

データの品質、 どう測る？

下山 紗代子

政府CIO補佐官
(内閣官房IT室)

2020.10.13 (Tue.)

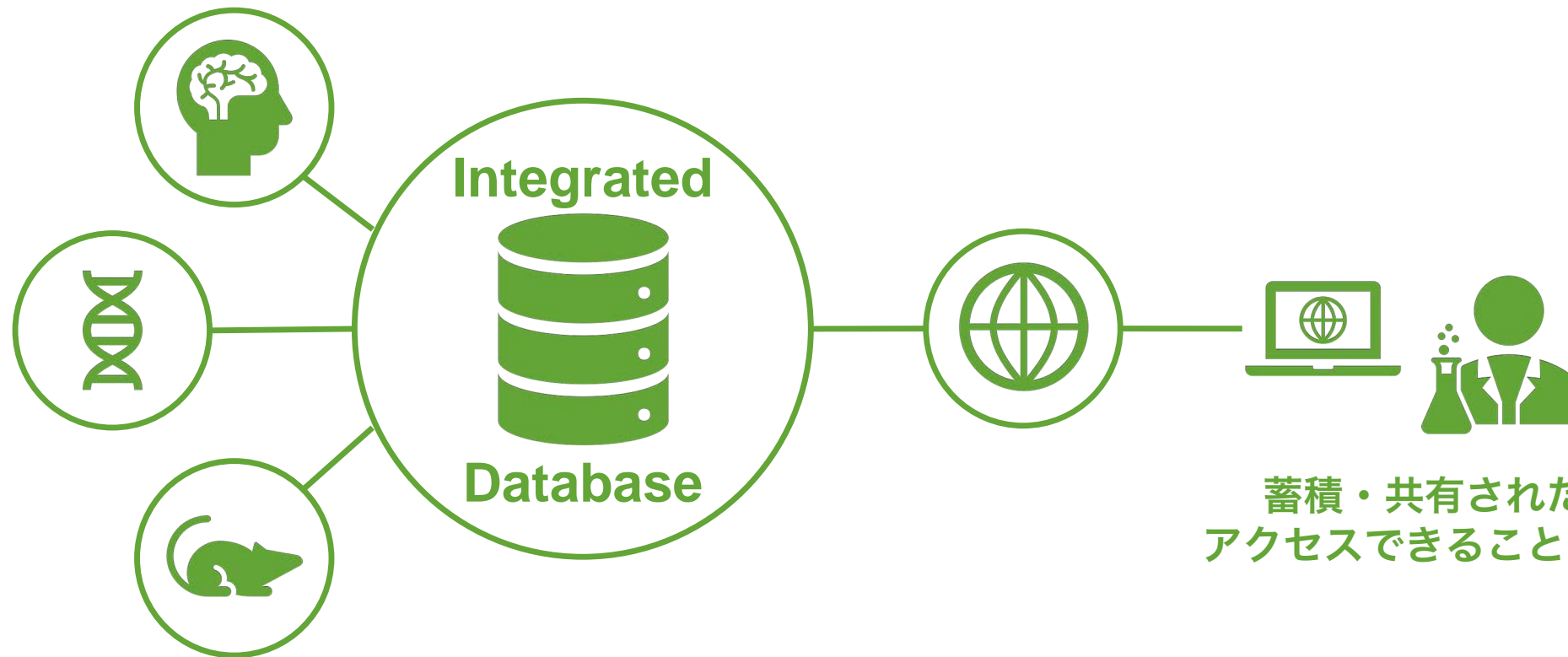
SCI-Japanウェビナー

日本型データ社会加速の
シナリオ





Sayoko Shimoyama



蓄積・共有されたデータに
アクセスできることで研究が可能

Background:

分子生物学→バイオインフォマティクス

原点は学部時代の卒論提出3ヶ月前の事件

卒業した先輩から引き継いだデータが間違っていたことが判明



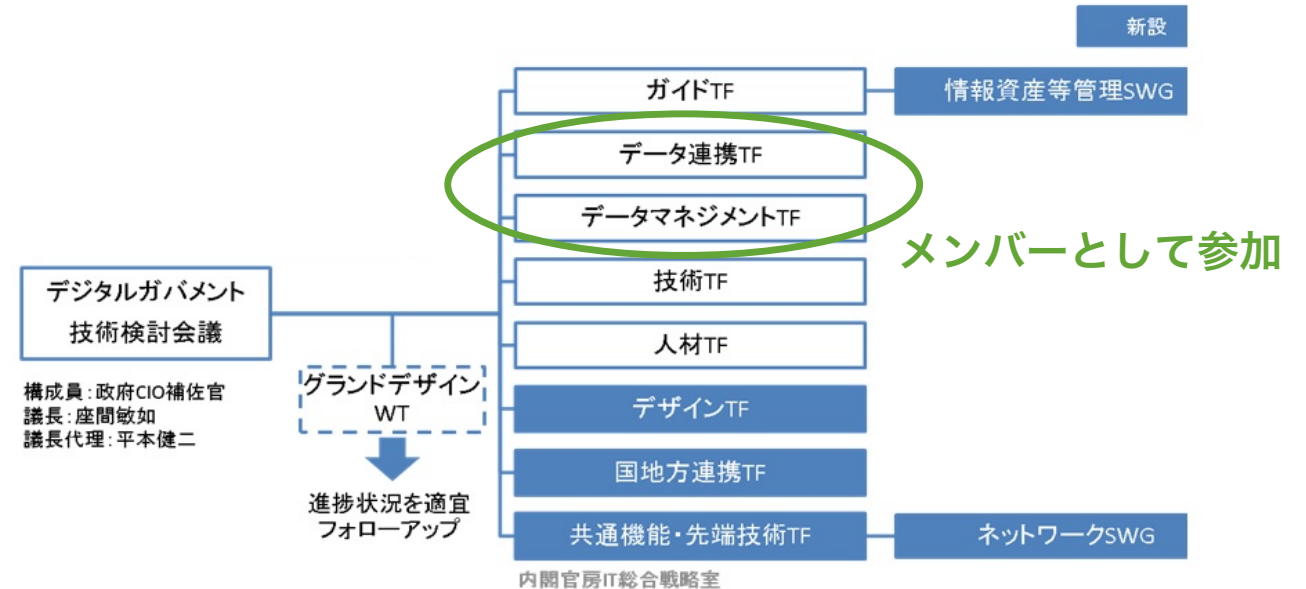
実験全部
やりなおし

データは見ただけでは
間違いに気付けないことが多い…

政府におけるタスクフォース活動

デジタル・ガバメント技術検討会議

- 政府CIO補佐官で構成される会議体
- 政府職員だけでは解決が困難な技術的、専門的な課題等について検討するために設立された
- 現在8つのタスクフォースが稼動中



民間企業における 「データスチュワード」という役割

データスチュワード (Data Steward) の役割

(DAMA-DMBOK : データマネジメント知識体系ガイドによる)

1. 主要メタデータの作成と管理 (Creating and managing core Metadata)
2. ルールと標準の文書化 (Documenting rules and standards)
3. データ品質課題の管理 (Managing data quality issues)
4. データガバナンス業務の諸活動の実施 (Executing operational data governance activities)

1では、なにを重要なメタデータとするか決めて、その値と意味を定義し、管理する

2のルールと標準には、データとその品質に関するものだけでなく、業務に関するルールも含まれる。(業務ルールが明確でないと、そこで作られるデータの品質を保証するのは難しくなるため)

3では、データ品質の課題を特定し、解決に取り組む。

4では、全体最適化の視点を持って、プロジェクト等の個別活動を統制する。

データの
品質管理の
専門職

データエンジニアリングの知識
+データ化されている対象の知識
が必要

参考：医療系ベンチャー「ミーカンパニー」における活動

<https://mecompany.me/member/>



開発部

データスチュワード

大学院では生命理工学系を専攻。国立研究所で統合DBの研究開発に携わり異文化のデータを組み合わせる面白味に目覚める。

その後開発したデータの公開支援を目的に一般社団法人リンクデータを設立。代表理事として全国の公共団体等のオープンデータ公開をサポートする傍ら、2017年6月よりデータスチュワードとしてミーカンパニーにJoinする。

