYP	WIDTH
AUTHOR10	CHAR	40

EAUTHOR10	CHAR	30
SHOZOKU10	CHAR	100
ESHOZOKU10	CHAR	200
MAGAZINE	CHAR	60
VOLUME	CHAR	6
GNUMBER	CHAR	4
PDATE	CHAR	8
LANGUAGE	CHAR	2
COUNTRY	CHAR	3
CHART	CHAR	3

CNAME	COLTYP	WIDTH
THESIS	CHAR	3
KEYWORD1	CHAR	50
KEYWORD2	CHAR	50
KEYWORD3	CHAR	50
KEYWORD4	CHAR	50
KEYWORD5	CHAR	50
KEYWORD6	CHAR	50
KEYWORD7	CHAR	50
KEYWORD8	CHAR	50
KEYWORD9	CHAR	50
KEYWORD10	CHAR	50

CNAME	COLTYP	WIDTH
EXCERPT1	CHAR	80
EXCERPT2	CHAR	80
EXCERPT3	CHAR	80
EXCERPT4	CHAR	80
EXCERPT5	CHAR	80
EXCERPT6	CHAR	80
EXCERPT7	CHAR	80
EXCERPT8	CHAR	80
EXCERPT9	CHAR	80
EXCERPT10	CHAR	80
EXCERPT11	CHAR	80

CNAME	COLTYP	WIDTH
EXCERPT12	CHAR	80
EXCERPT13	CHAR	80
EXCERPT14	CHAR	80
EXCERPT15	CHAR	80
EXCERPT16	CHAR	80
EXCERPT17	CHAR	80
EXCERPT18	CHAR	80
EXCERPT19	CHAR	80
EXCERPT20	CHAR	80
CLASS	CHAR	30
PAGE	CHAR	6

CNAME	COLTYP	WIDTH
ARTICLE	CHAR	12
THANKS1	CHAR	80
THANKS2	CHAR	80
THANKS3	CHAR	80
THANKS4	CHAR	80
THANKS5	CHAR	80
THANKS6	CHAR	80
THANKS7	CHAR	80
THANKS8	CHAR	80
THANKS9	CHAR	80
THANKS10	CHAR	80

```
CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
THANKS11                       CHAR           80
THANKS12                       CHAR           80
THANKS13                       CHAR           80
THANKS14                       CHAR           30

103 records selected.

SQL> change /JOUHOU/MODEL_INFO/
  2* where tname='MODEL_INFO'
SQL> run
  1  select cname,coltype,width from columns
  2* where tname='MODEL_INFO'

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
FILENAME                       CHAR           12
LEFTTOP_X                      NUMBER         10
LEFTTOP_Y                      NUMBER         10
RIGHTBOTTOM_X                  NUMBER         10
RIGHTBOTTOM_Y                  NUMBER         10
NODE_LEVEL                     NUMBER         10
NODE_NO                        NUMBER         10
PNODE_LEVEL                    NUMBER         10
PNODE_NO                       NUMBER         10
BLOCK_NAME                     CHAR           16
MOJI_SIZE                      NUMBER         10

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
GYO_PITCH                      NUMBER         10

12 records selected.

SQL> change /MODEL_INFO/MODEL_ATT/
  2* where tname='MODEL_ATT'
SQL> run
  1  select cname,coltype,width from columns
  2* where tname='MODEL_ATT'

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
FILENAME                       CHAR           12
NODE_LEVEL                     NUMBER         10
NODE_NO                        NUMBER         10
ATTRIBUTE_NO                   NUMBER         10

SQL> change /MODEL_ATT/IJOU_DATA/
  2* where tname='IJOU_DATA'
SQL> run
  1  select cname,coltype,width from columns
  2* where tname='IJOU_DATA'

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
FILENAME                       CHAR           12

SQL> change /IJOU_DATA/GYO_INFO/
  2* where tname='GYO_INFO'
SQL> run
  1  select cname,coltype,width from columns
  2* where tname='GYO_INFO'

CNAME                          COLTYP      WIDTH
```

```
FILENAME                       CHAR            12
REC_NO                         NUMBER          10
LEFTTOP_X                      NUMBER          10
LEFTTOP_Y                      NUMBER          10
RIGHTBOTTOM_X                  NUMBER          10
RIGHTBOTTOM_Y                  NUMBER          10
MOJI_SIZE                      NUMBER          10
BASE_LINE                      NUMBER          10
PNODE_LEVEL                    NUMBER          10
PNODE_NO                       NUMBER          10
DNODE_NO                       NUMBER          10

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
MOJI_RETSU                     CHAR           200

12 records selected.

SQL> change /GYO_INFO/COMPONENT/
  2* where tname='COMPONENT'
SQL> run
  1  select cname, coltype, width from columns
  2* where tname='COMPONENT'

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
FILENAME                       CHAR            12
REC_NO                         NUMBER          10
LEFTTOP_X                      NUMBER          10
LEFTTOP_Y                      NUMBER          10
RIGHTBOTTOM_X                  NUMBER          10
RIGHTBOTTOM_Y                  NUMBER          10

6 records selected.

SQL> change /COMPONENT/DANGUMI_INFO/
  2* where tname='DANGUMI_INFO'
SQL> run
  1  select cname, coltype, width from columns
  2* where tname='DANGUMI_INFO'

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
FILENAME                       CHAR            12
DNODE_NO                       NUMBER          10
LEFTTOP_X                      NUMBER          10
LEFTTOP_Y                      NUMBER          10
RIGHTBOTTOM_X                  NUMBER          10
RIGHTBOTTOM_Y                  NUMBER          10

6 records selected.

SQL> change /DANGUMI_INFO/EVAL_TABLE/
  2* where tname='EVAL_TABLE'
SQL> run
  1  select cname, coltype, width from columns
  2* where tname='EVAL_TABLE'

CNAME                          COLTYP      WIDTH
------------------------------ ------ ----------
FILENAME                       CHAR            12
DNODE_NO                       NUMBER          10
REC_NO                         NUMBER          10
MOJI_SIZE_SA                   NUMBER          10
GYO_KAN_HI                     NUMBER          10
```

GYO_JOUGE	NUMBER	10
BASE_LINE	NUMBER	10
BLOCK_NO	NUMBER	10

8 records selected.