D4
研究データの品質表示を考える

企業データの観点から見た
データ品質と保証についての必要性

研究データ利活用協議会（RDUF)
データ共有・公開制度検討部会
委員 岡山将也

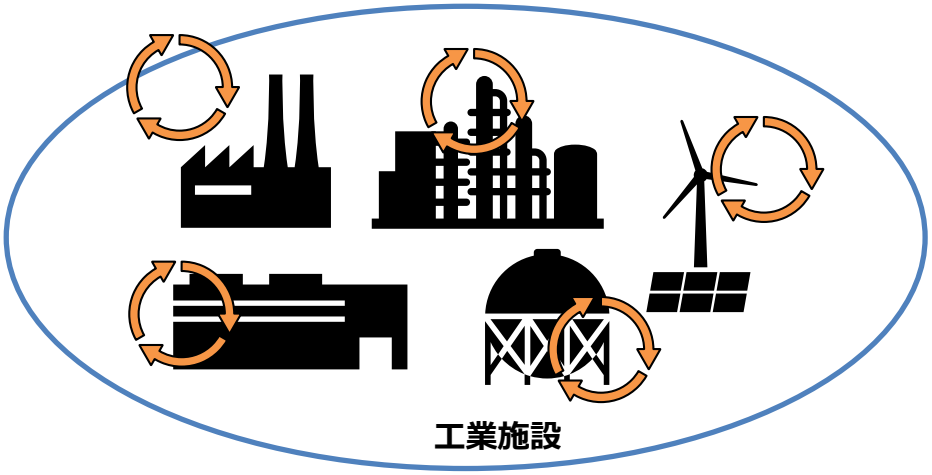
本セッションの概要

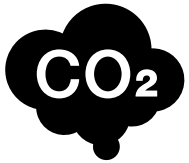
■ テーマ
企業側から見たデータのあり方についての一考察として、
どのようなデータが必要かという観点ではなく、

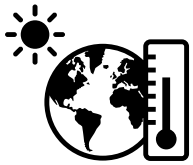
“企業が利用するデータには何が必要か”