オープンデータを活用し付加価値を生み出している環境に出会い、スタッフが専門性を活かしリスペクトし合っている社風に魅力を感じた。

一般社団法人リンクデータで多くの方々をサポートしていく中、もっと技術やコンサルティングスキルを積む必要があると実感していた私。そんなある日、友人だったミーカンパニーのCDOより紹介されたのがミーカンパニーでした。Joinした決め手は、理事を務めるリンクデータの業務と掛け持ちで、専門職として関わられるフレキシブルな働き方が可能だったこと。またSler等にありがちな営業vs技術者という図式がなく、対等な立場でリスペクトし合いながら仕事が進められるフラットな風土が、とても魅力的に映りました。

「SCUEL」で用いる医療系データの品質管理や精度保証を担う中で実現できていないことに挑戦する企業風土に感銘を覚える日々。

データスチュワード……なかなか聞き慣れない職種名ですが、簡単に言えばデータの品質管理や精度保証を担う人のことを言います。現在私は、ミーカンパニーが取り扱う医療系データのデータスチュワードとして事業に携わっていますが、少数のデータベース企業でここまで徹底的に、データの品質や精度にこだわっている例を他に知りません。また他では実現できていないことに挑戦していくという姿勢が素晴らしく、難しい課題も多いですが、それがやりがいにつながっています。また私はミーカンパニー以外にも複数の組織に所属しており、常駐ではない分、外で得た知見をフィードバックできる立場にいますが、そんな自由な関わり方が可能なミーカンパニーは、本当に稀有な存在だと思っています。

より良い仕組みを作るためのディスカッションが盛んな環境ですので、判断軸を持ちつつ相手を尊重でき

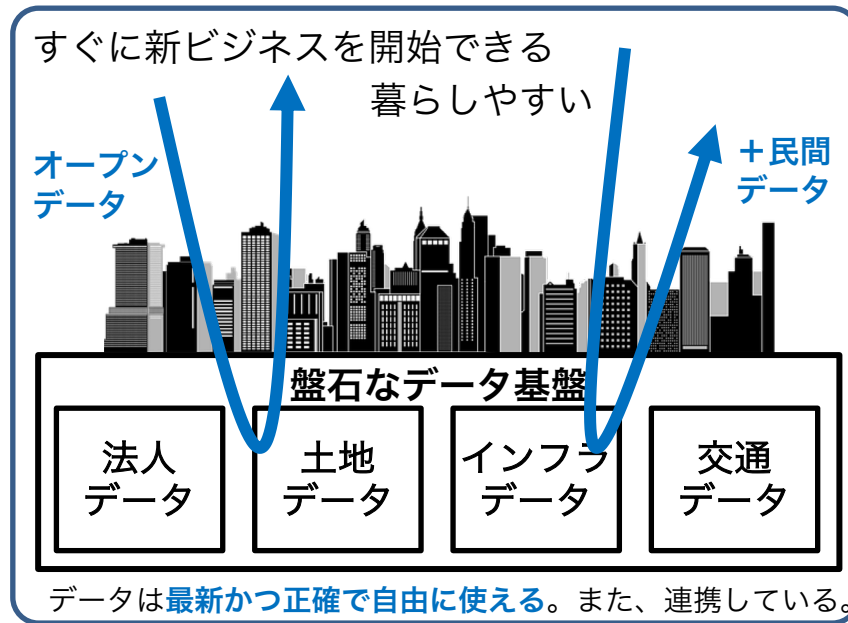
1. なぜ、 データ品質管理は 重要なのか

データ環境が国の競争力の源泉になっている

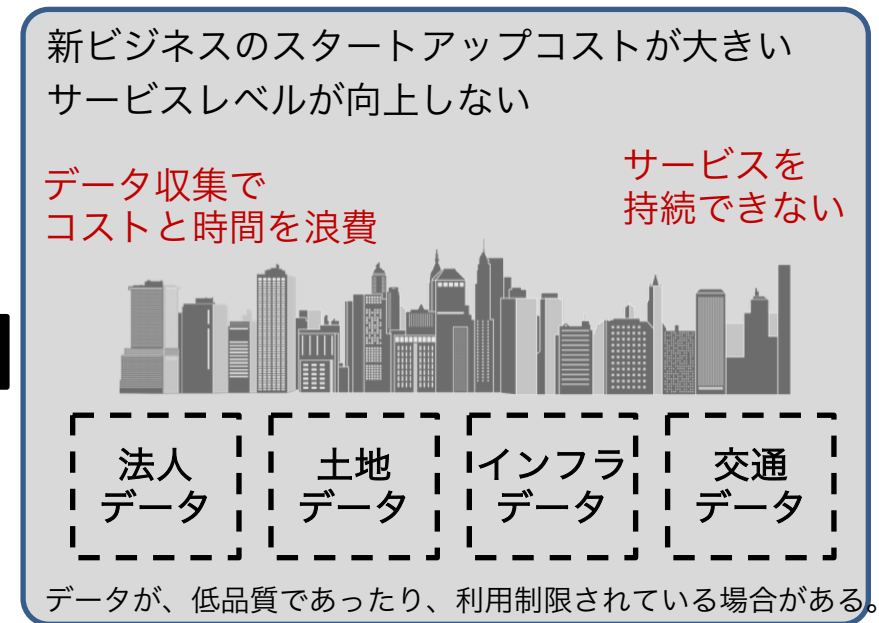
■ データが安価に安定的に供給される持続可能なエコシステムが必要

- 社会の基本データはデジタル時代のインフラであり、地力（ポテンシャル）である

データ基盤が整備済みの国・都市



データ基盤の整備が遅れている国・都市



人や企業、投資は、より魅力的な場所へ移動

- ## ■ 50年後、100年後のデジタル社会を展望したデジタル社会の基盤として、エストニアは20年、デンマーク、オランダは10年以上かけて整備してきている

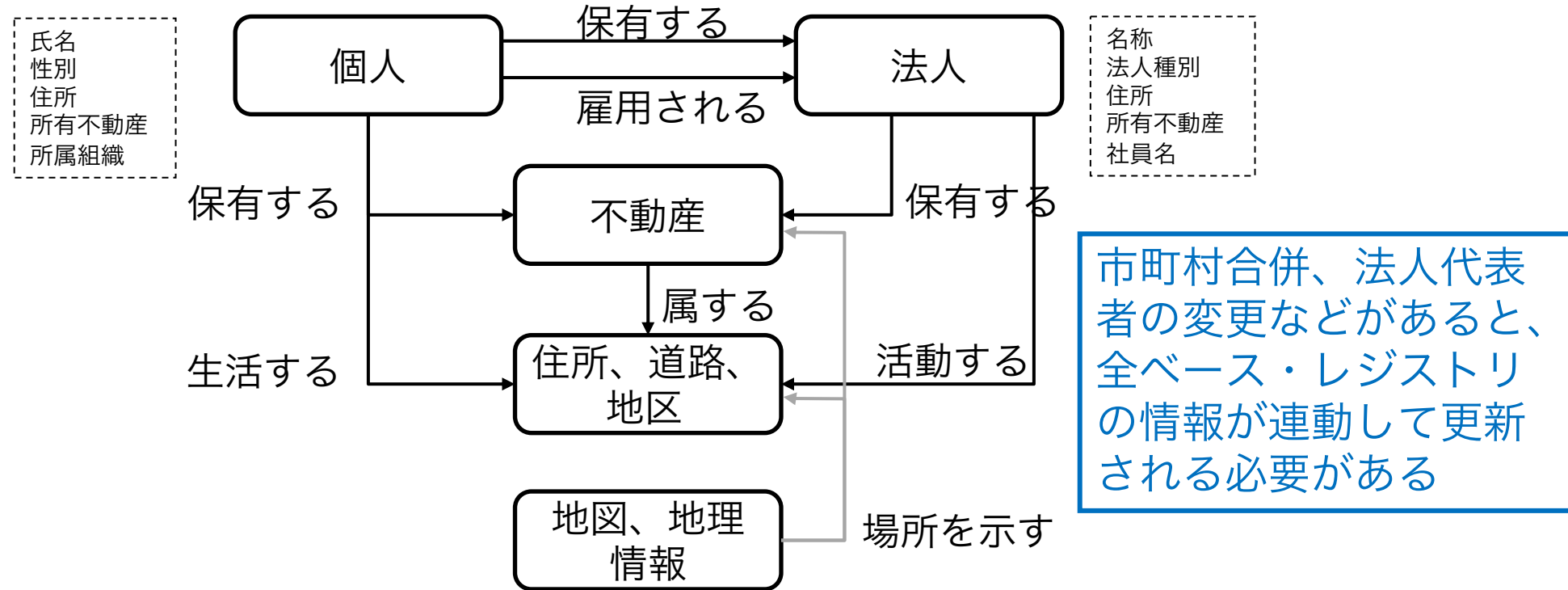
デジタル時代の必須インフラ：ベース・レジストリ

- ベース・レジストリとは、公的機関等で登録・公開され、様々な場面で参照される正確性や最新性が確保された社会の基幹となるデータベース
- 人、法人、土地、建物、資格等の社会の基本データであり、日本では台帳等が相当する場合が多い
- 全ての社会活動の土台であり、デジタル社会における必須の環境
- ベース・レジストリの有無が、国の競争力を左右する
- AIやドローン、自動運転等の最新のデジタルテクノロジーを支える基盤



