(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5340689号
(P5340689)
(45)発行日　平成25年11月13日(2013.11.13)　(24)登録日　平成25年8月16日(2013.8.16)

(51)Int.Cl.		F I		
G06Q 10/00	(2012.01)	G06Q	10/00	140
G06F 17/30	(2006.01)	G06F	17/30	230Z
G06F 12/00	(2006.01)	G06F	12/00	513Z
G06Q 10/10	(2012.01)	G06Q	10/10	130Z

請求項の数 9　(全 23 頁)

(21)出願番号　特願2008-265354 (P2008-265354)
(22)出願日　平成20年10月14日 (2008.10.14)
(65)公開番号　特開2010-97263 (P2010-97263A)
(43)公開日　平成22年4月30日 (2010.4.30)
審査請求日　平成23年9月16日 (2011.9.16)

(73)特許権者　390024350
株式会社ジャストシステム
徳島県徳島市川内町平石若松１０８番地４
(74)代理人　100095407
弁理士　木村　満
(72)発明者　岡　智幸
大阪府大阪市東淀川区東中島１丁目３番１４号　株式会社キーエンス内

審査官　大野　朋也

最終頁に続く

(54)【発明の名称】データベース生成装置、データベース生成方法及びコンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
表データが含まれる一又は複数の電子文書ファイル中から抽出したデータに基づいて新たなデータベースを生成するデータベース生成装置において、
一又は複数の前記電子文書ファイルを取得する電子文書ファイル取得手段と、
取得した一又は複数の前記電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する罫線情報抽出手段と、
抽出した複数の罫線情報に基づいて前記電子文書ファイルの内容を解析する解析手段と、
前記罫線情報に基づいた前記電子文書ファイルの内容の解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定するデータ抽出規則特定手段と、
特定したデータベース項目及びデータ抽出規則にて、一又は複数の前記電子文書ファイルから前記データベース項目及び対応するデータを抽出するデータ抽出手段と、
抽出されたデータベース項目及び対応するデータを一覧表示する表示手段と、
表示されたデータベース項目及び対応するデータが適正である旨を示す確定情報の入力を受け付ける確定情報受付手段と
を備えることを特徴とするデータベース生成装置。
【請求項２】
前記データ抽出規則特定手段は、

一又は複数の前記電子文書ファイルの指定を受け付けるファイル指定受付手段と、
指定を受け付けた一又は複数の電子文書ファイルの表データの範囲指定を受け付ける範囲指定受付手段と、
受け付けた範囲指定に従って、前記データベース項目及び前記データ抽出規則を特定する特定手段と
を備えることを特徴とする請求項１記載のデータベース生成装置。
【請求項３】
異なる電子文書ファイルから抽出した表データの位置の相違に関する情報、及び／又は異なる電子文書ファイルから抽出したデータベース項目の相違に関する情報を少なくとも含む表データの相違に関するゆらぎ情報を抽出するゆらぎ情報抽出手段と、
前記データベース項目、前記データ抽出規則及び抽出されたゆらぎ情報に基づいて、前記データ抽出規則の変更部分が存在するか否かを判断する判断手段と、
該判断手段で変更部分が存在すると判断した場合、同一の前記データベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与するタグ情報付与手段と
を備え、
前記データ抽出手段は、前記データ抽出規則の変更部分を反映して前記データベース項目及び対応するデータを抽出するようにしてあり、
前記表示手段は、前記データベース項目に付与されているタグ情報に従って前記データベース項目を配列して、前記データベース項目及び対応するデータを一覧表示するようにしてあることを特徴とする請求項１又は２記載のデータベース生成装置。
【請求項４】
前記解析手段は、
罫線により区切られた区画が複数列又は複数行にわたって同一であるか否かを判断する手段を備え、
該手段で同一であると判断した場合、前記データ抽出手段は、複数列又は複数行にわたって同一である最初の行又は列での区画に相当するデータベース項目にてデータを抽出するようにしてあることを特徴とする請求項１乃至３のいずれか一項に記載のデータベース生成装置。
【請求項５】
表データが含まれる一又は複数の電子文書ファイル中から抽出したデータに基づいて新たなデータベースを生成するデータベース生成装置で実行することが可能なデータベース生成方法において、
前記データベース生成装置は、
一又は複数の前記電子文書ファイルを取得し、
取得した一又は複数の前記電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出し、
抽出した複数の罫線情報に基づいて前記電子文書ファイルの内容を解析し、
前記罫線情報に基づいた前記電子文書ファイルの内容の解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定し、
特定したデータベース項目及びデータ抽出規則にて、一又は複数の前記電子文書ファイルから前記データベース項目及び対応するデータを抽出し、
抽出されたデータベース項目及び対応するデータを一覧表示し、
表示されたデータベース項目及び対応するデータが適正である旨を示す確定情報の入力を受け付けることを特徴とするデータベース生成方法。
【請求項６】
前記データベース生成装置は、
一又は複数の前記電子文書ファイルの指定を受け付け、
指定を受け付けた一又は複数の電子文書ファイルの表データの範囲指定を受け付け、
受け付けた範囲指定に従って、前記データベース項目及び前記データ抽出規則を特定す

ることを特徴とする請求項５記載のデータベース生成方法。
【請求項７】
前記データベース生成装置は、
異なる電子文書ファイルから抽出した表データの位置の相違に関する情報、及び／又は異なる電子文書ファイルから抽出したデータベース項目の相違に関する情報を少なくとも含む表データの相違に関するゆらぎ情報を抽出し、
前記データベース項目、前記データ抽出規則及び抽出されたゆらぎ情報に基づいて、前記データ抽出規則の変更部分が存在するか否かを判断し、
変更部分が存在すると判断した場合、同一の前記データベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与し、
前記データ抽出規則の変更部分を反映して前記データベース項目及び対応するデータを抽出し、
前記データベース項目に付与されているタグ情報に従って前記データベース項目を配列して、前記データベース項目及び対応するデータを一覧表示することを特徴とする請求項５又は６記載のデータベース生成方法。
【請求項８】
前記データベース生成装置は、
罫線により区切られた区画が複数列又は複数行にわたって同一であるか否かを判断し、
同一であると判断した場合、複数列又は複数行にわたって同一である最初の行又は列での区画に相当するデータベース項目にてデータを抽出することを特徴とする請求項５乃至７のいずれか一項に記載のデータベース生成方法。
【請求項９】
表データが含まれる一又は複数の電子文書ファイル中から抽出したデータに基づいて新たなデータベースを生成するデータベース生成装置で実行することが可能なコンピュータプログラムにおいて、
前記データベース生成装置を、
一又は複数の前記電子文書ファイルを取得する電子文書ファイル取得手段、
取得した一又は複数の前記電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する罫線情報抽出手段、
抽出した複数の罫線情報に基づいて前記電子文書ファイルの内容を解析する解析手段、
前記罫線情報に基づいた前記電子文書ファイルの内容の解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定するデータ抽出規則特定手段、
特定したデータベース項目及びデータ抽出規則にて、一又は複数の前記電子文書ファイルから前記データベース項目及び対応するデータを抽出するデータ抽出手段、
抽出されたデータベース項目及び対応するデータを一覧表示する表示手段、及び
表示されたデータベース項目及び対応するデータが適正である旨を示す確定情報の入力を受け付ける確定情報受付手段
として機能させることを特徴とするコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、異なるデータ形式を有する複数の表データが存在する場合であっても、容易に一のデータベースを生成することができるデータベース生成装置、データベース生成方法及びコンピュータプログラムに関する。
【背景技術】
【０００２】
関係データベースを生成する場合、事前に生成されている表データを利用することが多い。表データのデータベース項目が一致している場合には、複数の表データをマージすることにより容易に新たなデータベースを生成することができる。

【０００３】
しかし、表データのデータ形式が標準化されていない場合、表データのデータ形式は作成者に依存しており、また同一のアプリケーションであっても使用するソフトウェアプログラムによってデータベース項目の相違、データベース項目の配列の相違等が存在しており、そのままマージすることができない。斯かる問題を解決するべく、従来は中間ファイルフォーマットを用いて、データベース項目が相違している、あるいはデータベース項目の配列順序が相違している複数の表データをマージして１つの表データを生成している。
【０００４】
例えば特許文献１では、表データの中間ファイルとして良く用いられているＣＳＶファイルを用い、複数の表データファイルから１つのデータベースを生成しているデータベース管理システムが開示されている。
【特許文献１】特開２００６－０５９１３５号公報
【発明の開示】
【発明が解決しようとする課題】
【０００５】
しかし、特許文献１のように、ＣＳＶファイルを介して複数の表データをマージする場合、どのデータベース項目が相違しているか、どのデータベース項目の配列順序が相違しているか等に関する情報を事前に知っておく必要が有り、これらの情報に応じて適切な変換手順を設定しておかないと、所望の表データを生成することができないという問題点があった。
【０００６】
また、表データのセル位置に基づいて自動的に表データをマージする方法も考えられているが、表データの開始セル位置が一定ではなく、開始セル位置の相違に基づいてマージ対象となるセルの位置補正を行わないと、複数の異なるファイル又は異なるシート上の表データを正しくマージすることはできない。
【０００７】
本発明は斯かる事情に鑑みてなされたものであり、異なるデータ形式を有する複数の表データが存在する場合であっても、新たなデータベースを正しく生成することができるデータベース生成装置、データベース生成方法及びコンピュータプログラムを提供することを目的とする。
【課題を解決するための手段】
【０００８】
上記目的を達成するために第１発明に係るデータベース生成装置は、表データが含まれる一又は複数の電子文書ファイル中から抽出したデータに基づいて新たなデータベースを生成するデータベース生成装置において、一又は複数の前記電子文書ファイルを取得する電子文書ファイル取得手段と、取得した一又は複数の前記電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する罫線情報抽出手段と、抽出した複数の罫線情報に基づいて前記電子文書ファイルの内容を解析する解析手段と、前記罫線情報に基づいた前記電子文書ファイルの内容の解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定するデータ抽出規則特定手段と、特定したデータベース項目及びデータ抽出規則にて、一又は複数の前記電子文書ファイルから前記データベース項目及び対応するデータを抽出するデータ抽出手段と、抽出されたデータベース項目及び対応するデータを一覧表示する表示手段と、表示されたデータベース項目及び対応するデータが適正である旨を示す確定情報の入力を受け付ける確定情報受付手段とを備えることを特徴とする。
【０００９】
また、第２発明に係るデータベース生成装置は、第１発明において、前記データ抽出規則特定手段は、一又は複数の前記電子文書ファイルの指定を受け付けるファイル指定受付手段と、指定を受け付けた一又は複数の電子文書ファイルの表データの範囲指定を受け付ける範囲指定受付手段と、受け付けた範囲指定に従って、前記データベース項目及び前記