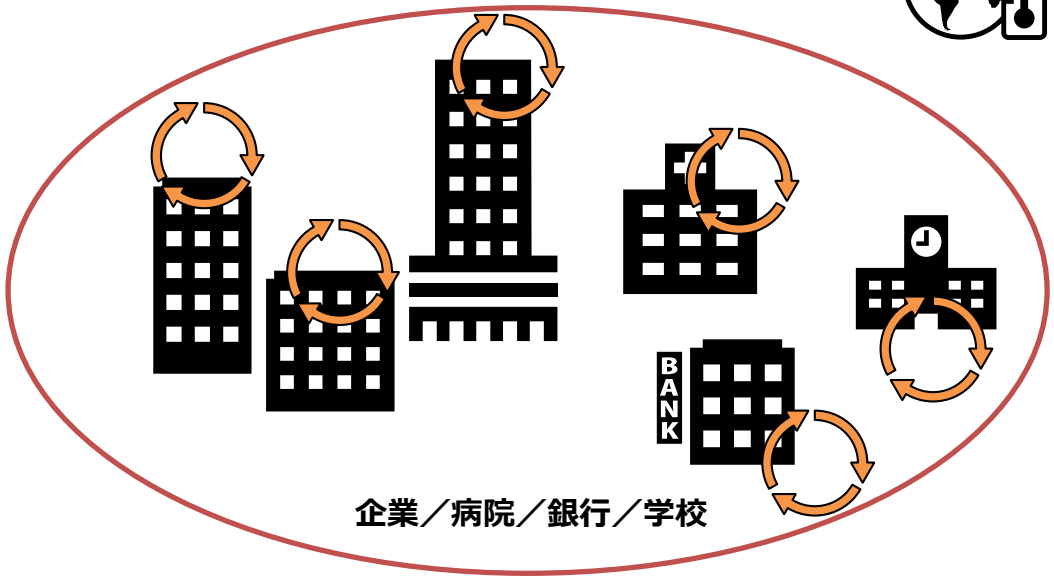
について考えたいと思います。

これまでの(企業の)データとは

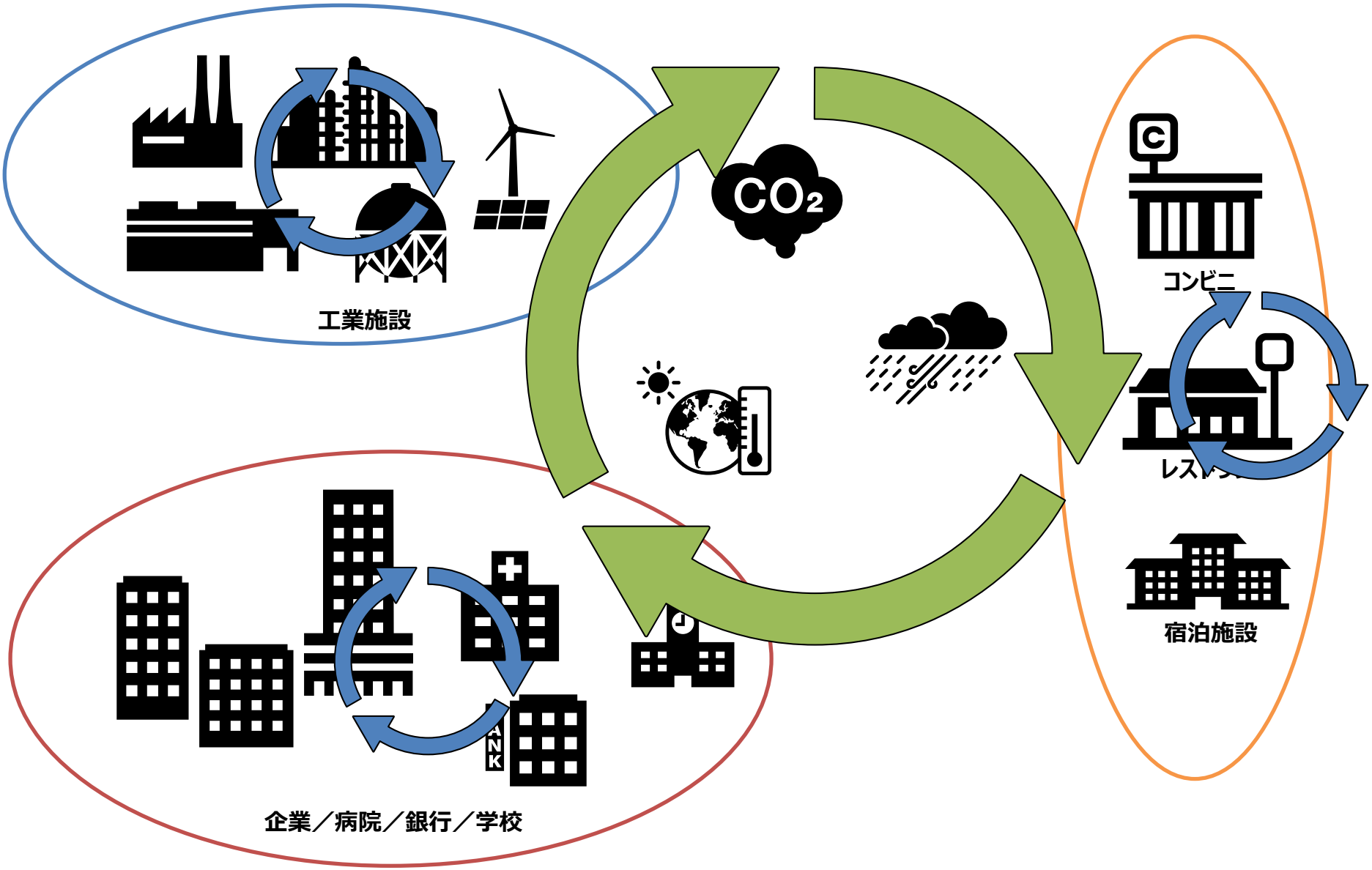






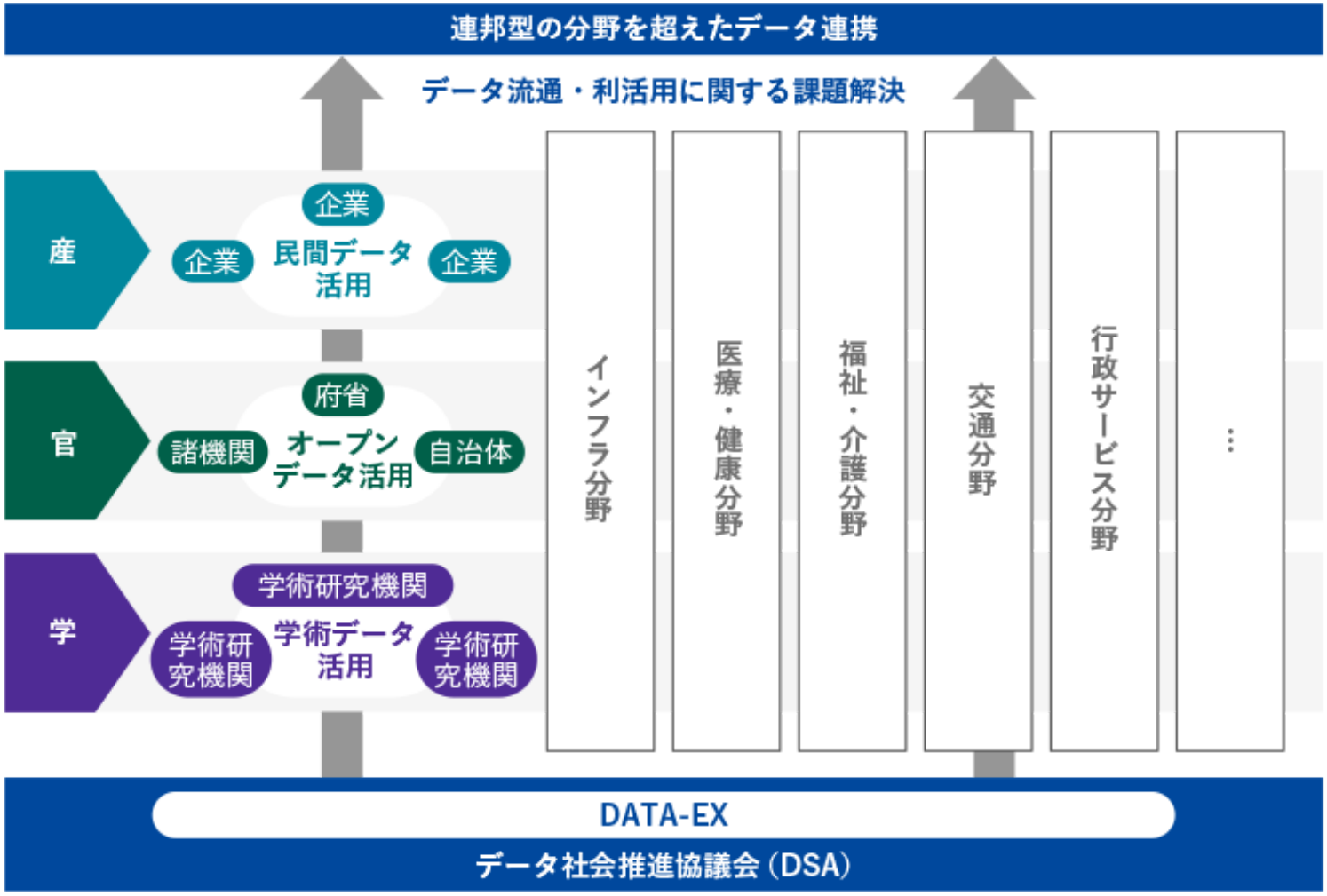


これからの(企業の)データとは



データ社会推進協議会の発足

データ社会推進協議会（Data Society Alliance：DSA）は、あらゆる分野におけるデータ流通・利活用の課題を産官学を越えた企業・団体の連携により解決する取組みを推進



DSAの取り組み方針より引用　https://data-society-alliance.org/about/vision-mission/

RESASとV-RESAS

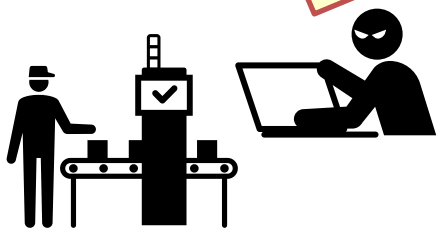
RESAS:地方創生の様々な取り組みを情報面から支援するために、経済産業省と内閣官房デジタル田園都市国家構想実現会議事務局がデータを提供する地域経済分析システム