ベース・レジストリは連携している必要がある

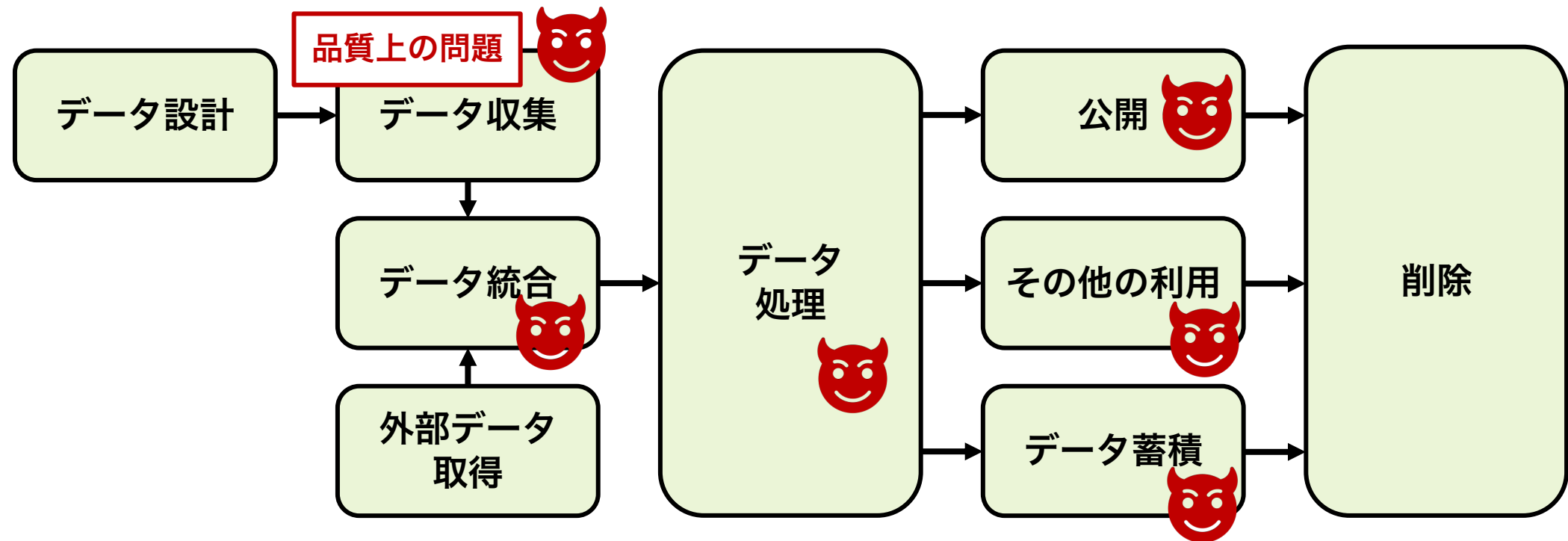
- ベース・レジストリは単体でも効果はあるが、複数のベース・レジストリを相互参照することで、その効果を飛躍的に増大させることができる



- 相互にエラーレポートを報告することも重要
 - サービスの中でベースレジストリ情報に間違いがあったときに修正を反映するプロセスを決める

低品質のデータがあると サービス全体に最後まで影響が残る

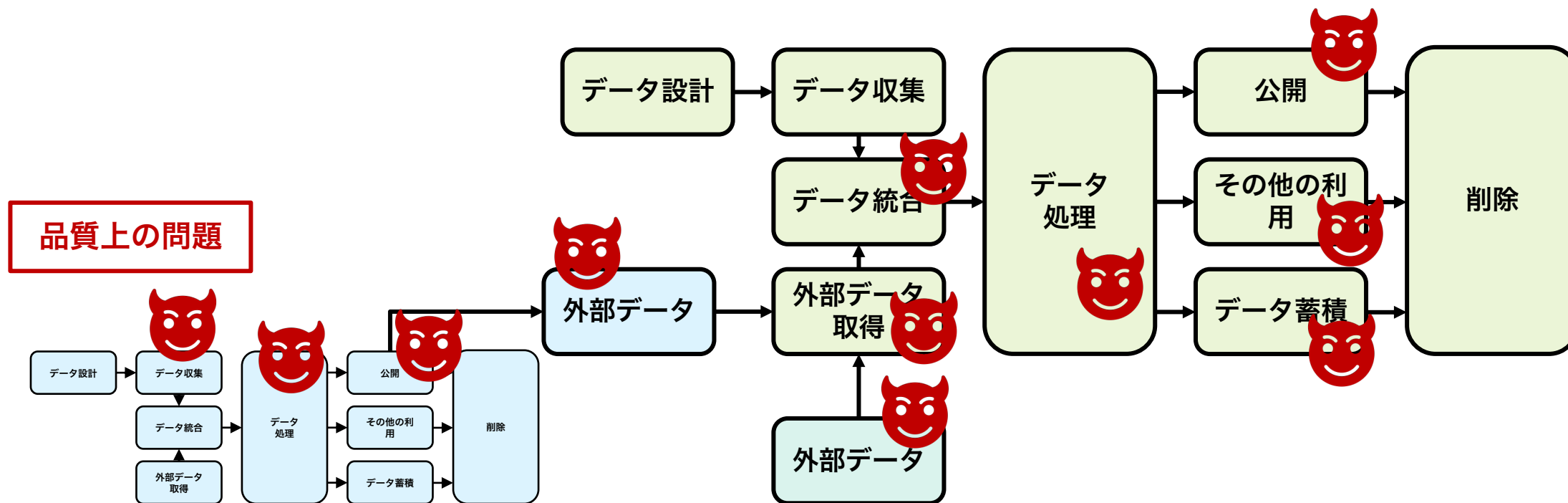
品質の低いデータを使ってサービスを実現すると、途中のサービスをどんなに品質高く作っても、最後まで低品質なデータの影響を受け続けることになる



ISO/IEC 25024:2015 Systems and software engineering -- Measurement of data qualityより翻訳

低品質のデータが複数種類混ざると品質改善のコストが増大

複数のデータソースを使っていたり、データの加工プロセスが複雑な場合、原因を特定して対処するまでに膨大なコストを要することがある



事例：某店舗MAP

某キャンペーンに参加している店舗を地図上に表示するサービスが公開されたが...