データ抽出規則を特定する特定手段とを備えることを特徴とする。

【００１０】

また、第３発明に係るデータベース生成装置は、第１又は第２発明において、異なる電子文書ファイルから抽出した表データの位置の相違に関する情報、及び／又は異なる電子文書ファイルから抽出したデータベース項目の相違に関する情報を少なくとも含む表データの相違に関するゆらぎ情報を抽出するゆらぎ情報抽出手段と、前記データベース項目、前記データ抽出規則及び抽出されたゆらぎ情報に基づいて、前記データ抽出規則の変更部分が存在するか否かを判断する判断手段と、該判断手段で変更部分が存在すると判断した場合、同一の前記データベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与するタグ情報付与手段とを備え、前記データ抽出手段は、前記データ抽出規則の変更部分を反映して前記データベース項目及び対応するデータを抽出するようにしてあり、前記表示手段は、前記データベース項目に付与されているタグ情報に従って前記データベース項目を配列して、前記データベース項目及び対応するデータを一覧表示するようにしてあることを特徴とする。

【００１１】

また、第４発明に係るデータベース生成装置は、第１乃至第３発明のいずれか１つにおいて、前記解析手段は、罫線により区切られた区画が複数列又は複数行にわたって同一であるか否かを判断する手段を備え、該手段で同一であると判断した場合、前記データ抽出手段は、複数列又は複数行にわたって同一である最初の行又は列での区画に相当するデータベース項目にてデータを抽出するようにしてあることを特徴とする。

【００１２】

次に、上記目的を達成するために第５発明に係るデータベース生成方法は、表データが含まれる一又は複数の電子文書ファイル中から抽出したデータに基づいて新たなデータベースを生成するデータベース生成装置で実行することが可能なデータベース生成方法において、前記データベース生成装置は、一又は複数の前記電子文書ファイルを取得し、取得した一又は複数の前記電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出し、抽出した複数の罫線情報に基づいて前記電子文書ファイルの内容を解析し、前記罫線情報に基づいた前記電子文書ファイルの内容の解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定し、特定したデータベース項目及びデータ抽出規則にて、一又は複数の前記電子文書ファイルから前記データベース項目及び対応するデータを抽出し、抽出されたデータベース項目及び対応するデータを一覧表示し、表示されたデータベース項目及び対応するデータが適正である旨を示す確定情報の入力を受け付けることを特徴とする。

【００１３】

また、第６発明に係るデータベース生成方法は、第５発明において、前記データベース生成装置は、一又は複数の前記電子文書ファイルの指定を受け付け、指定を受け付けた一又は複数の電子文書ファイルの表データの範囲指定を受け付け、受け付けた範囲指定に従って、前記データベース項目及び前記データ抽出規則を特定することを特徴とする。

【００１４】

また、第７発明に係るデータベース生成方法は、第５又は第６発明において、前記データベース生成装置は、異なる電子文書ファイルから抽出した表データの位置の相違に関する情報、及び／又は異なる電子文書ファイルから抽出したデータベース項目の相違に関する情報を少なくとも含む表データの相違に関するゆらぎ情報を抽出し、前記データベース項目、前記データ抽出規則及び抽出されたゆらぎ情報に基づいて、前記データ抽出規則の変更部分が存在するか否かを判断し、変更部分が存在すると判断した場合、同一の前記データベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与し、前記データ抽出規則の変更部分を反映して前記データベース項目及び対応するデータを抽出し、前記データベース項目に付与されているタグ情報に従って前記データベース項目を配列して、前記データベース項目及び対応するデータを一覧表示することを特徴とする。

【００１５】
また、第８発明に係るデータベース生成方法は、第５乃至第７発明のいずれか１つにおいて、前記データベース生成装置は、罫線により区切られた区画が複数列又は複数行にわたって同一であるか否かを判断し、同一であると判断した場合、複数列又は複数行にわたって同一である最初の行又は列での区画に相当するデータベース項目にてデータを抽出することを特徴とする。
【００１６】
次に、上記目的を達成するために第９発明に係るコンピュータプログラムは、表データが含まれる一又は複数の電子文書ファイル中から抽出したデータに基づいて新たなデータベースを生成するデータベース生成装置で実行することが可能なコンピュータプログラムにおいて、前記データベース生成装置を、一又は複数の前記電子文書ファイルを取得する電子文書ファイル取得手段、取得した一又は複数の前記電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する罫線情報抽出手段、抽出した複数の罫線情報に基づいて前記電子文書ファイルの内容を解析する解析手段、前記罫線情報に基づいた前記電子文書ファイルの内容の解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定するデータ抽出規則特定手段、特定したデータベース項目及びデータ抽出規則にて、一又は複数の前記電子文書ファイルから前記データベース項目及び対応するデータを抽出するデータ抽出手段、抽出されたデータベース項目及び対応するデータを一覧表示する表示手段、及び表示されたデータベース項目及び対応するデータが適正である旨を示す確定情報の入力を受け付ける確定情報受付手段として機能させることを特徴とする。
【００２０】
第１発明、第５発明、及び第９発明では、一又は複数の電子文書ファイルを取得し、取得した一又は複数の電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出して、抽出した複数の罫線情報に基づいて電子文書ファイルの内容を解析する。罫線情報に基づいた電子文書ファイルの内容の解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定し、特定したデータベース項目及びデータ抽出規則にて、一又は複数の電子文書ファイルからデータベース項目及び対応するデータを抽出する。抽出されたデータベース項目及び対応するデータを一覧表示し、表示されたデータベース項目及び対応するデータが適正である旨を示す確定情報の入力を受け付ける。罫線情報に基づいて電子文書ファイル中の表データの位置を特定することができ、新たに生成するデータベースの基礎となるデータベース項目及びデータを抽出するデータ抽出規則を正しく特定することができる。
【００２１】
第２発明及び第６発明では、一又は複数の電子文書ファイルの指定を受け付け、指定を受け付けた一又は複数の電子文書ファイルの表データの範囲指定を受け付ける。受け付けた範囲指定に従って、データベース項目及びデータ抽出規則を特定する。これにより、新たに生成するデータベースに使用するデータベース項目及びデータ抽出規則に不備が生じた場合であっても、ユーザの範囲指定により適切なデータベース項目及びデータ抽出規則を特定することができ、表データをマージすることを半自動化することができる。
【００２２】
第３発明及び第７発明では、異なる電子文書ファイルから抽出した表データの位置の相違に関する情報、及び／又は異なる電子文書ファイルから抽出したデータベース項目の相違に関する情報を少なくとも含む表データの相違に関するゆらぎ情報を抽出する。データベース項目、データ抽出規則及び抽出されたゆらぎ情報に基づいて、データ抽出規則の変更部分が存在するか否かを判断し、変更部分が存在すると判断した場合、同一のデータベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与する。データ抽出規則の変更部分を反映してデータベース項目及び対応するデータを抽出し、データベース項目に付与されているタグ情報に従ってデータベース項目を配列して、データベース項目及び対応するデータを一覧表示する。これにより、複数のファイル上で表

データの位置が相違する場合、データベース項目が相違する場合、データベース項目の配列順序が相違する場合等であっても、斯かる相違に起因して変更されたデータ抽出規則に従ってデータを抽出し、同一のデータベース項目については同一のタグ情報をキー情報として集約することができ、新たな異なるデータベース項目については、異なるタグ情報により新規のデータベース項目として追加することができる。したがって、ユーザがデータベース項目の相違点を事前にすべて知ることができない場合であっても、データベース項目が重複又は欠落することなく新たなデータベースを生成して一覧表示することが可能となる。

【００２３】

第４発明及び第８発明では、罫線により区切られた区画が複数列又は複数行にわたって同一であるか否かを判断し、同一であると判断した場合、複数列又は複数行にわたって同一である最初の行又は列での区画に相当するデータベース項目にてデータを抽出する。これにより、連続して区画が同一である場合には一の項目がさらに複数の項目に分割される可能性が少ないことから、各データベース項目に対応するデータを漏れなく抽出することが可能となる。

【発明の効果】

【００２４】

本発明によれば、罫線情報に基づいて電子文書ファイル中の表データの位置を特定することができ、新たに生成するデータベースの基礎となるデータベース項目及び対応するデータを抽出するデータ抽出規則を正しく特定することができる。

【発明を実施するための最良の形態】

【００２５】

以下、本発明の実施の形態に係るデータベース生成装置について、図面に基づいて具体的に説明する。以下の実施の形態は、特許請求の範囲に記載された発明を限定するものではなく、実施の形態の中で説明されている特徴的事項の組み合わせの全てが解決手段の必須事項であるとは限らないことは言うまでもない。

