

# 業務文書の品質向上を支援する文書間整合性診断

## 不統一な用語、定義の重複、記載内容の相違をすばやく発見

社内規程やソフトウェア開発に関する文書、及び行政機関へ提出する各種申請書などの文書は、複数の文書が密接に関連し合っています。このような文書群に不整合があると、作業や手続きの誤り、更にはコンプライアンス違反につながるリスクがあります。大量の文書間の不整合を手で発見し修正するには、膨大な時間と人的な労力が必要です。

東芝ソリューション(株)は、文書間の整合性チェックを支援する文書間整合性診断技術を開発しています。この技術は、大量の文書間から網羅的に不整合箇所を発見できます。

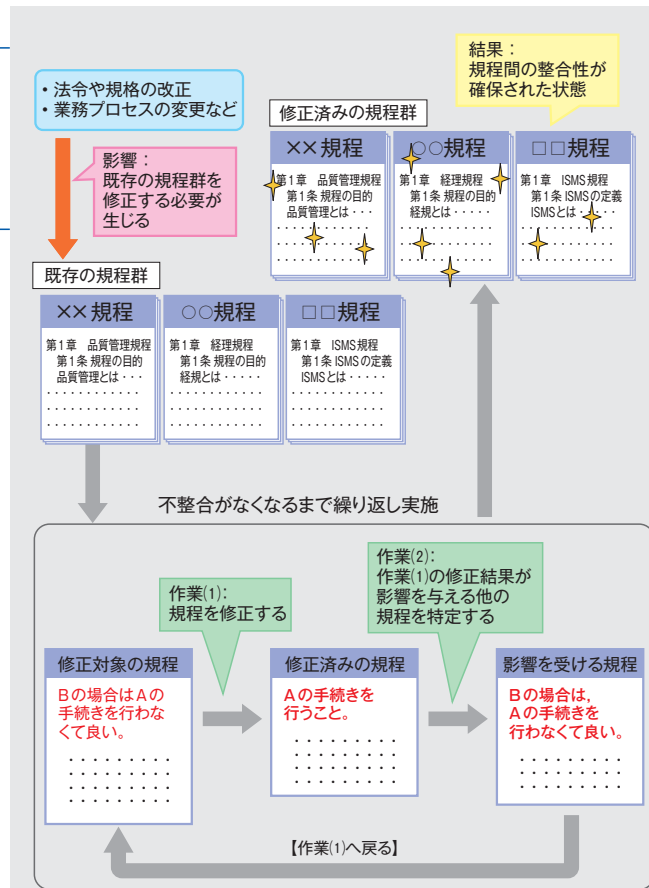


図1. 整合性チェック及び修正作業の流れ — 修正が発生するたびに、大量の文書群に対して繰り返し不整合チェックを行う必要があるため、膨大な時間が掛かっています。

e203393.txt	e3090397001.txt
(排出禁止物) 第33条占有者は、市長が行う家庭廃棄物の収集に際して、次の各号に掲げる家庭廃棄物を排出してはならない。 (1) 有害性の物 (2) 危険性のある物 (3) 引火性のある物 (4) 著しく悪臭を発する物 (5) 特別管理一般廃棄物に指定されている物 (6) 前各号に掲げるもののほか、家庭廃棄物の処理を著しく困難にし、又は家庭廃棄物の処理機能に支障が生ずる物 2 占有者は、前各号に掲げる家庭廃棄物を処分しようとするときは、市長の指示に従わなければならない。	(排出禁止物) 第13条占有者は、市が行う一般廃棄物の収集に際して、次の各号に掲げる一般廃棄物を排出してはならない。 (1) 有害性のある物 (2) 危険性のある物 (3) 引火性のある物 (4) 著しく悪臭を発する物 (5) 特別管理一般廃棄物 (6) 前各号に掲げるもののほか、市が行う処理に支障を及ぼすおそれがあるもの 2 占有者等は、前項各号に掲げる一般廃棄物を処分しようとするときは、市長の指示に従わなければならない。
「排出してはならない」に係る、「掲げる家庭廃棄物」という記述が、e3090397001.txtの「排出してはならない」と異なっています	
差異箇所	類似文章