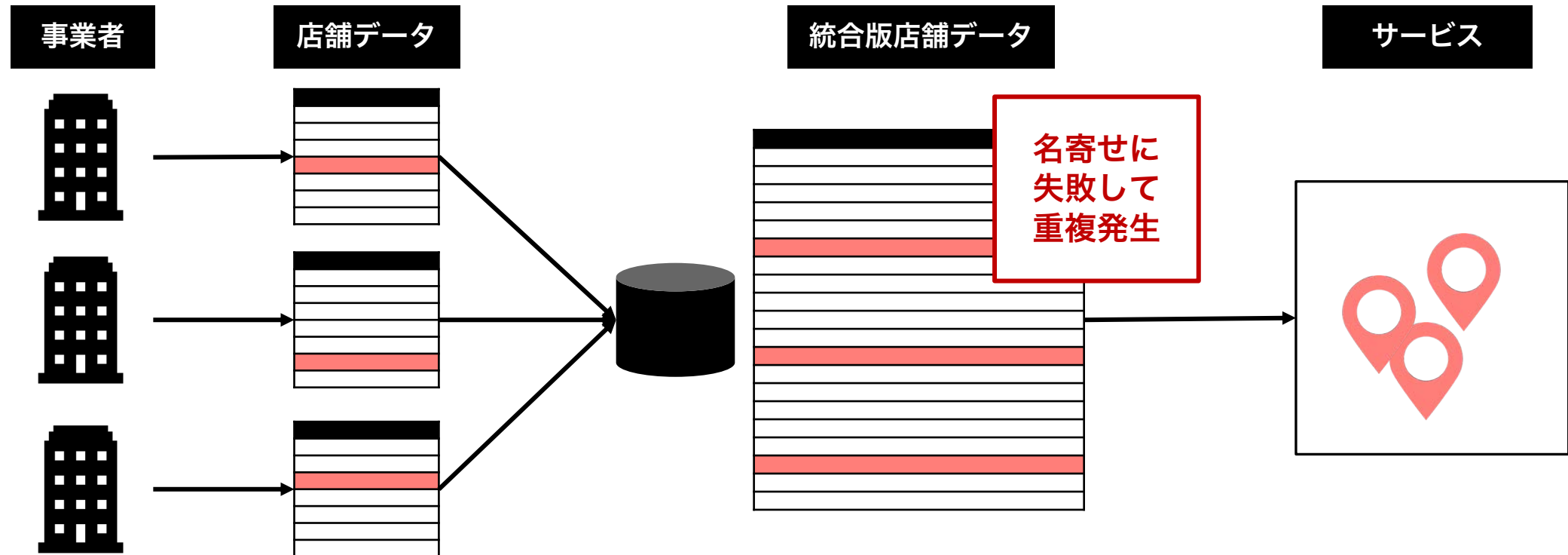
同じ店舗名が複数

実際の位置とずれた場所にピンが立ってしまったり（道路上とか）

地図表示のアプリケーションの
UI/UXを工夫したところで
サービス品質は向上しない

サービスの裏側で起こっていた可能性のあるデータの問題

複数のデータソース由来の店舗情報の「名寄せ」が完全に出ていなかった
(名寄せ：同じ対象を示すデータが重複しないように処理すること)



参考：名寄せは意外と厄介

名寄せの受託サービス等ではかなり高度なロジックを組んでいたりする

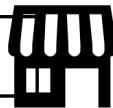
店舗名だけでは無理

同じ名前の別店舗があったり



スナック恭子

スナック恭子



表記揺れがあったり

スナック恭子

すなっく恭子

スナックきょうこ

店舗名+住所？

表記が何パターンもあるので
まず住所の処理が必要

東京都千代田区霞が関3-3-1
東京都千代田区霞が関3-3-1
東京都霞が関3-3-1
千代田区霞が関3-3-1
港区霞が関3-3-1
東京都千代田区霞が関三丁目3番1号
東京都千代田区霞が関三丁目3-1
東京都千代田区霞が関3丁目3番1号
東京都千代田区霞が関3丁目3-1
東京都千代田区霞が関3丁目3-1

店舗名+電話番号？

市外局番あり/なしの混在

03-5253-2111

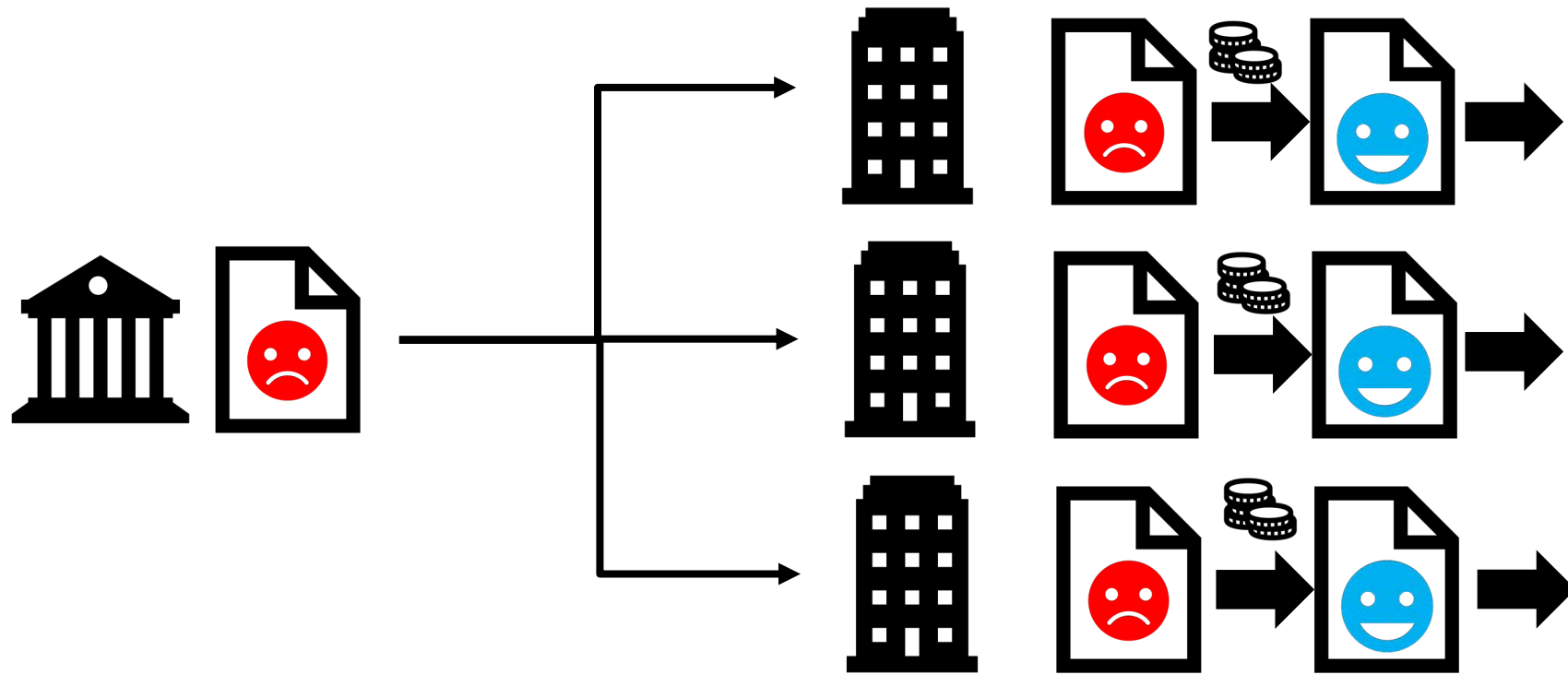
5253-2111

移転の関係で変わっていたり
(データソースごとの
収集のタイミングの違い)