【００２６】

また、本発明は多くの異なる態様にて実施することが可能であり、実施の形態の記載内容に限定して解釈されるべきものではない。実施の形態を通じて同じ要素には同一の符号を付している。

【００２７】

以下の実施の形態では、コンピュータシステムにコンピュータプログラムを導入したデータベース生成装置について説明するが、当業者であれば明らかな通り、本発明はその一部をコンピュータで実行することが可能なコンピュータプログラムとして実施することができる。したがって、本発明は、データベース生成装置というハードウェアとしての実施の形態、ソフトウェアとしての実施の形態、又はソフトウェアとハードウェアとの組み合わせの実施の形態をとることができる。コンピュータプログラムは、ハードディスク、ＤＶＤ、ＣＤ、光記憶装置、磁気記憶装置等の任意のコンピュータで読み取ることが可能な記録媒体に記録することができる。

【００２８】

（実施の形態１）

図１は、本発明の実施の形態１に係るデータベース生成装置の構成例を示すブロック図である。本発明の実施の形態１に係るデータベース生成装置１は、少なくともＣＰＵ（中央演算装置）１１、メモリ１２、記憶装置１３、Ｉ／Ｏインタフェース１４、ビデオインタフェース１５、可搬型ディスクドライブ１６、通信インタフェース１７及び上述したハードウェアを接続する内部バス１８で構成されている。

【００２９】

ＣＰＵ１１は、内部バス１８を介してデータベース生成装置１の上述したようなハードウェア各部と接続されており、上述したハードウェア各部の動作を制御するとともに、記憶装置１３に記憶されているコンピュータプログラム１００に従って、種々のソフトウェ

ア的機能を実行する。メモリ１２は、ＳＲＡＭ、ＳＤＲＡＭ等の揮発性メモリで構成され、コンピュータプログラム１００の実行時にロードモジュールが展開され、コンピュータプログラム１００の実行時に発生する一時的なデータ等を記憶する。

【００３０】

記憶装置１３は、内蔵される固定型記憶装置（ハードディスク）、ＳＲＡＭ等の揮発性メモリ、ＲＯＭ等の不揮発性メモリ等で構成されている。記憶装置１３に記憶されているコンピュータプログラム１００は、プログラム及びデータ等の情報を記録したＤＶＤ、ＣＤ－ＲＯＭ等の可搬型記録媒体９０から、可搬型ディスクドライブ１６によりダウンロードされ、実行時には記憶装置１３からメモリ１２へ展開して実行される。もちろん、通信インタフェース１７を介してネットワーク２に接続されている外部のコンピュータからダウンロードされたコンピュータプログラムであっても良い。

【００３１】

また記憶装置１３は、電子文書ファイル記憶部１３１、データ抽出規則記憶部１３２、データベース記憶部１３３及びゆらぎ情報記憶部１３４を備えている。電子文書ファイル記憶部１３１には、表データを内容に含み、新たなデータベースを生成するための基礎となる電子文書ファイルを記憶する。

【００３２】

データ抽出規則記憶部１３２には、例えば電子文書ファイルに含まれる表データのうち最大のサイズを有する表データを選択する、ファイルの先頭からｎ（ｎは自然数）番目の表データを選択する等の、表データからデータベース項目及び対応するデータを抽出するデータ抽出規則を記憶している。

【００３３】

データベース記憶部１３３には、複数の電子文書ファイルに含まれる表データをマージして新たに生成されたデータベースを記憶する。ゆらぎ情報記憶部１３４には、マージする対象となる表データ間の相違に関する情報、いわゆるゆらぎ情報を記憶する。ゆらぎ情報としては、例えば表データの開始セルの位置の相違に関する表位置ゆらぎ情報、表データの項目の順序が相違する、新規項目の存在、項目の抜けの存在等の項目の相違に関する項目ゆらぎ情報等がある。また、英語表記での大文字と小文字との相違、全角と半角との相違等も含む広い概念である。

【００３４】

通信インタフェース１７は内部バス１８に接続されており、インターネット、ＬＡＮ、ＷＡＮ等の外部のネットワーク２に接続されることにより、外部のコンピュータ等とデータ送受信を行うことが可能となっている。電子文書ファイル記憶部１３１は、データベース生成装置１の記憶装置１３に備えることに限定されるものではなく、外部のコンピュータの記憶装置に記憶されることによりネットワーク２上に点在していても良い。

【００３５】

Ｉ／Ｏインタフェース１４は、キーボード２１、マウス２２等のデータ入力媒体と接続され、データの入力を受け付ける。また、ビデオインタフェース１５は、ＣＲＴモニタ、ＬＣＤ等の表示装置２３と接続され、所定の画像を表示する。

【００３６】

図２は、本発明の実施の形態１に係るデータベース生成装置１の機能ブロック図である。電子文書ファイル取得部２０１は、一又は複数の表データを含む電子文書ファイルを取得する。電子文書ファイルは、記憶装置１３内に記憶されている電子文書ファイルを電子文書ファイル記憶部１３１に集約しても良いし、ネットワーク２を介して外部のコンピュータから取得しても良い。また、キーボード２１、マウス２２等の入力装置を介して入力しても良い。

【００３７】

罫線情報抽出部２０２は、取得した一又は複数の電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する。具体的には、罫線で囲まれている部分を表データと認識し、その他の罫線がどのように配置されているかに関する情報を取得する。

【００３８】
解析部２０３は、抽出した複数の罫線情報に基づいて電子文書ファイルの内容を解析する。具体的には、罫線によりレコード単位で項目がどのように区分けされているかを判断し、見出し部とデータ部とを区別する。
【００３９】
データ抽出規則特定部２０４は、解析部２０３での解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定する。特定されたデータベース項目及びデータ抽出規則は、データ抽出規則記憶部１３２に記憶される。
【００４０】
データ抽出規則特定部２０４は、ファイル指定受付部２０５、範囲指定受付部２０６、及び特定部２０７を備えても良い。解析部２０３の解析結果だけでは、データベース項目を特定することができない場合もありうるからである。このような場合、手動にてデータベース項目及びデータ抽出規則の特定を受け付ける。
【００４１】
データ抽出部２０８は、ファイル指定受付部２０５にて一又は複数の電子文書ファイルの指定を受け付け、範囲指定受付部２０６にて複数のシートが存在する場合にはシートの指定、及びシートに含まれる表データ中にて該表データと他の表データとのマージ対象となる範囲指定を受け付ける。特定部２０７は、受け付けた範囲指定に従って、データベース項目及びデータ抽出規則を特定する。
【００４２】
データ抽出部２０８は、特定したデータベース項目及びデータ抽出規則にて、一又は複数の電子文書ファイルからデータベース項目及び対応するデータを抽出する。抽出されたデータベース項目及び対応するデータはデータベース記憶部１３３に記憶される。
【００４３】
表示部２０９は、抽出されたデータベース項目及び対応するデータを表示装置２３にて一覧表示し、確定情報受付部２１０は、表示されたデータベース項目及び対応するデータが適正であるか否かをユーザが判断し、ユーザが適正であると判断した場合、すなわちデータベース項目に重複、抜け等が生じておらず、適正にマージされていると判断した場合には、ユーザによる適正であると判断した旨を示す確定情報の入力を受け付ける。
【００４４】
図３は、本発明の実施の形態１に係るデータベース生成装置１のＣＰＵ１１のデータベース生成処理の手順を示すフローチャートである。図３において、データベース生成装置１のＣＰＵ１１は、一又は複数の表データを含む電子文書ファイルを取得する（ステップＳ３０１）。電子文書ファイルは、記憶装置１３の電子文書ファイル記憶部１３１内に記憶されている電子文書ファイルを読み出しても良いし、ネットワーク２を介して外部のコンピュータから読み出しても良い。また、キーボード２１、マウス２２等の入力装置を介して入力を受け付けても良い。
【００４５】
ＣＰＵ１１は、取得した一又は複数の電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する（ステップＳ３０２）。具体的には、罫線で囲まれている部分を表データと認識し、その他の罫線がどのように配置されているかに関する情報を取得する。
【００４６】
ＣＰＵ１１は、抽出した複数の罫線情報に基づいて電子文書ファイルの内容を解析する（ステップＳ３０３）。具体的には、罫線によりレコード単位で項目がどのように区分けされているかを判断し、見出し部とデータ部とを区別する。
【００４７】
図４は、罫線情報に基づいて表データの抽出を行う処理の例示図である。具体的には、電子文書ファイルのデータの走査方向につきユーザの指定を受け付け、項目が階層化されているか否かを１行ずつ判定する。図４（ａ）では、表データを下方向４２へ走査する場合を示しており、項目領域４１の１行目には項目「材料名」、「重量」、「物質」、「比

率」が存在することを検出することができる。２行目では、項目「比率」が「平均重量」、「最大重量」に分割され、項目数が増加していることを検出することができる。

【００４８】

３行目では、項目名を検出することはできないものの、２行目の項目とセル位置及び項目数が同一であることを検出することができる。したがって、２行目まで見出し部であり、３行目以降がデータ部であることを自動認識することができ、新たなデータベース生成のためのデータ抽出は、３行目以降のデータ部から行うことができる。

【００４９】

図４（ｂ）では、表データを右方向４４へ走査する場合を示しており、項目領域４３の１列目には項目「材料名」、「重量」、「物質」、「比率」が存在することを検出することができる。２列目では、項目「比率」が「平均重量」、「最大重量」に分割され、項目数が増加していることを検出することができる。

