データクレンジングの重要性（IRSME13027）

平成 26 年 2 月 20 日　原田長州

企業内のデータは、さまざまな状況で取り扱われる。確実に規則を守っているつもりでもデータの重複・不整合・表記のゆれなどが発生する。これらの不整合は想定外のトラブルやクレームの発生につながり、場合によっては会社の屋台骨を揺さぶりかねない事態に発展する。定期的な見直しをすることで不整合を取り除いていきたい。

以下に代表的な不整合の例を取り上げチェックしていくが、この稿では Office ソフトで管理されているような規模の小さいデータを対象にする。企業内で基幹システム構築はされていても、付随業務などでシステム構築するほどのデータ量がない場合は、エクセルやアクセスで管理されているケースが多く、不整合の多くはそこを起因としているからだ。

■　データの重複

顧客リストで同一人物が複数登録されているなどのデータの重複があると、全顧客に一斉に案内する場合に同じ顧客に同じ案内をしてしまう問題が生じる。そればかりか、本来は一人の顧客に対するものとして残しておくべき取引の記録や売掛金の状況が分散してしまい、顧客との連絡に大きな問題が発生する場合も少なくない。

この重複を見つけ出すには、顧客リストなどの場合、名前、住所、電話番号などをキーに重複していないかをチェックする例が多い。注意したいのは電話番号で、登録されているデータそのままで単純に重複がないか比較すると、-(ハイフン)の位置によって別の番号と見なされてしまうことがあることだ。文字置換機能や関数などを利用して一時的にハイフンを取り除いてから重複がないかを調べる。

年賀状などの宛名ソフトでは、CSV 形式のデータを読み込んだ際に自動的にデータに重複がないかを調べる機能があるものがある。この機能を利用してデータの重複を発見するのも一つの方法だ。

同一人物（取引先）が複数の人物（企業）として登録されている場合に、一人の人物としてまとめることを「名寄せ」と呼び、定期的なチェックが欠かせない作業だ。

■ 不整合

データを記録したシステムが複数あり、一方のシステムではデータは A となっており、もう一方では B になっているケースを不整合と呼ぶ。同じであるはずのデータが一致しないことにより、誤った住所に送付してしまう、しなければならない作業が抜ける、どちらが正しいのかわからない、などの不具合が生じる恐れがある。

チェック方法は、データ同士の突合だ。ただしこれで不整合が発見されても、どちらを正として扱うかは人間の判断が必要になる。

根本的な対応方法は、不整合が生じる理由を突き止めることだ。IT システム上の不具合なのか、業務運用上で不整合を防ぐ仕組みの漏れなのかを突き止め、不整合が生じていることを検知できる仕組み、不整合データは登録できないようにする仕組みを構築する必要がある。

■ 表記の揺れ

企業名においては、アルファベットで記載、・（ナカグロ）の有無など表記の揺れが生じる恐れがある。また、住所の場合は次のような揺れが少なくない。

A：XX 市○○町 1 － 1 － 1

B：XX 市○○町 1 丁目 1 番地 1 号

この場合ではどちらの表記であっても郵便物は届くため機械的な対応は困難だ。人間が見て判断する場合には大きな問題には至らないことが多いが、データを二次加工して利用する場合に「株式会社」を除いたつもりが「㈱」が残っている、顧客分析で正確に統計をとれない、などの問題が発生する。

郵便番号との組み合わせや住所を分割してデータを保存することで一定程度まで絞り込みをかけた状態で比較することで表記の揺れに対応できる。

なお、市町村合併があれば住所の表記揺れ、不整合が発生しているものと考えたほうがよい。

■ まとめ

今回は取り上げていないが、人名などに利用される漢字の「高」と「はしごだか」や、取引記録の時系列データなどを扱う場合にも注意が必要だ。

データクレンジングを行う意味は、情報を正確な状態に保つことだ。情報を正確に保つことで初めてデータを分析し、経営戦略に役立てることができる。（了）