03-5253-2111

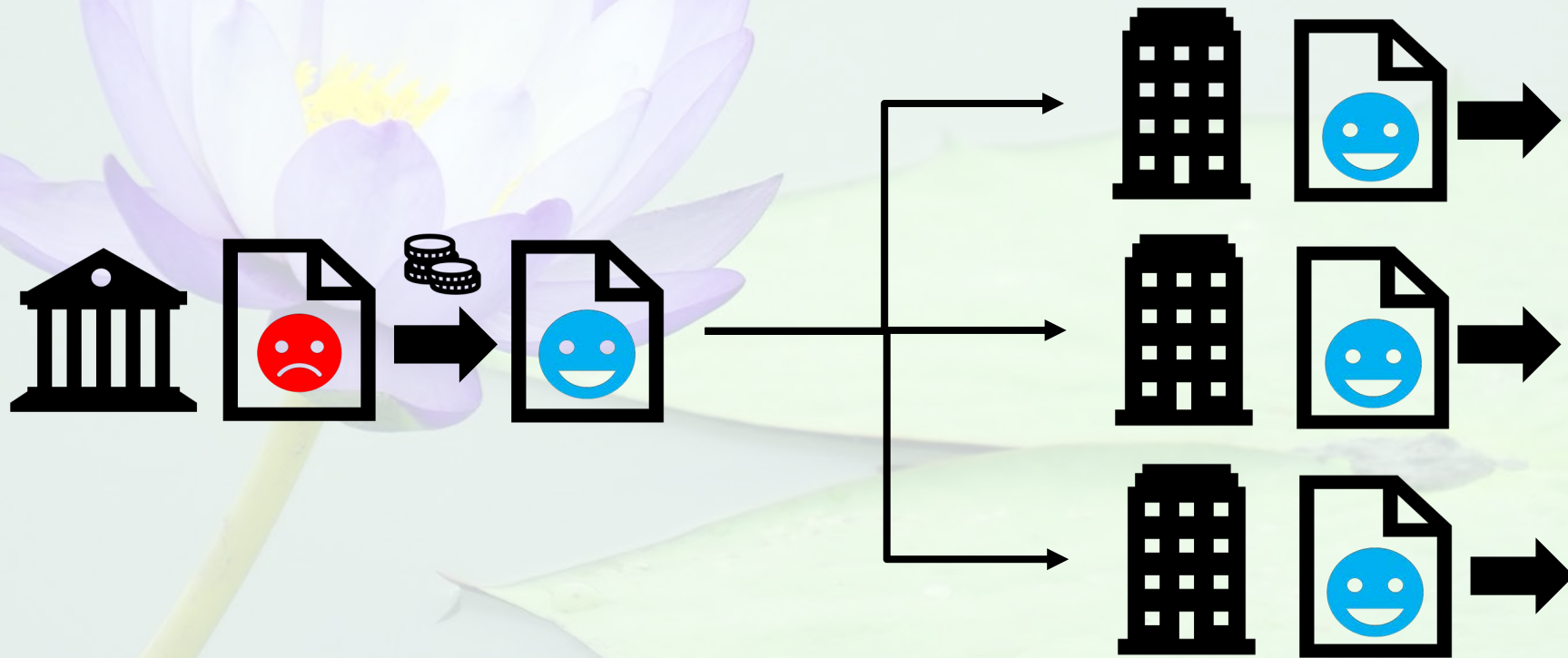
03-6910-0256

低品質データがもたらす経済損失



利用者側で都度加工コストがかかる

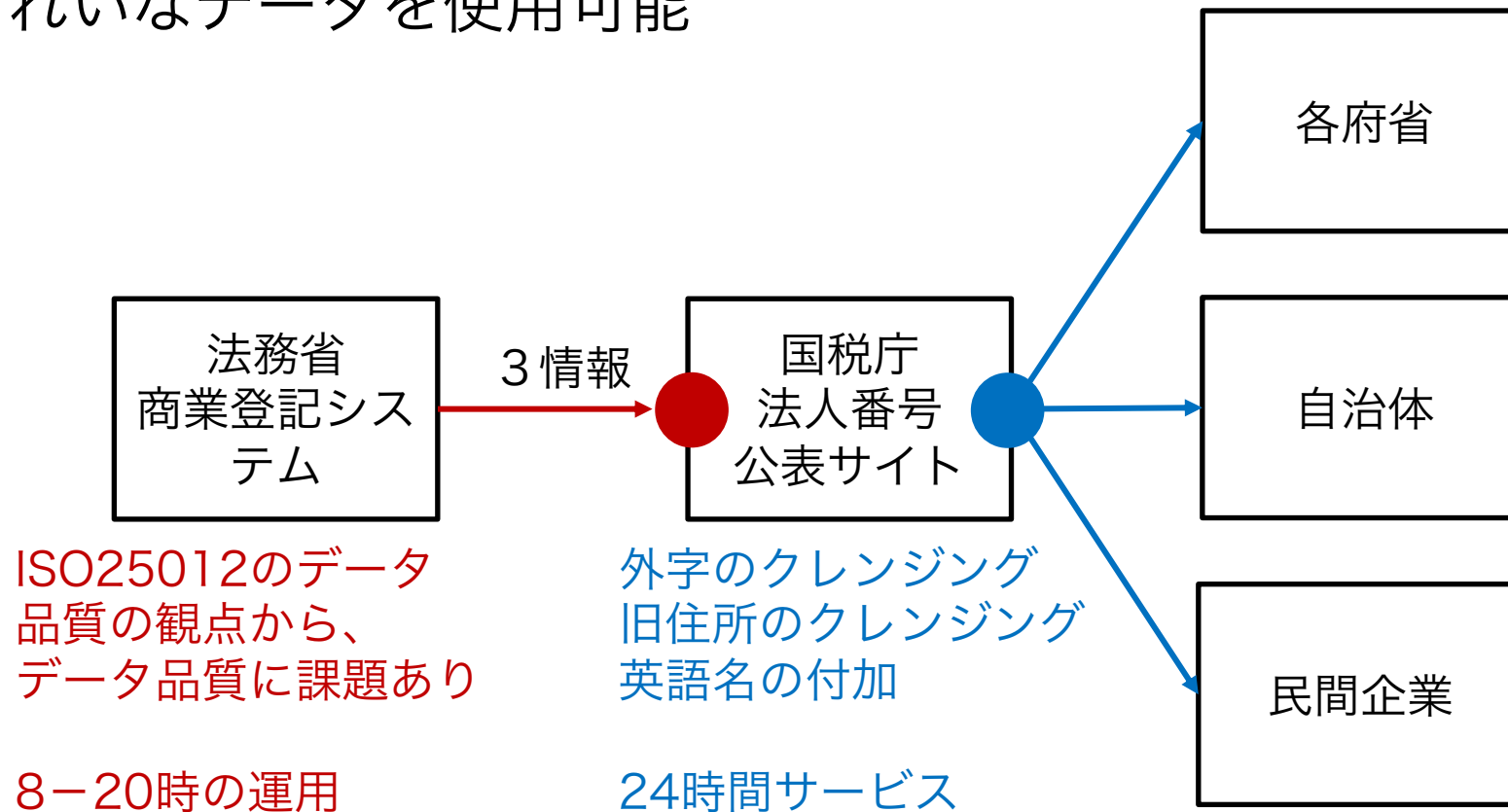
公開元で高品質データに変換して公開



社会全体で無駄なコストを減らせる

社会全体のコスト削減に成功した例：法人番号公表サイト

- 国税庁がデータ入手時にデータをクレンジングしているため、利用者である各府省、自治体、民間企業はクレンジングにコストをかけずきれいなデータを使用可能

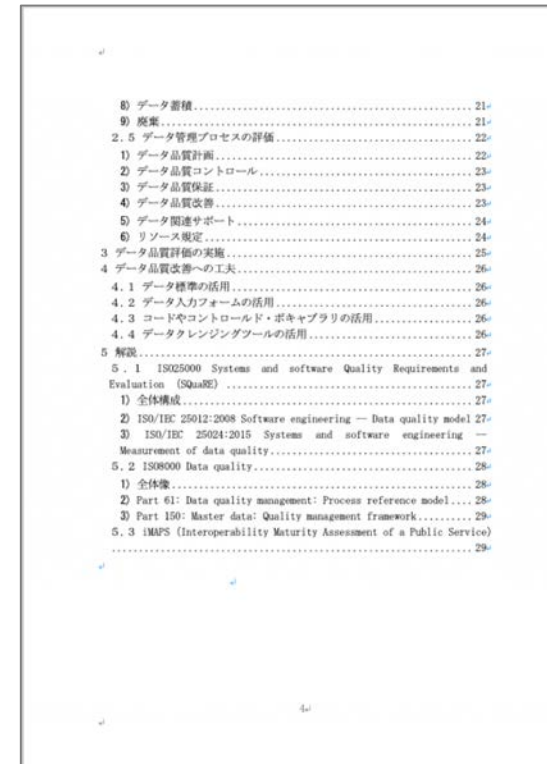
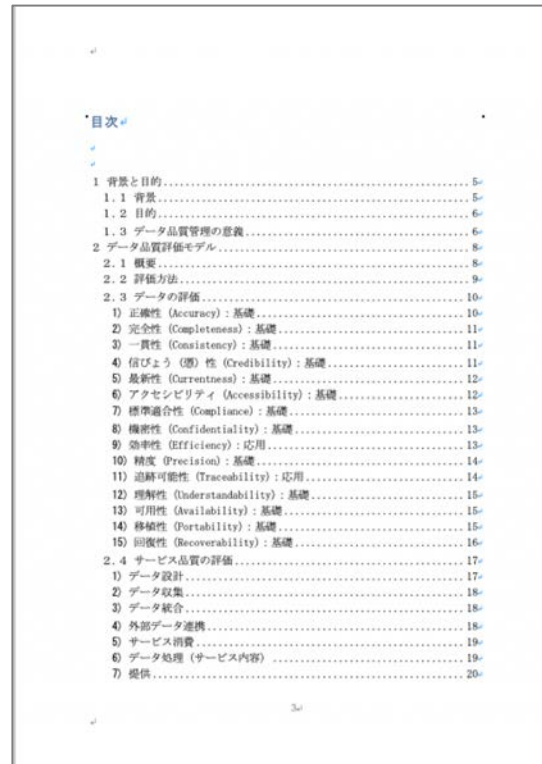
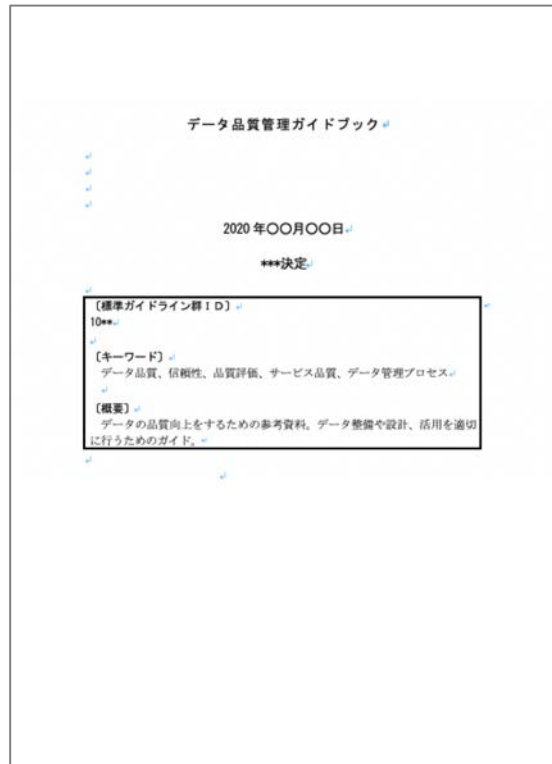