【００５０】

３列目では、項目名を検出することはできないものの、２列目の項目とセル位置及び項目数が同一であることを検出することができる。したがって、２列目まで見出し部であり、３列目以降がデータ部であることを自動認識することができ、新たなデータベース生成のためのデータ抽出は、３列目以降のデータ部から行うことができる。

【００５１】

このように走査方向によらず、罫線情報に基づいて、データベース生成時に抽出するべきデータベース項目及び対応するデータのセル位置を正確に検出することができるので、表データの行と列とが反転している場合であっても一のデータベースとしてマージすることが可能となる。

【００５２】

図３に戻って、データベース生成装置１のＣＰＵ１１は、解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定する（ステップＳ３０４）。ＣＰＵ１１は、一又は複数の電子文書ファイル又は該電子文書ファイル中の表データから、新たに生成するデータベースのデータベース項目及び対応するデータを抽出する（ステップＳ３０５）。データを抽出する規則は、データ抽出規則記憶部１３２に記憶されているデータ抽出規則に従う。

【００５３】

なお、罫線情報の解析結果だけでは正しくデータベース項目等が特定できない場合も生じうる。この場合、手動にてデータベース項目及びデータ抽出規則の特定を受け付ける。図５は、本発明の実施の形態１に係るデータベース生成装置１のＣＰＵ１１の手動特定処理の手順を示すフローチャートである。

【００５４】

図５において、データベース生成装置１のＣＰＵ１１は、図３のステップＳ３０３の処理の実行終了後、一又は複数の電子文書ファイルの指定を受け付け（ステップＳ５０１）、複数のシートが存在する場合にはシートの指定、及びシートに含まれる表データ中にて該表データと他の表データとのマージ対象となる範囲指定を受け付ける（ステップＳ５０２）。ＣＰＵ１１は、受け付けた範囲指定に従って、データベース項目及びデータ抽出規則を特定し（ステップＳ５０３）、処理を図３のステップＳ３０５へ進める。

【００５５】

図６は、範囲指定が必要となる場合の例示図である。図６（ａ）は、表データの構造が特段の規則性を有していない場合の例示図である。この場合、キーボード２１、マウス２２等の入力装置により、表データとして使用する領域６１のみを範囲指定として受け付ける。指定を受け付けた範囲に、例えば他の表データのデータベース項目とリンクするようなタグ情報を付加することにより、新たなデータベースのデータとして用いることができる。

【００５６】

図６（ｂ）は、表データとして認識できない領域区分となっている場合の例示図である

。図６（ｂ）の例では、見出し部として認識するべき領域６２が表として認識できる領域、すなわち矩形領域となっていない。この場合、キーボード２１、マウス２２等の入力装置により、領域６２を含めて列ごとの領域６３の範囲指定を受け付け、見出し部「材料」、「重量」、「比率１」、「比率２」に対して、他の表データのデータベース項目とリンクするようにタグ情報を付加する。これにより、新たなデータベースのデータとして用いることができる。

【００５７】

図３に戻って、データベース生成装置１のＣＰＵ１１は、抽出されたデータベース項目及び対応するデータを表示装置２３にて一覧表示し（ステップＳ３０６）、表示されたデータベース項目及び対応するデータが適正である旨の確認情報を受け付けたか否かを判断する（ステップＳ３０７）。ＣＰＵ１１が、確認情報を受け付けていないと判断した場合（ステップＳ３０７：ＮＯ）、再度新たなデータベースを生成するべく処理を終了し、確認情報を受け付けたと判断した場合（ステップＳ３０７：ＹＥＳ）、ユーザがデータベース項目に重複、抜け等が生じておらず、適正にマージされていると判断したので、生成された新たなデータベースを検索処理の対象等として記憶装置１３に記憶する（ステップＳ３０８）。

【００５８】

このように、罫線情報に基づいて抽出対象となる表データの存在位置を特定することができ、表データからデータベース項目及び対応するデータを抽出するためのデータ抽出規則を正しく特定することができるので、新たなデータベースを適正に生成することが可能となる。

【００５９】

また、マージ対象となる表データ間に、いわゆるゆらぎ情報が存在する場合がある。ここで、「ゆらぎ情報」とは、表データ間の相違に関する情報の総称である。例えば表データの開始セルの位置の相違に関する表位置ゆらぎ情報、表データの項目の順序が相違する、新規項目の存在、項目の抜けの存在等の項目の相違に関する項目ゆらぎ情報等がある。

【００６０】

図７は、表データの位置に相違が存在する「表位置ゆらぎ情報」の説明図である。図７（ａ）から図７（ｃ）に示すように、表データを示す罫線が存在する領域の左上のセル７１、７２、７３のサイズが相違することにより、電子文書ファイル中の表データのセル位置がそれぞれ相違している。表位置ゆらぎ情報が存在する場合、例えばデータ抽出規則を「上から１番目の表」等に特定しておく、又は記憶してあるデータ抽出規則から選択することにより、図７（ａ）から図７（ｃ）に示すすべての表データを抽出の対象とすることができる。つまり、表データの開始位置（開始セル位置）をあらかじめ定めるのではなく、罫線情報に基づいて表データを抽出し、一の電子文書ファイル中又は一の電子文書ファイルに含まれる一のシート中に複数の表データが抽出された場合には、「上から１番目の表」、「最大の表」、「特定の項目を含む表」等の直接的な位置ではない表の特徴をデータ抽出規則として設定して記憶しておくことにより、表位置のゆらぎに対応して表データを抽出することができる。

【００６１】

図８は、表データの項目に相違が存在する「項目ゆらぎ情報」の説明図である。図８（ａ）を基準とした場合、図８（ｂ）は項目Ｃと項目Ｂとの順序が入れ替わっている。従来のＣＳＶファイルを用いて表データをマージする場合には、項目Ｃと項目Ｂとの順序が入れ替わっていることを事前にユーザが知っている状態で、入れ替え指示を出す必要があった。

【００６２】

本実施の形態１では、項目が入れ替わっていることを検出して、項目名にリンクしたタグ情報を付与する。すなわち図８（ａ）の見出し部８１では、例えば項目Ａに対してタグ情報‘ａ’を、項目Ｂに対してタグ情報‘ｂ’を、項目Ｃに対してタグ情報‘ｃ’を、それぞれ付与する。図８（ｂ）の見出し部８２では、項目Ｂと項目Ｃとの順序が入れ替わっ

ているが、タグ情報は図８（ａ）と同様の対応関係で付与しておく。データ抽出時にはタグ情報‘ａ’、タグ情報‘ｂ’及びタグ情報‘ｃ’を基礎としてデータを集約するので、基礎となる表データで項目がどのように配置されていても、新たなデータベースでは、タグ情報の順に集約することができる。したがって、項目Ｃと項目Ｂとの順序が入れ替わっていることを事前にユーザが知らなくても、表データを適正にマージすることが可能となる。

【００６３】

また、図８（ｃ）では、見出し部８３に新たな項目である項目Ｄ、項目Ｅが存在するのに対し、項目Ｃが欠落している。この場合も、項目Ｄに対してタグ情報‘ｄ’を、項目Ｅに対してタグ情報‘ｅ’を、それぞれ付与することにより、データ抽出時にタグ情報を基礎としてデータを抽出する限り、誤った項目を集約するおそれはない。すなわち、新規に追加された項目は独立して集約することができるし、欠落している項目については本表データからはデータを抽出することがない。

【００６４】

以上のように、表データ間にゆらぎ情報が存在する場合であっても、タグ情報を基礎として同一項目についてはデータを集約することができ、ユーザが項目の入れ替わりに関する情報等を正確に把握することなく、新たなデータベースを生成することが可能となる。つまり、抽出するべき項目、項目の順序等をあらかじめ定めるのではなく、レコードを抽出する場合に項目名も抽出し、項目をタグ情報として例えばＸＭＬデータベース形式で記憶し、タグ情報（項目名）によって表を再構成することで、項目のゆらぎに対応して表データを抽出することができる。また、何らかの原因により本来は同じ項目として取り扱うべき項目を異なる項目として認識した場合であっても、表を再構成するときに、同じ項目として取り扱うべき項目を同一化する指示を受け付けることが可能な構成とすることにより、事後的に項目ゆらぎを修正することも可能となる。

【００６５】

図９は、本発明の実施の形態１に係るデータベース生成装置１のＣＰＵ１１のゆらぎ補正処理の手順を示すフローチャートである。

【００６６】

図９において、データベース生成装置１のＣＰＵ１１は、図３のステップＳ３０４の処理実行後、異なる電子文書ファイルから抽出した表データの位置の相違に関する情報、及び／又は異なる電子文書ファイルから抽出したデータベース項目の相違に関する情報を少なくとも含む表データの相違に関するゆらぎ情報を抽出する（ステップＳ９０１）。抽出するゆらぎ情報は、上述した２つに限定されるものではない。

【００６７】

ＣＰＵ１１は、データベース項目、データ抽出規則及び抽出されたゆらぎ情報に基づいて、データ抽出規則の変更部分が存在するか否かを判断する（ステップＳ９０２）。ＣＰＵ１１が、変更部分が存在しないと判断した場合（ステップＳ９０２：ＮＯ）、ＣＰＵ１１は、ゆらぎ情報に起因するデータ抽出規則に対する何らかの補正処理（以下、ゆらぎ補正）が実行されていないと判断して、処理を図３のステップＳ３０７へ進める。ＣＰＵ１１が、変更部分が存在すると判断した場合（ステップＳ９０２：ＹＥＳ）、ＣＰＵ１１は、同一のデータベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与する（ステップＳ９０３）。