図2. 抽出された記載内容の相違箇所 — 前後の文と説明文を確認しながら、修正すべき記載内容の相違を発見できます。

第32条の4 占有者は、資源物を排出しようとするときは、当該資源物が飛散し、流出し、及びその悪臭が発生しないようにしなければならない。 2 前項の規定に基づき排出された資源物(次項において単に「排出資源物」という。)の所有権は、市に帰属する。 3 市(市長が収集運搬する事業者として指定した場合は、当該事業者を含む。)以外の者は、排出資源物を収集し、又は運搬してはならない。	第3章 廃棄物の適正な処理 (一般廃棄物処理計画) 第12条 市は、法第6条第1項の規定に基づき一般廃棄物処理計画を地方自治法(昭和22年法律第67号)第2条第4項の基本構想に即して定めるものとする。 2 一般廃棄物処理計画は、基本的事項について定める基本計画及び基本計画の実施のために必要な各年度の事業について定める実施計画に分けて定めるものとする。 3 市長は、一般廃棄物処理計画を定めたとき、又は変更したときは、これを告示する。(平12条例12・一部改正)
(排出禁止物) 第33条占有者は、市長が行う家庭廃棄物の収集に際して、次の各号に掲げる家庭廃棄物を排出してはならない。 (1) 有害性のある物 (2) 危険性のある物 (3) 引火性のある物 (4) 著しく悪臭を発する物 (5) 特別管理一般廃棄物に指定されている物 (6) 前各号に掲げるもののほか、家庭廃棄物の処理を著しく困難にし、又は家庭廃棄物の処理機能に支障が生ずる物 2 占有者は、前各号に掲げる家庭廃棄物を処分しようとするときは、市長の指示に従わなければならない。	(排出禁止物) 第13条占有者等は、市が行う一般廃棄物の収集に際して、次の各号に掲げる一般廃棄物を排出してはならない。 (1) 有害性のある物 (2) 危険性のある物 (3) 引火性のある物 (4) 著しく悪臭を発する物 (5) 特別管理一般廃棄物 (6) 前各号に掲げるもののほか、市が行う処理に支障を及ぼすおそれがあるもの 2 占有者等は、前項各号に掲げる一般廃棄物を処分しようとするときは、市長の指示に従わなければならない。
(一般廃棄物管理票) 第40条 規則で定める事業者(以下「特定事業者」という。)は、事業系一般廃棄物を市長の指定する処理施設に運搬する場合には、規則で定めるところにより、事業系一般廃棄物の種類、排出場所等を記載した一般廃棄物管理票を市長に提出しなければならない。	(処理施設の受入基準) 第14条 市の廃棄物処理施設に一般廃棄物を搬入しようとする者(次項において「搬入者」という。)は、別に定める規則に従わなければならない。
文書1	文書2
類似文	類似文章

図3. 類似文章の抽出方法 — 一定間隔で文書中に出現する類似文をまとめ、類似文章を抽出します。文章単位に抽出することで、記載内容の相違だけでなく、記載の抜けも確認できます。

### 文書間の整合性チェックの課題

企業で作成される文書は、複数の文書間の記載事項が常に整合性の取れた状態にあることが要求されます。そのため、文書の追加や、変更、削除などが行われた場合、各文書に不整合がないことを確認する必要があります。

一例として、規程管理者が社内規程の不整合箇所を確認する作業を図1に示します。ここで不整合とは、複数の文書にまたがって出現する不統一な用語、用語定義の重複、及び記載内容の相違です。この作業のほとんどが人手で行われているため、確認漏れが生じるだけでなく、作業に膨大な時間と労力が必要になります。また、文書管理者が文書ごとに異なっていると、担当以外

の文書を含めた文書群全体に不整合が生じやすくなります。不整合が含まれた状態で文書を管理していると、業務の誤りや最悪のケースではコンプライアンス違反につながるリスクがあります。

### 文書間整合性診断技術

東芝ソリューション(株)は、前述の課題を解決するため、文書間に生じる不整合を抽出する文書間整合性診断技術を開発しています。この技術は、次の3種類の抽出技術から構成されています。

- (1) 用語の不統一抽出技術 用語表記の類似性を比較し、表記が不統一な用語を抽出
- (2) 定義の重複抽出技術 各文書で重複して定義されている用語とその内容を抽出

- (3) 記載内容の相違抽出技術 不整合が含まれやすい、類似した内容が記載される箇所を抽出  
ここでは、これら3種類の抽出技術の中でもっとも特徴的な抽出技術である、(3)の詳細を以下に述べます。

### 記載内容の相違抽出技術の特長

記載内容の相違抽出技術では、まず、内容が類似した箇所を類似した複数の文から構成される文章単位で抽出します。次に、この技術は類似文章間の差異もユーザーに提示します。抽出結果の例を図2に示します。

記載内容の相違抽出技術を構成する類似文章抽出と差異箇所抽出について、次に述べます。

類似文章抽出では、単語とその出現

頻度に基づき、複数の文書から類似文を抽出します。しかし、単文で類似した文を抽出しても、不整合箇所と判断するには、前後の文も含めて抽出元の文書を確認しなければなりません。そこで、この技術では一定間隔で出現する類似文をまとめた類似文章を抽出します(図3)。

差異箇所抽出では、抽出された類似文章から修正すべき不整合を発見しやすくするため、類似文章間に生じる差異を抽出します。また、差異を抽出するだけでなく、差異の説明文を自動生成します。抽出された差異と自動生成された説明文を参照することで、差異箇所だけでなく、この技術が差異と判定した根拠をユーザーに提示できます。

以上のように、記載内容の相違抽出

技術では、類似文章抽出と差異箇所抽出による抽出結果をユーザーが参照することで、文章に生じる不整合箇所を効率的に発見できます。

### 今後の展望

文書間整合性診断技術を用いることで、用語レベルから文章レベルまでの不整合を、網羅的かつ効率的に発見できます。その結果、人手で行っていた作業の労力を、大幅に軽減できます。

今後、この技術を次の3点についても対応させていきます。

- (1) 多言語文書への対応 人が翻訳した文書は、翻訳前後で対応が取れていないことがあります。このように翻訳前後の文書など、多言語で書かれた文書間の整合性

- チェックを実現します。
- (2) Big Dataへの対応 収集される大規模データには、文章を含むデータも含まれます。このようなBig Dataから高速に不整合箇所を抽出します。
- (3) 精度の向上 行政機関へ申請する書類は、高い厳密性が求められます。こういった文書に対しても、精度良く不整合箇所を抽出します。

小林 幹門

東芝ソリューション(株)  
IT技術研究所  
研究開発部