- 上場企業4000社が、データクレンジングで年間100万円コスト削減できたと仮定すると、年間40億円のコスト削減効果が見込める

2. 「データ品質管理 ガイドブック」 ドラフト版のご紹介

「データ品質管理ガイドブック」 鋭意作成中

「データ連携タスクフォース」で執筆を進めている



注意点

- 「データ品質管理ガイドブック」は現在ドラフト版のため、公開までに内容は変更になる可能性があります
- 「こんな内容が入っていると良い」といったフィードバックを歓迎いたします！
(必ず反映できるとはお約束できないのですが、一つ一つきちんと検討させていただきます)
- 本日は話す内容は、私の個人的な解釈であり、政府全体の見解ではありません

何をするためのガイドブック？

データ品質の評価モデルを提示することで、**行政機関**や**民間企業**においてデータの利活用や管理が効率的にできる環境の実現を可能にし、**デジタル社会を支える盤石なデータ基盤の構築**につなげる。

そのために、以下の観点で整理を行った評価モデルを本ガイドブックで提示する。

- 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うことができる具体的なモデル
- データの提供者と利用者の中で共有可能なモデル
- 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モデル

データ品質管理によって 誰にどんなメリットが生まれるのか

データ所有者（データオーナー）

- 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
- データ更新が容易にできる
- データ更新にまつわる問題を回避できる
- 組織内部でのデータ活用が容易にできる
- データ公開が容易にできる
- 利用者でのデータ活用が進む

データ利用者

- 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
- データ更新が容易にできる
- データ更新にまつわる問題を回避できる
- 組織内部でのデータ活用が容易にできる
- 提供するサービスの信頼性を向上できる
- 意思決定における誤りを回避できる

データ品質を 測る 3つの 国際標準



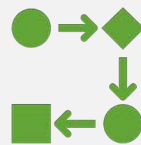
ISO/IEC 25012 (データ品質の評価)

: データそのものの品質



ISO/IEC 25024 (サービス品質の評価)

: データを使ったサービス実現プロセスに関する品質



ISO/TS 8000-61 (データ管理プロセスの評価)

: データの整備から活用までの管理プロセスに関する品質



ISO/IEC 25012 (データ品質の評価)

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)

利用者の視点から外形的に評価することが可能



ISO/IEC 25012 (データ品質の評価)

1. 正確性 (Accuracy)

データの基本は正確であること。
データの正しさを以下の点に着目して評価する。

評価項目

- 書式が正しいか
- 誤字脱字などはないか
- 意味的な誤りがないか
- データに誤りはないか

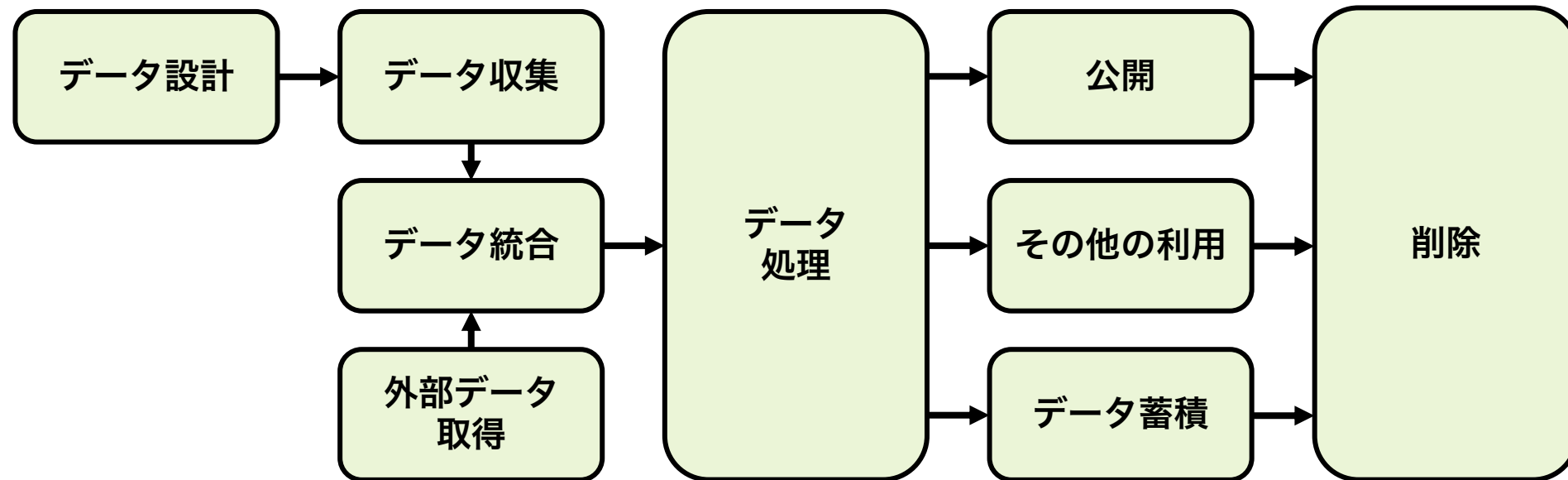
問題となる例

- 「同上」「//」などの記述がある
- 日付や数字が入るべき欄に「不明」など数字以外の文字列が記述されている
- 住所が入るべき欄に電話番号が入っている
- フリガナにカタカナとひらがなが混在している



ISO/IEC 25024 (サービス品質の評価)

- サービスを提供するプロセスの品質をデータライフサイクルのステップごとに評価
- EC (European Commission) で検討されているインターオペラビリティのための品質評価モデル：iMAPS (Interoperability Maturity Assessment of a Public Service) も参考に評価項目を設定





ISO/IEC 25024 (サービス品質の評価)

3) データ統合

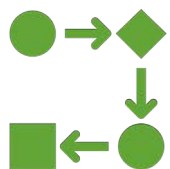
- 複数データの統合時に、データ項目のマッチング、精度や単位の確認、コードのコンバージョンなどを正確に行う必要がある。データ統合時に品質の劣化を起こさない仕組みが整備できているかを評価する。