【００６８】

ＣＰＵ１１は、データ抽出規則の変更部分を反映してデータベース項目及び対応するデータを抽出し（ステップＳ９０４）、データベース項目に付与されているタグ情報に従ってデータベース項目を配列して、データベース項目及び対応するデータを一覧表示し（ステップＳ９０５）、処理を図３のステップＳ３０７へ進める。

【００６９】

このようにすることで、マージ対象となる表データ間にゆらぎ情報が存在する場合であっても、ユーザは事前にその存在を把握しておく必要がなく、適正なゆらぎ補正を実行し

た状態で新たなデータベース項目及び対応するデータを視認することができる。したがって、ユーザに過剰な負荷がかかることなく、容易に新たなデータベースを生成することが可能となる。

【００７０】

（実施の形態２）

本発明の実施の形態２に係るデータベース生成装置の構成は、実施の形態１と同様であることから、同一の符号を付することにより、詳細な説明を省略する。本実施の形態２では、異なるデータ形式を有する複数の表データが存在する場合であっても、容易に新たなデータベースを生成することができ、ゆらぎ補正の内容を視認することができる点で実施の形態１と相違する。

【００７１】

図１０は、本発明の実施の形態２に係るデータベース生成装置１の機能ブロック図である。図１０では、実施の形態１と同様の機能ブロックについては同一の符号を付している。電子文書ファイル取得部２０１は、一又は複数の表データを含む電子文書ファイルを取得する。電子文書ファイルは、記憶装置１３内に記憶されている電子文書ファイルを電子文書ファイル記憶部１３１に集約しても良いし、ネットワーク２を介して外部のコンピュータから取得しても良い。また、キーボード２１、マウス２２等の入力装置を介して入力しても良い。

【００７２】

罫線情報抽出部２０２は、取得した一又は複数の電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する。具体的には、罫線で囲まれている部分を表データと認識し、その他の罫線がどのように配置されているかに関する情報を取得する。

【００７３】

解析部２０３は、抽出した複数の罫線情報に基づいて電子文書ファイルの内容を解析する。具体的には、罫線によりレコード単位で項目がどのように区分けされているかを判断し、見出し部とデータ部とを区別する。

【００７４】

データ抽出規則特定部２０４は、解析部２０３での解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定する。特定されたデータベース項目及びデータ抽出規則は、データ抽出規則記憶部１３２に記憶される。

【００７５】

データ抽出規則特定部２０４は、ファイル指定受付部２０５、範囲指定受付部２０６、及び特定部２０７を備えても良い。解析部２０３の解析結果だけでは、データベース項目を特定することができない場合もありうるからである。このような場合、手動にてデータベース項目及びデータ抽出規則の特定を受け付ける。

【００７６】

ゆらぎ情報抽出部１００１は、解析部２０３での解析結果に基づいて、いわゆるゆらぎ情報が存在する場合には、存在するゆらぎ情報を抽出する。例えば、抽出対象となる表データ間において項目ゆらぎ情報が存在する場合、項目の順序の相違、項目の相違等に関する情報を抽出する。

【００７７】

判断部１００２は、データベース項目、データ抽出規則及び抽出されたゆらぎ情報に基づいて、データ抽出規則の変更部分が存在するか否かを判断する。すなわち、ゆらぎ情報の存在によって、タグ情報に基づく表データの抽出規則が変更されるので、変更部分が存在すると判断した場合にはデータ抽出規則に対して何らかのゆらぎ補正が実行されていると判断することができる。

【００７８】

タグ情報付与部１００３は、判断部１００２で変更部分が存在すると判断した場合、同一のデータベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与する。すなわち項目の配列、順序、種類等が異なる表データであっても、タグ

情報が同一である場合には同一の項目であることを担保する。これにより、タグ情報に基づいてデータベースを構成することにより、ゆらぎ情報が存在する場合であっても適切に新たなデータベースを生成することができる。

【００７９】

データ抽出部１００４は、特定したデータベース項目及びデータ抽出規則にて、一又は複数の電子文書ファイルからデータベース項目及び対応するデータを抽出する。データ抽出規則にはタグ情報に関する項目が含まれ、タグ情報に応じてデータベース項目及び対応するデータが抽出され、データベース記憶部１３３に記憶される。

【００８０】

表示部２０９は、抽出されたデータベース項目及び対応するデータを表示装置２３にて一覧表示する。一方、ゆらぎ情報表示部１００５は、データ抽出の対象となる電子文書ファイルに関するゆらぎ情報の種類に関する情報、及びゆらぎ情報が存在するシートを特定する情報を表示する。

【００８１】

図１１は、本発明の実施の形態２に係るデータベース生成装置１のＣＰＵ１１のデータベース生成処理の手順を示すフローチャートである。図１１において、データベース生成装置１のＣＰＵ１１は、一又は複数の表データを含む電子文書ファイルを取得する（ステップＳ１１０１）。電子文書ファイルは、記憶装置１３内の電子文書ファイル記憶部１３１に記憶されている電子文書ファイルを読み出しても良いし、ネットワーク２を介して外部のコンピュータから読み出しても良い。また、キーボード２１、マウス２２等の入力装置を介して入力を受け付けても良い。

【００８２】

ＣＰＵ１１は、取得した一又は複数の電子文書ファイルに含まれる罫線に関する罫線情報をそれぞれ抽出する（ステップＳ１１０２）。具体的には、罫線で囲まれている部分を表データと認識し、その他の罫線がどのように配置されているかに関する情報を取得する。

【００８３】

ＣＰＵ１１は、抽出した複数の罫線情報に基づいて電子文書ファイルの内容を解析する（ステップＳ１１０３）。具体的には、罫線によりレコード単位で項目がどのように区分けされているかを判断し、見出し部とデータ部とを区別する。区別する方法は実施の形態１と同様であることから、詳細な説明は省略する。

【００８４】

ＣＰＵ１１は、解析結果に基づいて、生成するデータベースのデータベース項目及びデータ抽出規則を特定する（ステップＳ１１０４）。ＣＰＵ１１は、解析結果に基づいて、いわゆるゆらぎ情報が存在する場合には、存在するゆらぎ情報を抽出する（ステップＳ１１０５）。すなわち、抽出対象となる表データ間において、項目の順序の相違、項目の相違等に関する情報を抽出する。

【００８５】

ＣＰＵ１１は、データベース項目、データ抽出規則及び抽出されたゆらぎ情報に基づいて、データ抽出規則の変更部分が存在するか否かを判断する（ステップＳ１１０６）。すなわち、ゆらぎ情報の存在によって、タグ情報に基づく表データの抽出規則が変更されるので、変更部分が存在すると判断した場合にはデータ抽出規則に対して何らかのゆらぎ補正が実行されていると判断することができる。

【００８６】

ＣＰＵ１１が、データ抽出規則の変更部分が存在すると判断した場合（ステップＳ１１０６：ＹＥＳ）、ＣＰＵ１１は、同一のデータベース項目に対しては同一の、異なるデータベース項目に対しては異なるタグ情報を付与する（ステップＳ１１０７）。すなわち項目の配列、順序、種類等が異なる表データであっても、タグ情報が同一である場合には同一の項目であることを担保する。これにより、タグ情報に基づいてデータベースを構成することで、ゆらぎ情報が存在する場合であっても適切に新たなデータベースを生成するこ

とができる。

【００８７】

ＣＰＵ１１が、データ抽出規則の変更部分が存在しないと判断した場合（ステップＳ１１０６：ＮＯ）、ＣＰＵ１１は、ステップＳ１１０７をスキップし、特定したデータベース項目及びデータ抽出規則にて、一又は複数の電子文書ファイルからデータベース項目及び対応するデータを抽出する（ステップＳ１１０８）。データ抽出規則にはタグ情報に関する項目が含まれ、タグ情報に応じてデータベース項目及び対応するデータが抽出され、データベース記憶部１３３に記憶される。

【００８８】

ＣＰＵ１１は、抽出されたデータベース項目及び対応するデータを表示装置２３にて一覧表示するとともに（ステップＳ１１０９）、データ抽出の対象となる電子文書ファイルに関するゆらぎ情報の種類に関する情報、及びゆらぎ情報が存在する位置に関する情報、例えばゆらぎ情報が存在する電子文書ファイル名、シート名、セル位置等を表示する（ステップＳ１１１０）。

【００８９】

このように、ゆらぎ情報が存在する場合であっても、ゆらぎ情報に応じてデータ抽出規則を変更する、すなわちタグ情報を基礎としたデータ抽出規則を変更することにより、正しくデータベース項目及び対応するデータを集約することが可能となる。また、ゆらぎ情報が生じた電子文書ファイル、シート、セル位置等を視認することができ、ゆらぎ補正をどのように行うかを判断することができる。

【００９０】

図１２は、抽出されたゆらぎ情報の一覧を表示する表示画面の例示図である。図１２の例では、「表位置ゆらぎ」、「項目ゆらぎ」、「文字ゆらぎ」、「エラー」の各項目について、その総数、電子文書ファイル名、シート名等の一覧を、それぞれ領域１２１、１２２、１２３、１２４に表示している。

【００９１】

また、ゆらぎ情報の一覧表示画面から、ゆらぎ補正がどのように実行されたか確認できることが好ましい。図１３は、ゆらぎ補正の内容を確認するための表示画面の例示図である。図１３（ａ）は、図１２と同様、抽出されたゆらぎ情報の一覧を表示する表示画面の例示図であり、図１３（ｂ）は、ゆらぎ情報の内容を示す表示画面の例示図であり、図１３（ｃ）は、ゆらぎ情報抽出の対象となった表データを並列に表示する表示画面の例示図である。

