# データ品質ガイドブック

* https://cio.go.jp/sites/default/files/uploads/documents/data\_hinshitu\_guide\_beta.pdf
* 2021-06-04
* β版
* キーワード: データ品質、信頼性、品質評価、サービス品質、データ管理プロセス、データ品質管理フレームワーク
* 概要: データの品質向上をするための参考資料。データ整備や設計、活用を適切に行うためのガイドブック。

## 1. 背景と目的

### 1.2. 目的

本ガイドブックは、データの利活用や管理が効率的に行われるようにするため、以下の観点でデータ品質管理のフレームワークと評価モデルを示します。 - データの提供者や利用者が容易にデータ品質の評価を行えるようにする。 - データの提供者と利用者の間で品質の評価を共有できるようにする。 - 海外とのデータ連携や海外サービスによる利用の際にも活用できるようにする。

### 1.3. データ品質管理の意義

データ利活用が社会活動の基盤となる中で、品質に問題のあるデータがあると、新たなサービスの創出にかかる時間やコストに悪影響を及ぼします。

特に、データクレンジングと言われるデータ品質の改善作業が、データ提供者側ではなくデータ利用者側で行われていることは、社会的な非効率につながります。

本来、データクレンジングはデータ提供側で行われるべき作業ですが、多くのデータオーナーがその問題やデータ品質管理の意義に気が付いていません。データオーナーにとってのデータ品質管理の意義は以下のように整理されます。

* データ収集コストを低減できる。
* データ収集を迅速化できる。
* データ更新を容易にできる。
* データ更新にまつわる問題を回避できる。
* 内部でのデータ活用を容易にできる。
* データ公開を容易にできる。
* 利用者側でのデータ活用が進む。

## 2. データ品質評価モデル

### 2.2. 評価方法

本ガイドブックでは、データ品質を単純に評価するだけではなく、各項目の評価に成熟度モデル（マチュリティモデル）を採用しています。

マチュリティモデルとは、減点法ではなく、ステップごとの目標レベルを定め、目標に対し現在どこに位置しているのか、次に何をやるべきかを見定められる形で成熟度を測るという評価方法です。

具体的な評価に当たっては、各評価項目それぞれの達成度を４つの成熟度レベルで評価します。各データが基本レベルに達することを目指して取組を進めていき、最終的には、サステイナブル・レベルを目指します。

* アドホック・レベル
  + 独自の方法で実施している。当該サービスの目的を達成するための基本機能が実装されていない。
  + データ作成者による改善がされなければ、データとして利用することが困難である。
* 部分対応レベル
  + 基本機能が一部で実装されている。
  + 利用者側にてある程度の修復が可能であり、限定的な範囲でデータとして利用可能である。
* 基本レベル
  + 基本機能が実装されている。
  + データとして利用可能な品質である。
* サステイナブル・レベル
  + 基本機能を継続的に提供し、フィードバックをかけている。
  + 継続性のあるデータ、コンピュータシステムとして統合が可能である。

### 2.3. データの評価

データ自体の品質の確保は、データライフサイクルにおけるプロセスの整備や管理体制の整備などにより実現されます。実際に活用するデータそのものの品質ですので重要な評価項目になります。ショッピングサイトでは、製品の評価と出店者の評価が行われますが、データの評価は、製品（特に原材料）の評価に相当するものです。

ただし、データ自体の評価は評価時点でのスナップショットにすぎず、品質の良いデータであり続けるためには、継続的にデータが更新される必要があります。ある時、品質の良いデータを提供したとしても、その後データの品質が下がってしまっては意味がありません。そのため、データの品質を繰り返し確認することが重要になります。

データの品質は、網羅性の確認等、コンピュータによる検証が有効な項目が多いため、積極的にコンピュータによる確認を導入するなどの対策が効果的です。

本ガイドブックでは、データ自体の品質を、ISO/IEC 25012 に沿って評価します。

1. 正確性 (Accuracy)
2. 完全性 (Completeness)
3. 一貫性 (Consistency)
4. 信憑性 (Credibility)
5. 最新性 (Currentness)
6. アクセシビリティ (Accessibility)
7. 標準適合性 (Compliance)
8. 機密性 (Confidentiality)
9. 効率性 (Efficiency)
10. 精度 (Precision)
11. 追跡可能性 (Traceability)
12. 理解性 (Understandability)
13. 可用性 (Availability)
14. 移植性 (Portability)
15. 回復性 (Recoverability)

## 4. データ品質改善への工夫

### 4.1. データ標準の活用

データ設計時にデータ標準を活用することで、設計品質や相互運用性の高いデータ設計をすることができます。

データ設計に当たっては、政府 CIO ポータルサイトに掲載されている行政データ連携標準、IMI 共通語彙基盤、推奨データセットや schema.org 等を参考にすることができます

### 4.2. データ入力フォームの活用

データ入力フォームを使うことで、データ収集時のデータ形式の揺らぎや誤データの混入を防ぐことができます。フォーム作成の外部サービスを、目的、データのセキュリティレベルに合わせて選択して活用することも検討します。

### 4.3. コードやコントロールド・ボキャブラリの活用

選択肢があるデータ項目は、自由記述ではなく、コード入力や選択肢入力にすることで誤データの混入を防ぐことができます。ただし、東京都と入力する代わりに東京都の地方公共団体コードを入力してくださいというのは利用者には不便ですので、選択肢で東京都を選択するとデータにはコードが書き込まれるといった工夫をしていくことが重要です。この時の選択肢のことをコントロールド・ボキャブラリと呼びます。

コードやコントロールド・ボキャブラリは、データ設計時に独自に設定するのではなく、できるだけ標準的な選択肢を設定してください。そうすることで、他データとの連携も容易に行えるようになります。また、コントロールド・ボキャブラリを作る時には、データ項目定義を明確にすることが必要です。他データと連携する時に検討が容易になります。

コードやコントロールド・ボキャブラリは、国際標準、総務省統計局、政府CIO ポータルサイトのコード一覧等で整理されている既存の体系をできるだけ引用するようにしてください。

### 4.4. データクレンジングツールの活用

データのクレンジングには、機械的に処理できる工程がたくさんあります。OpenRefine や IMI コンポーネント等のツールを必要に応じて活用してください。 ただし、ツールを使って一括処理する際には、誤変換を起こさないように細心の注意を払う必要があります。

紙のデータを OCR やパンチでデジタル化したデータは、誤データが含まれる可能性があることから、既存のデータとの照合、計算による検証、他社の OCR結果と認識結果の照合をする等、必要に応じて確認プロセスを検討します。