評価項目

- 重複データの確認プロセスがあるか
- コード変換表など統合に関する情報を公開しているか
- 統合元データの出所を明確にしているか

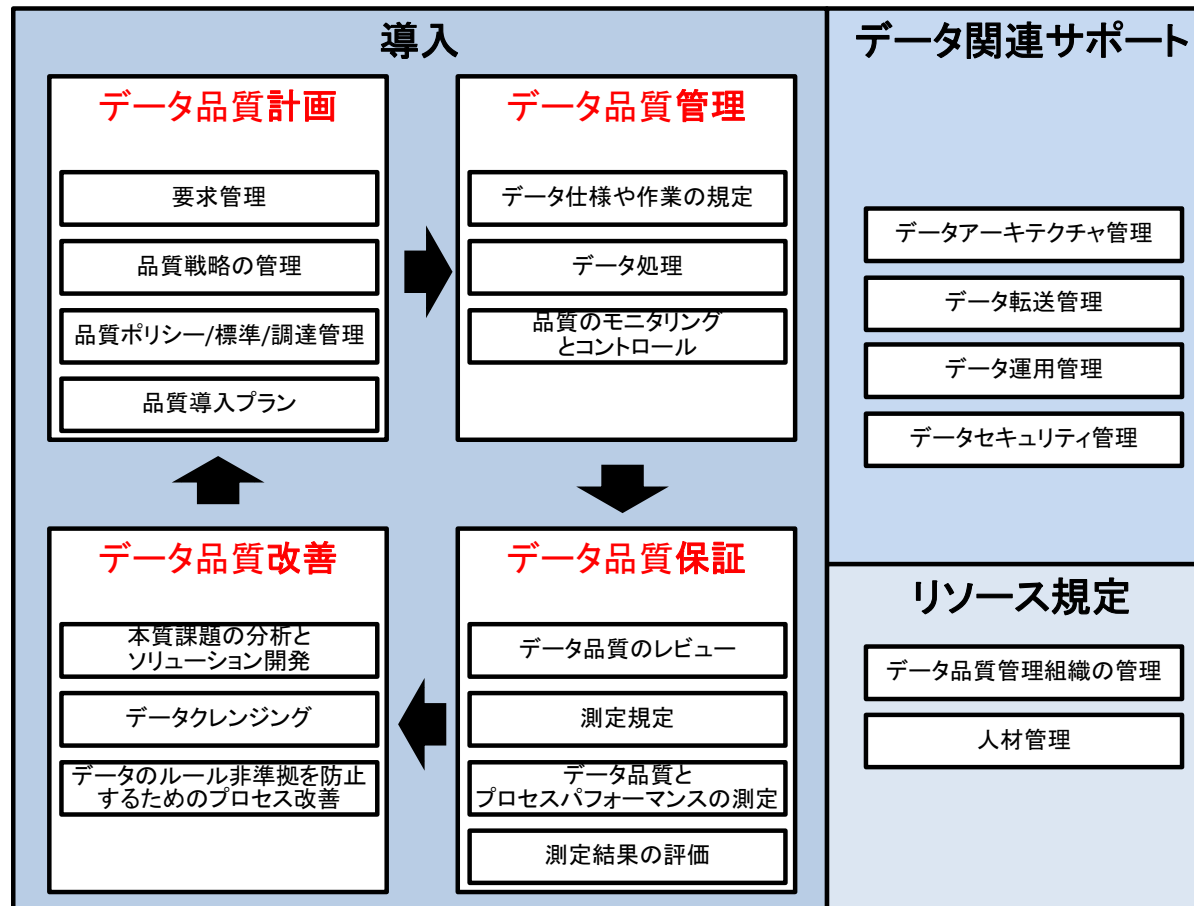
問題となる例

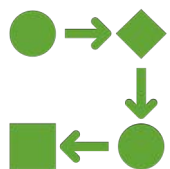
- 有効数字の違う数字を単純に加算してしまう
- コード変換表を公開していないので、詳細分析ができない
- 同一名の別法人のデータを結合してしまう



ISO/TS 8000-61 (データ管理プロセスの評価)

- データ品質やサービス品質を維持、向上していくためにはデータ管理プロセス及び運用体制が重要
- ISO/TS 8000のデータクオリティのプロセス管理の規格である part61 に沿って評価





ISO/TS 8000-61 (データ管理プロセスの評価)

1) データ品質計画

- データの品質を維持・管理していくためには、計画をたてて実行することが重要です。データ品質管理が計画的に行われているかを評価します。

評価項目

- (応用) データ品質に対する要求の管理をしているか
- (応用) データ品質を維持・向上させる戦略の管理ができているか
- データ品質ポリシー／標準／調達管理等のルール化ができているか
- データ品質導入プランを持っているか

問題となる例

- 品質の要求要件や基準が明確になっていないため、何から手を付けて良いのか分からない。

「成熟度モデル」方式による評価

成熟度モデル (Maturity Model) :

単純にできている/できていないで加点/減点する方式ではなく、現在の自分の組織の立ち位置を確認し、次のステップを目指せるようにするためのモデル

アドホック・レベル

- ・独自の方法で実施している。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なければ、データとして利用するのが困難である。

部分対応レベル

- ・基本の条件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

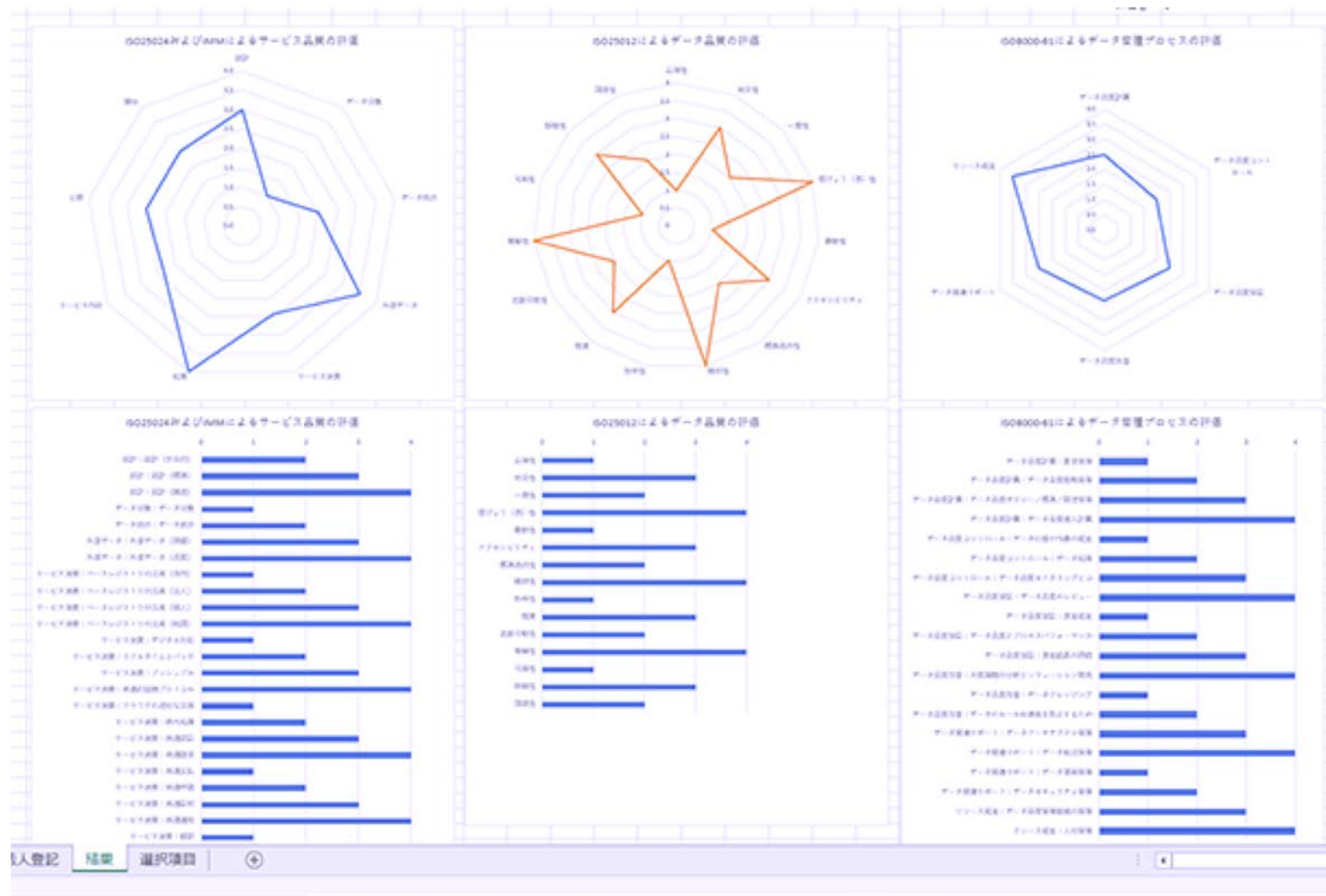
- ・実現すべき機能・レベルである。
- ・データとして利用可能な十分な品質である。