【００９２】

図１３（ａ）の表示画面において、ゆらぎ補正がどのように行われたのか知りたい項目１３５について、マウス等の入力装置による選択を受け付ける。項目１３５の選択を受け付けた場合、図１３（ｂ）に示すように、対応する表データのシート自体を表示する表データ表示領域１３６とゆらぎ情報の内容を表示するゆらぎ情報表示領域１３８を含む表示画面が表示される。表データ表示領域１３６には、ゆらぎ情報の対象となった表データのシートが表示され、ゆらぎ情報表示領域１３８には、表位置ゆらぎ情報に対するゆらぎ補正の内容が表示されている。

【００９３】

図１３の例では、表データが開始するセル位置にゆらぎが生じており、図１３（ｂ）の表示画面のようにデータ抽出規則における開始するセル位置１３７を変更した旨を示す情報が表示されている。すなわち、データ抽出規則では、セル位置「Ｃ：３」からデータを抽出するよう指示されていたのに対し、ゆらぎ情報により、データ抽出規則を、セル位置「Ｄ：５」からデータを抽出するよう変更した旨を示している。

【００９４】

また、シート自体を比較することが可能なように並列して表示しても良い。この場合、図１３（ｃ）に示すように、表データ表示領域１３６の横に並列して、ゆらぎ元表データ表示領域１３９を表示しておき、ゆらぎの対象となるセル位置１３７、１４０をそれぞれ

強調表示することにより、どのようにゆらぎ補正が実行されているのか、明確に視認することができる。

【００９５】

また、ゆらぎ情報の一覧表示画面において、エラー表示がなされている表データは、事前に想定しているゆらぎ補正では対応できなかった表データの存在を示している。図１４は、エラー表示の内容を示す表示画面の例示図である。図１４（ａ）は、図１２と同様、抽出されたゆらぎ情報の一覧を表示する表示画面の例示図であり、図１４（ｂ）は、エラー情報の内容を示す表示画面の例示図であり、図１４（ｃ）は、エラー情報抽出の対象となった表データを並列に表示する表示画面の例示図である。

【００９６】

図１４（ａ）の表示画面において、エラー表示がなされている項目１４１について、マウス等の入力装置による選択を受け付ける。項目１４１の選択を受け付けた場合、図１４（ｂ）に示すように、対応する表データのシート自体を表示する表データ表示領域１４２とエラー情報の内容を表示するエラー情報表示領域１４４を含む表示画面が表示される。表データ表示領域１４２には、表データのシート自体が表示され、エラー情報表示領域１４４には、エラーが生じた原因を示す原因情報、例えば表データ内の項目がすべて一致していない旨を示す情報が表示されている。

【００９７】

図１４の例では、表データの項目が、データ抽出規則に規定されている新たなデータベースのデータベース項目とすべて一致していないことから、ゆらぎ情報の補正処理を行うことができない。したがって、項目をデータベース項目にあわせるよう補正するには、シート自体を比較することが可能なように並列して表示することが好ましい。

【００９８】

図１４（ｃ）では、表データ表示領域１４２の横に並列して、ゆらぎ情報を抽出できなかった元表データ表示領域１４５を表示している。両者の項目領域１４３と項目領域１４６とを比較できるよう、例えば強調表示することにより、項目名がどのように相違しているのかを、明確に視認することができる。

【００９９】

なお、原因情報としては、少なくともエラー情報が生じている電子文書ファイルを特定する情報、エラー情報の内容に関する情報、及びエラー情報が生じているシートを特定する情報を含むことが好ましい。また、原因情報は、データ抽出規則の変更部分に関する情報を含むことも望ましい。これは、ゆらぎ情報を表示する場合にゆらぎ情報が生じている電子文書ファイルを特定する情報、ゆらぎ情報に関する情報、ゆらぎ情報が生じているシートを特定する情報を含むことに準ずる。

【０１００】

エラー情報が生じている電子文書ファイルを特定する情報、エラー情報の内容に関する情報、及びエラー情報が生じているシートを特定する情報を含むことにより、ゆらぎ情報を補正しきれなかった電子文書ファイル及びシートを特定することができ、補正ができなかった原因を視認することが可能となる。また、データ抽出規則の相違を視認することができ、データ抽出規則を変更することにより新たなデータベース生成に用いることができるか否かを判断することができる。

【０１０１】

さらに、原因情報を解析して、エラー情報表示領域１４４に、エラー情報が生じている表データを抽出することが可能な他のデータ抽出規則を検索して表示することが好ましい。データ抽出規則を変更することが可能である場合には、エラーを解消することができる可能性もあり、エラー情報が生じている表データであってもエラーを解消することで新たなデータベースに取り込むことができるからである。

【０１０２】

図１５は、他のデータ抽出規則を示唆する旨の情報を表示する表示画面の例示図である。図１５の例では、表データ表示領域１５２の横に並列して、ゆらぎ情報を抽出できなか

った元表データ表示領域１５５を表示している。両方の表データ領域１５３と表データ領域１５６とを比較し、いずれの表も表のサイズが最も大きい表であることを検出した場合、エラー情報表示領域１５４には、いずれの表も１番大きな表である旨のメッセージを表示する。斯かる情報が表示されることにより、データ抽出規則として「最も大きいサイズの表データ」を採用した場合には、表データ領域１５６に表示されている表データについても新たなデータベースに取り込むことが可能となる。

【０１０３】

図１６は、原因情報を表示する場合の本発明の実施の形態２に係るデータベース生成装置１のＣＰＵ１１の処理手順を示すフローチャートである。図１６において、データベース生成装置１のＣＰＵ１１は、図１１のステップＳ１１１０の処理の後、エラー情報が生じた電子文書ファイルが存在するか否かを判断する（ステップＳ１６０１）。

【０１０４】

ＣＰＵ１１が、エラー情報が生じた電子文書ファイルが存在しないと判断した場合（ステップＳ１６０１：ＮＯ）、ＣＰＵ１１は、処理を終了する。ＣＰＵ１１が、エラー情報が生じた電子文書ファイルが存在すると判断した場合（ステップＳ１６０１：ＹＥＳ）、ＣＰＵ１１は、エラー情報が生じた原因に関する原因情報を表示装置２３に表示し（ステップＳ１６０２）、他のデータ抽出規則が記憶装置１３のデータ抽出規則記憶部１３２に記憶されているか否かを判断する（ステップＳ１６０３）。

【０１０５】

ＣＰＵ１１が、他のデータ抽出規則が記憶されていないと判断した場合（ステップＳ１６０３：ＮＯ）、ＣＰＵ１１は、処理を終了する。ＣＰＵ１１が、他のデータ抽出規則が記憶されていると判断した場合（ステップＳ１６０３：ＹＥＳ）、ＣＰＵ１１は、記憶されている他のデータ抽出規則から一のデータ抽出規則を選択し（ステップＳ１６０４）、エラー情報が生じた表データから、データベース項目及び対応するデータを抽出する（ステップＳ１６０５）。

【０１０６】

ＣＰＵ１１は、再度エラー情報が生じたか否かを判断し（ステップＳ１６０６）、ＣＰＵ１１が、エラー情報が生じたと判断した場合（ステップＳ１６０６：ＹＥＳ）、ＣＰＵ１１は、記憶されているすべての他のデータ抽出規則を選択したか否かを判断する（ステップＳ１６０７）。ＣＰＵ１１が、すべてのデータ抽出規則を選択したと判断した場合（ステップＳ１６０７：ＹＥＳ）、ＣＰＵ１１は、処理を終了する。

【０１０７】

ＣＰＵ１１が、まだ選択されていない他のデータ抽出規則が存在すると判断した場合（ステップＳ１６０７：ＮＯ）、ＣＰＵ１１は、次の他のデータ抽出規則を選択し（ステップＳ１６０８）、処理をステップＳ１６０５に戻して、上述した処理を繰り返す。ＣＰＵ１１が、エラー情報が生じていないと判断した場合（ステップＳ１６０６：ＮＯ）、ＣＰＵ１１は、選択した他のデータ抽出規則に関するメッセージを表示装置２３に表示する（ステップＳ１６０９）。

【０１０８】

このようにすることで、複数の表データ間にゆらぎ情報が存在する場合であっても、ゆらぎ情報を補正して新たなデータベースを生成することができ、どのようにゆらぎ補正を実行したか、あるいはゆらぎ補正を実行することができなかった原因に関する原因情報を表示することにより、元の表データ又はデータ抽出規則を変更するための情報を得ることが可能となる。また、エラー情報が生じた電子文書ファイルを特定することができ、エラー情報が生じた原因を視認することができるとともに、データ抽出規則を変更することによりエラー情報が生じることなく新たなデータベース生成を行うことができるか否かを判断することが可能となる。

【０１０９】

なお、本発明は上記実施例に限定されるものではなく、本発明の趣旨の範囲内であれば多種の変更、改良等が可能である。例えば変更すべきデータ抽出規則が見つかった場合、

自動的にデータ抽出規則を変更しても良いし、ユーザによる変更指示の入力を受け付けても良い。また、データ抽出規則が変更された場合、自動的に再度データベース生成処理を実行するようにしても良いし、ユーザによるデータベース再生成指示の入力を受け付けても良い。