官民のデータを利用して
地方創生をサポート

https://resas.go.jp/ より引用

データ品質と保証に必要なものは

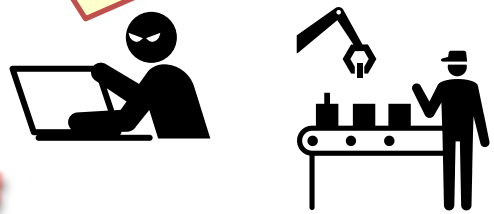
検査証明書のデータの書き換え等を
行い、製品仕様に適合していない
製品を適合していると偽って出荷した。
過去10年にさかのぼった調査でも
一部で品質データの改ざんがあった。



データ不正

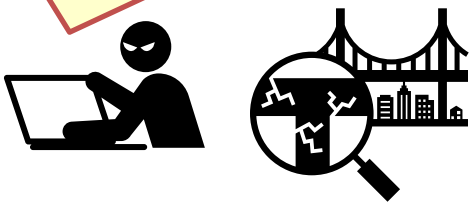
製造、販売した装置で、虚偽の
データを記載した検査成績書を
作成するなどの不正を行った。

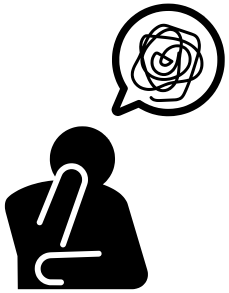
検査不正



法律違反

某製造販売業者がJISに反する
ものを製造し、この製品を利用して
建てられた物件が、違法建築に
なる恐れがあるとして、各自治体が
調査に乗り出した。





社会的信頼の失墜

データ品質と保証に必要なものは

文部科学省の研究活動における不正行為への対応等に関するガイドラインによると、対象とする不正行為は、故意又は研究者としてわきまえるべき基本的な注意義務を著しく怠ったことによる、投稿論文など発表された研究成果の中に示されたデータや調査結果等の捏造、改ざん及び盗用である（「特定不正行為」という）。

【捏　造】 存在しないデータ、研究結果等を作成すること。

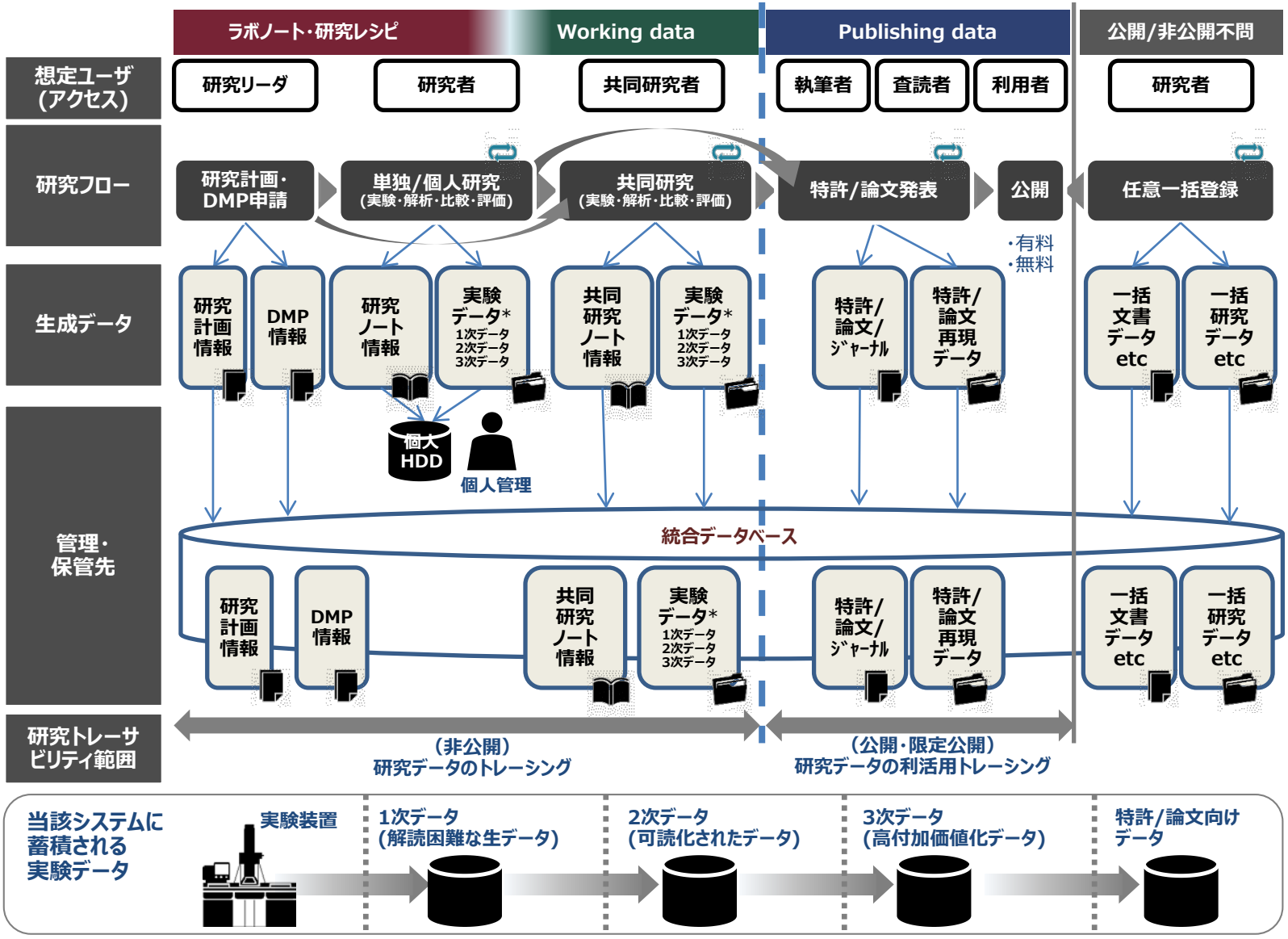
【改ざん】 研究資料・機器・過程を変更する操作を行い、データ、研究活動によって得られた結果等を真正でないものに加工すること。