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。

データ品質 評価ツール

- ガイドライン付属の評価ツール
- Excelファイルで各項目の質問に対する実現レベルを答えるとチャートで可視化される
- データ関連サービスの課題を明確にするとともに、他サービスと比較することでベンチマークが可能
- 年に1回実施するなど、その改善状況を公開するとともに、改善計画などを策定していくことが重要



データ品質改善への工夫

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、設計品質やインタオペラビリティの高いデータ設計をすることができる

■ データ設計には、政府CIOポータルサイトにある行政データ連携標準、IMI共通語彙基盤や、schema.org等の国際標準を参考にすると良い

○ データ連携モデル

行政基本情報データ連携モデル

略称 行政データ連携標準

最終改定 2020年5月14日

対象 各府省

概要 日付時刻、住所、電話番号等、手続や情報提供において分野を問わず使用される基本的なデータの形式について、データ連携を円滑に行えるよう、基本的なデータの記述形式を示したモデル。

・ 日付時刻	PDF 	DOCX 
・ 住所	PDF 	DOCX 
・ 郵便番号	PDF 	DOCX 
・ 地理情報	PDF 	DOCX 
・ 電話番号	PDF 	DOCX 
・ POIコード	PDF 	DOCX 
・ POIコード一覧	PDF 	XLSX 

<https://cio.go.jp/guides#renkeimodel>

データ品質改善への工夫

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、データ収集時の誤データの混入を防ぐことができる
- Googleフォームのような無料サービスも活用可能（ただしデータのセキュリティレベルを考慮する）
- バリデーション（不正値のチェック）を実装/設定することが重要



しゅうまいの消費に関するアンケート

以下の3つの設問に回答してください。

*必須

1. 居住地：あなたは2019年1月1日～12月31日にどこに住んでいましたか？*

引越した場合は長く住んでいた方を選択してください。

横浜市内

横浜市外

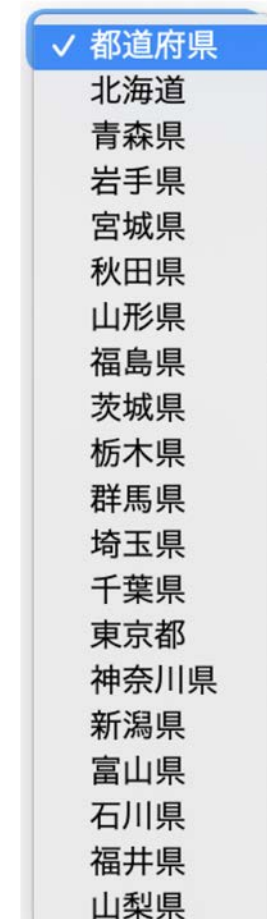
2. しゅうまい：あなたは2019年1月1日～12月31日にどのくらいしゅうまいを食べましたか？*

一番近いと思われる数を選択してください。

データ品質改善への工夫

③コードやコントロールド・ボキャブラリの活用

- 自由記述ではなく選択肢入力にすることで誤データの混入を防ぐ
例) 都道府県名、市区町村名
- 利用するコードやコントロールド・ボキャブラリは、標準的なものを採用することで、データ品質をより高められる
- ISO/JIS等の標準、総務省統計局、政府CIOポータルサイト等のコード一覧等で整理されている既存の体系を利用すると良い

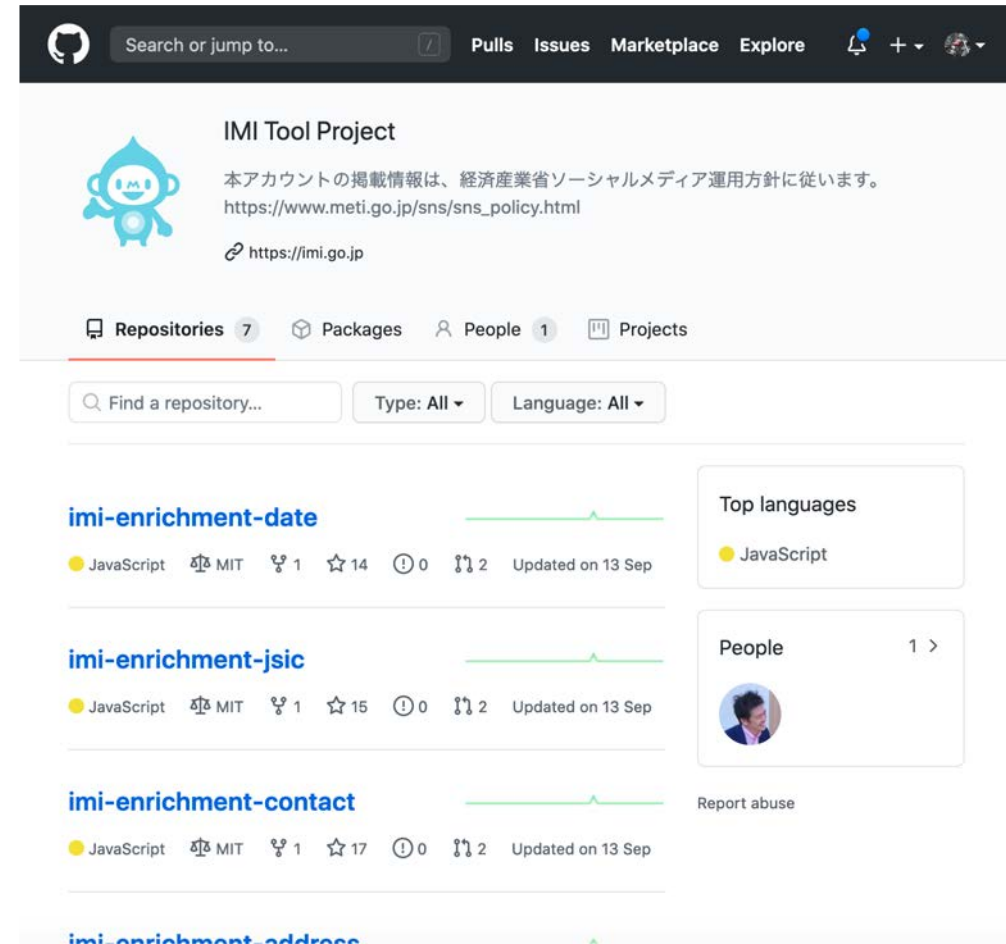


データ品質改善への工夫

④データクレンジングツールの活用

- 機械処理することで効率化を図る
- OpenRefineやIMIコンポーネント等、無償で使えたりオープンソースのツールもある
- 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注意する必要がある

<https://github.com/IMI-Tool-Project>



The screenshot shows the GitHub profile for the IMI Tool Project. The profile includes a search bar, navigation links for Pulls, Issues, Marketplace, and Explore, and a search bar for repositories. The main content area displays a list of repositories, including 'imi-enrichment-date', 'imi-enrichment-jsic', and 'imi-enrichment-contact'. Each repository entry shows the language (JavaScript), license (MIT), number of forks, stars, issues, and pull requests, along with the last update date (13 Sep). A sidebar on the right shows 'Top languages' (JavaScript) and 'People' (1).

Mapping the World of Data Problems

バッドデータ ハンドブック

データにまつわる
問題への19の処方箋



O'REILLY®
オライリー・ジャパン

Q. Ethan McCallum 著
磯 蘭水 監訳
笹井 崇司 訳

データ品質改善における 心得

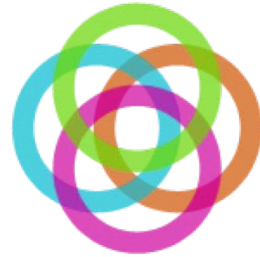
Fail early, fail often

(早く失敗しよう、たくさん失敗しよう)

後の工程に進むほど検証は複雑になるので、
できる限り早期に、たくさん洗い出しておくことが重要

<https://www.oreilly.co.jp/books/9784873116402/>

Link Data Now!!



LinkData.org

本資料に関するご質問・ご意見や、
データ活用研修・ワークショップ開催のご相談などは
こちらへお寄せください。

一般社団法人リンクデータ
代表理事 下山 紗代子
Email: support@linkdata.org