【図面の簡単な説明】

【０１１０】

【図１】本発明の実施の形態１に係るデータベース生成装置の構成例を示すブロック図である。

【図２】本発明の実施の形態１に係るデータベース生成装置の機能ブロック図である。

【図３】本発明の実施の形態１に係るデータベース生成装置のＣＰＵのデータベース生成処理の手順を示すフローチャートである。

【図４】罫線情報に基づいて表データの抽出を行う処理の例示図である。

【図５】本発明の実施の形態１に係るデータベース生成装置のＣＰＵの手動特定処理の手順を示すフローチャートである。

【図６】範囲指定が必要となる場合の例示図である。

【図７】表データの位置に相違が存在する「表位置ゆらぎ情報」の説明図である。

【図８】表データの項目に相違が存在する「項目ゆらぎ情報」の説明図である。

【図９】本発明の実施の形態１に係るデータベース生成装置のＣＰＵのゆらぎ補正処理の手順を示すフローチャートである。

【図１０】本発明の実施の形態２に係るデータベース生成装置の機能ブロック図である。

【図１１】本発明の実施の形態２に係るデータベース生成装置のＣＰＵのデータベース生成処理の手順を示すフローチャートである。

【図１２】抽出されたゆらぎ情報の一覧を表示する表示画面の例示図である。

【図１３】ゆらぎ補正の内容を確認するための表示画面の例示図である。

【図１４】エラー表示の内容を示す表示画面の例示図である。

【図１５】他のデータ抽出規則を示唆する旨の情報を表示する表示画面の例示図である。

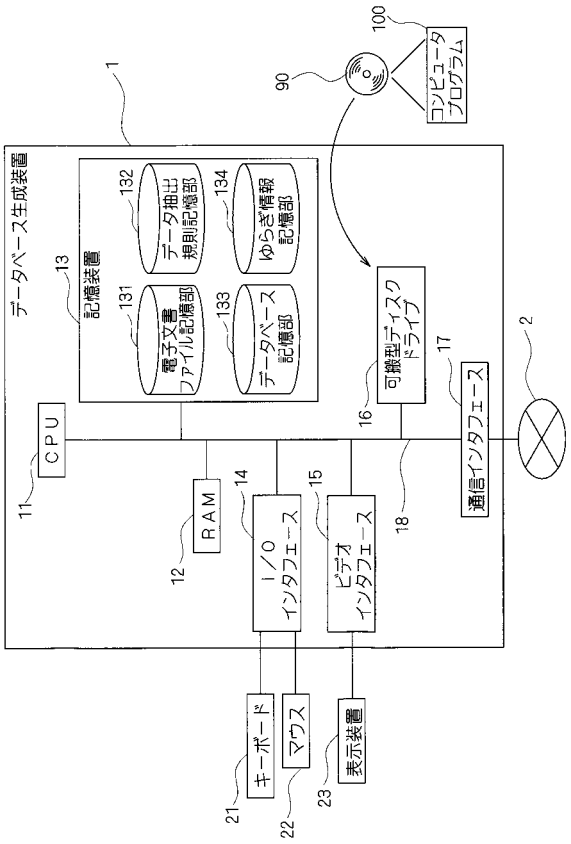
【図１６】原因情報を表示する場合の本発明の実施の形態２に係るデータベース生成装置のＣＰＵの処理手順を示すフローチャートである。

【符号の説明】

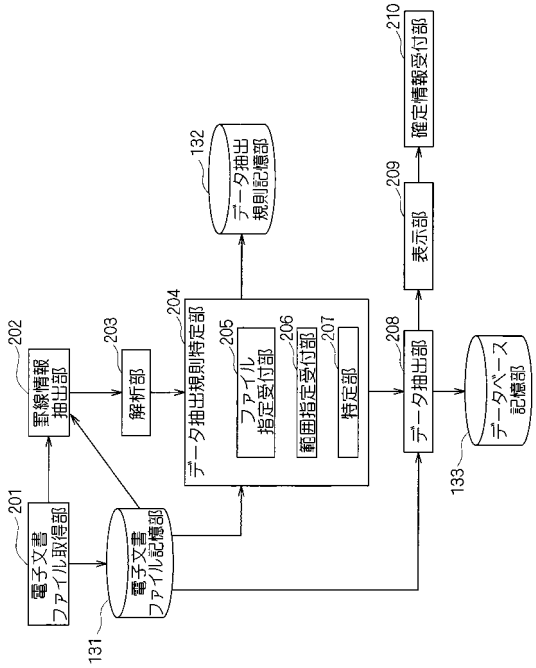
【０１１１】

１　データベース生成装置
２　ネットワーク
１１　ＣＰＵ
１２　ＲＡＭ
１３　記憶装置
１４　Ｉ／Ｏインタフェース
１５　ビデオインタフェース
１６　可搬型ディスクドライブ
１７　通信インタフェース
１８　内部バス
２３　表示装置
９０　可搬型記録媒体
１００　コンピュータプログラム
１３１　電子文書ファイル記憶部
１３２　データ抽出規則記憶部
１３３　データベース記憶部
１３４　ゆらぎ情報記憶部

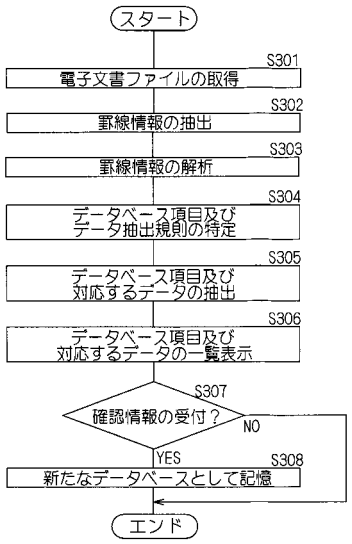
【図１】


データベース生成装置
1
記憶装置
13
データ抽出規則記憶部
132
ゆらぎ情報記憶部
134
電子文書ファイル記憶部
131
データベース記憶部
133
CPU
11
RAM
12
I／Oインタフェース
14
ビデオインタフェース
15
可搬型ディスクドライブ
16
通信インタフェース
17
18
2
90
100
コンピュータプログラム
キーボード
21
マウス
22
表示装置
23

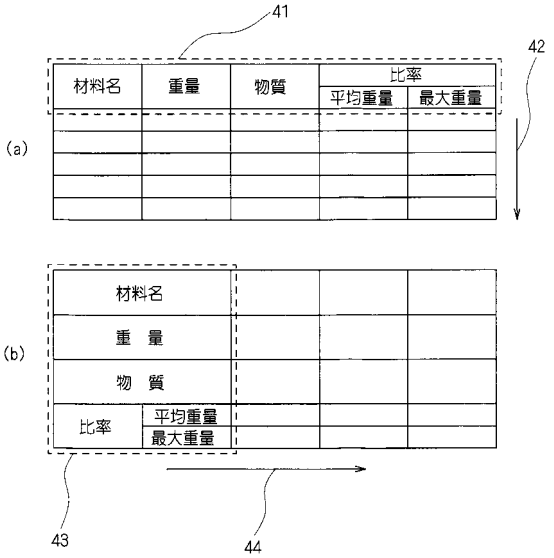
【図２】


罫線情報抽出部
202
解析部
203
データ抽出規則特定部
204
ファイル指定受付部
205
範囲指定受付部
206
特定部
207
データ抽出部
208
表示部
209
確定情報受付部
210
データ抽出規則記憶部
132
データベース記憶部
133
電子文書ファイル取得部
201
電子文書ファイル記憶部
131

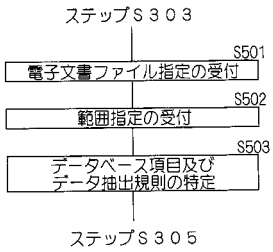
【図３】


スタート
S301
電子文書ファイルの取得
S302
罫線情報の抽出
S303
罫線情報の解析
S304
データベース項目及びデータ抽出規則の特定
S305
データベース項目及び対応するデータの抽出
S306
データベース項目及び対応するデータの一覧表示
S307
確認情報の受付？
NO
YES
S308
新たなデータベースとして記憶
エンド