【盗　用】 他の研究者のアイディア、分析・解析方法、データ、研究結果、論文又は用語を当該研究者の了解又は適切な表示なく流用すること。

研究者、技術者、社会人としての倫理規範

研究職、技術専門職の行動規範

R&Dを含めたデータ管理フローとデータの公開までの品質管理



研究倫理／研究公正の観点　　**データ管理／データ品質保証の観点**

データ自体の品質を確保するためには

- データ品質管理ガイドブックでは、データ自体の品質を考えるための指標として、ISO/IEC25012*に着目
- ISO/IEC25012では、データ品質の指標として15つの指標がある
- データそのものの品質の特性を持つ指標は、以下の５つであり、データの品質をチェックするには、まずはこの5点に着目する必要がある。

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)

また、データを流通させるためには、残り10指標のうち、すくなくとも、
アクセシビリティ、理解性、可用性、移植性の４つが必要と考える。

＊ソフトウェア製品の品質要求及び評価（SQuaRE）ーデータ品質モデル

ISO/IEC25012

#	データ品質の指標	評価項目	問題となる例
1	**正確性 (Accuracy)**	データの基本は正確であること。データの正しさは、以下の点に着目して評価する。 ● 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。	● 同上、〃などの記述がある。 ● 日付や数字が記述されるべき欄に「不明」など数字以外の文字列が記述されている。 ● 住所が記述されるべき欄に電話番号が記述されている。 ● フリガナ欄にカタカナとひらがなが混在している。
2	**完全性 (Completeness)**	データは目的に応じて抜け漏れなく存在することで、詳細な分析をすることができるようになる。データが完全であることを以下の点に着目して評価する。 ● 用途に応じて必要な項目が網羅されているか。 ● 必須項目に空欄が含まれていないか。	● 重要なデータ項目が定義されていない。 ● データが取得できないという理由で必須項目に空欄がある。
3	**一貫性 (Consistency)**	データには整合性や一貫性が必要で、データ内の項目や値に矛盾があるとエラー処理をする必要がある。データに矛盾がないことを以下の点に着目して評価する。 ● データセット内でデータに矛盾はないか。 ● データセット間でデータに矛盾はないか。	● 住所コードと住所が一致しない。 ● 外部参照に間違いがある。 ● 各項目の個別の値を集計した合計値と、元々データに含まれていた合計値が一致しない。
4	**信憑性 (Credibility)**	データは意思決定に使われることも多く、信頼できるデータである必要がある。データの信ぴょう性について以下の点に着目して評価する。 ● データの出所が明示されているか。 ● データの更新日が明示されているか。 ● 改ざん防止策が施してあるか。	● 特定のデータ作成者やデータ計測機器によるデータの誤りが複数発見された。 ● データがいつ作成されたものか分からない。 ● データの出典や収集方法が不明