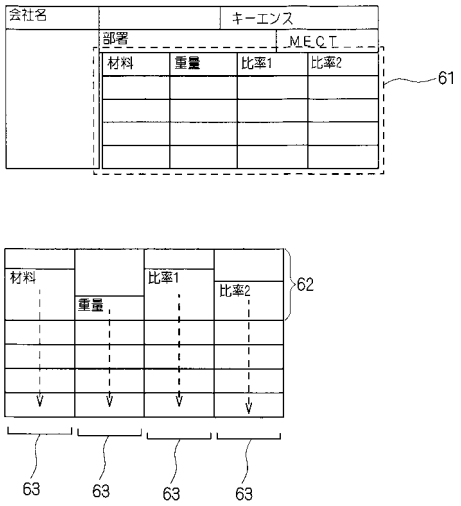
【図４】


41
42
(a)
材料名
重量
物質
比率
平均重量
最大重量
(b)
材料名
重　量
物　質
比率
平均重量
最大重量
43
44

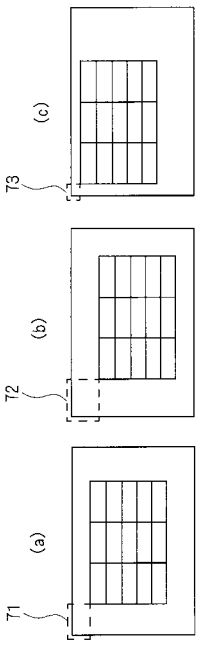
【図５】


ステップＳ３０３
S501
電子文書ファイル指定の受付
S502
範囲指定の受付
S503
データベース項目及び
データ抽出規則の特定
ステップＳ３０５

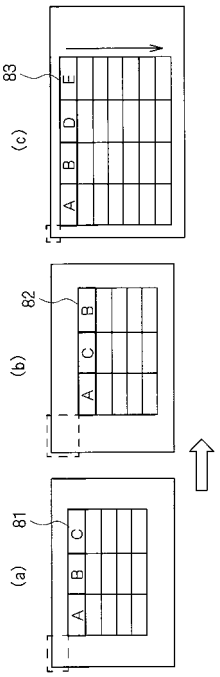
【図６】


会社名
キーエンス
部署
M.E.C.T.
材料
重量
比率1
比率2
61
材料
重量
比率1
比率2
62
63
63
63
63

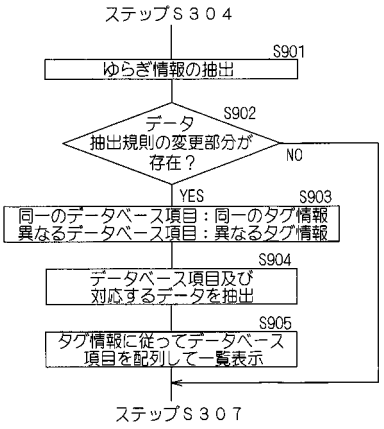
【図７】


(c)
73
(b)
72
(a)
71

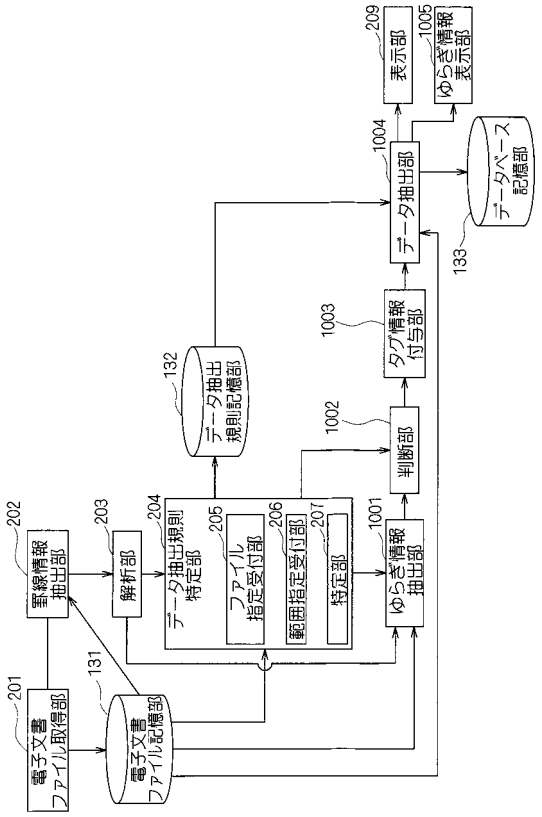
【図８】


83
(c)
E
D
B
A
82
(b)
B
C
A
81
(a)
C
B
A

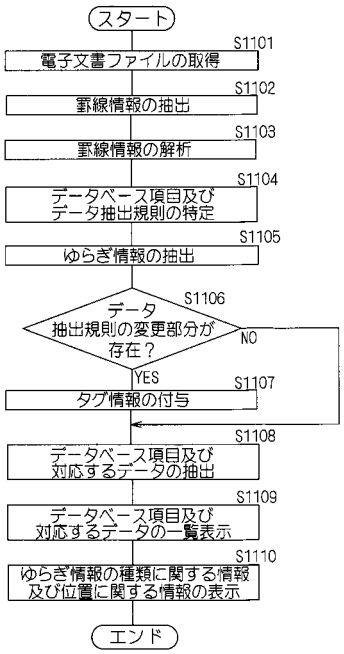
【図９】



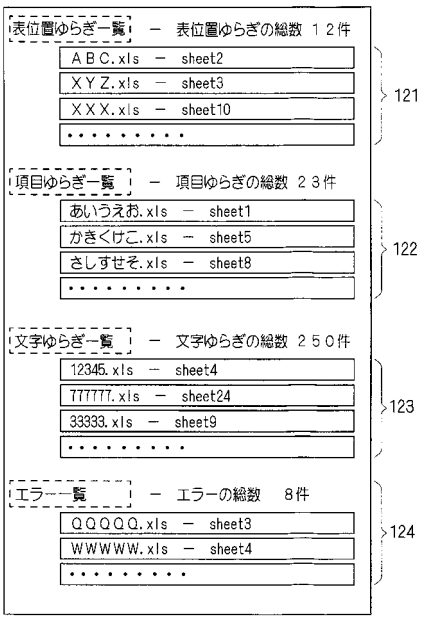
【図１０】



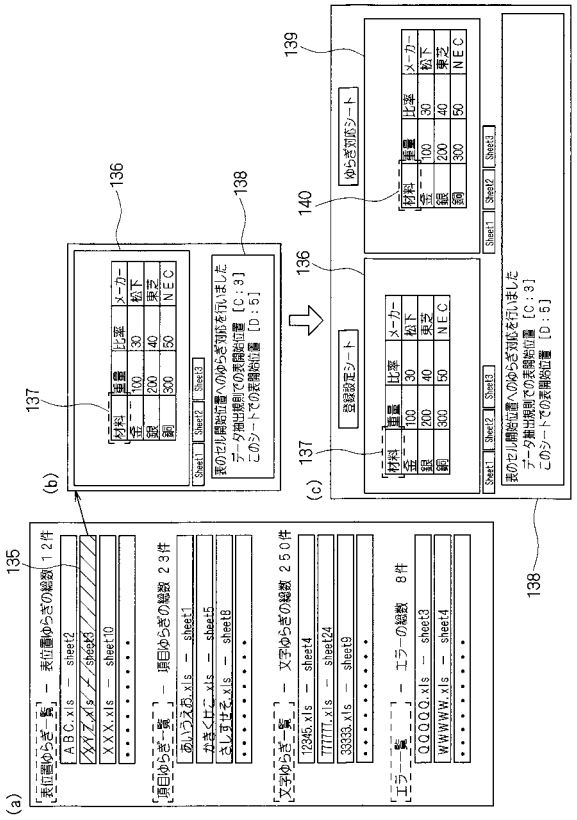
【図１１】



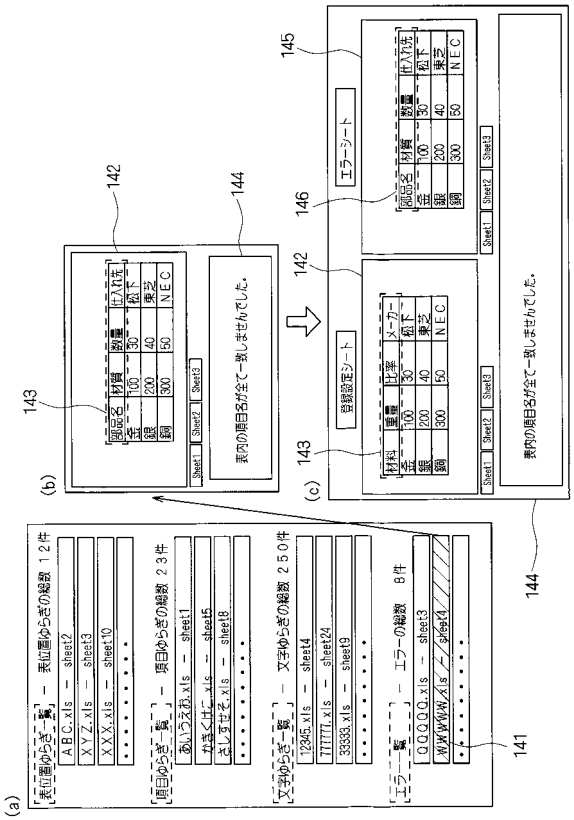
【図１２】



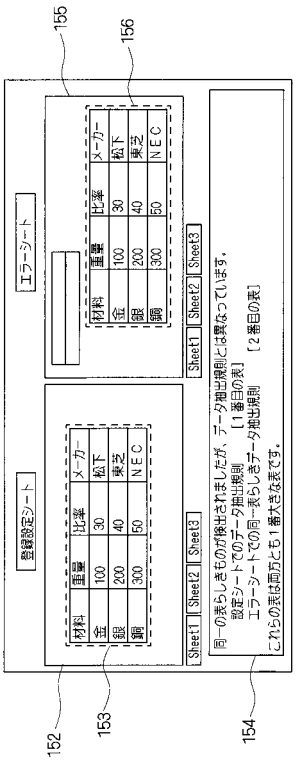
【図１３】



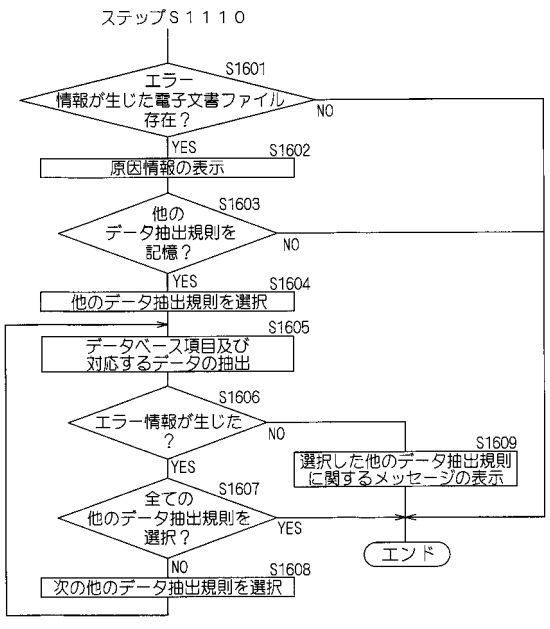
【図１４】



【図１５】



【図１６】



フロントページの続き

(56)参考文献 特開２００４－２５２５０９（ＪＰ，Ａ）
特開２００１－３３１７６４（ＪＰ，Ａ）
特開平０９－０５０５２７（ＪＰ，Ａ）
特開２００５－２４２５８７（ＪＰ，Ａ）
特開平０６－２４３１３０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｑ　１０／００－５０／３４
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　１７／３０