ISO/IEC25012（つづき）

#	データ品質の指標	評価項目	問題となる例
5	**最新性 (Currentness)**	データが最新のものに更新されていることで、誤処理や再処理を行う必要がなくなる。データが十分に新しいものに維持されていることを以下の点に着目して評価する。 ● 公開データの更新サイクルは元データの更新サイクルに対して適切か。 ● データは収集時から十分に短い期間で公開されているか。 ● ファイル等で提供される場合は、最終更新日時及び最新版の所在が明記されているなど、更新版の有無が確認できるようになっているか。	● データが更新されていない。データが古くなってしまった場合は、データの公開を終了することを検討する必要がある。 ● 各年で取りまとめを行っているデータの公開に半年近くかかる。 ● ダウンロードしたファイルの更新版の有無が確認できない。 ● 最新のデータにおいて住所だけ古いまま掲載されている(例：東京市)。
6	**アクセシビリティ (Accessibility)**	データを受け取った人がそのデータを活用できるようにする必要がある。ります。データが誰でも使用できるものになっているかを以下の点に着目して評価します。 ● ファイルで提供している場合、データの使用権を持つ全ての人が利用できるようになっているか。 ● ソフトウェアを通して提供している場合、そのソフトウェアはISO/IEC40500 に準拠しているか。 ● 使用している文字セット（常用漢字など）は正しいか。	● 特殊なファイル形式で公開されている。 ● 常用漢字が定められているにも関わらず、それ以外の漢字がフリガナを伴わず使用されている。 ● 環境依存の文字が使用されている。
7	**標準適合性 (Compliance)**	データは入力ルールなどの一定のルールにより管理されており、そのルールを守ることで円滑に処理をすることができる。データが標準に適合しているかを以下の点に着目して評価する。 ● データの書式は標準に準拠しているか。 ● 使用している文字セットは正しいか。 ● 選択項目に、指定された選択肢以外のデータが入っていないか。	● 年月日が西暦ではなく和暦で表記されている（例：R2.4.1）。 ● 環境依存の文字やユーザー定義文字が使用されている。 ● 都道府県名が略称で表記されている（例：「東京都」と表記すべきところを「東京」と表記）。

ISO/IEC25012（つづき）

#	データ品質の指標	評価項目	問題となる例
8	**機密性 (Confidentiality)**	データによっては機密情報を含むものもあり、目的に応じた機密性が確保される必要がある。データの機密性について以下の点に着目して評価する。 ● データにアクセスできるのは、アクセスを許可された者に限定されているか。 ● 利用者を制限する場合、暗号化やハッキング対策などが行われているか。	● データを提供しているソフトウェアに脆弱性がある。 ● データ管理ツールにおいて共有範囲が誤って設定されている。
9	**効率性 (Efficiency)**	データは効率的に処理される必要があり、そのためにコードを割り当てる 等の対応をします。データの効率性について以下の点に着目して評価します。 ● データの内容に重複などがないか。 ● データは効率的に処理できるようになっているか。 ● コードを効果的に使用しているか。 ● データに一貫性はあるか。	● データに全角と半角が混在するなど、データとデータを結び付ける際に正規化が必要となる。 ● 住所とビル名が別データ項目になっていないなど、データ活用するために分離処理が必要になる。 ● 表計算ソフトで作成されたデータに余分な罫線やスペースが入っている。 ● 他のデータと結合しやすくするための ID やコードが入っていない。
10	**精度 (Precision)**	データには使用目的に応じて必要な精度がある。また精度の違うデータを一体として扱う時に精度の調整が必要になる。データの精度について以下の点に着目して評価する。 ● データの精度は適正に設定されているか。 ● データの精度がそろっているか。 ● データの精度が示されているか。	● 各データが、小数点以下切捨て、小数点以下２桁まで記録など、精度にばらつきがあり、単純に加算できない。 ● 許容誤差範囲が異なるデータが混在している。 ● 正確な位置を特定する必要のあるデータにおいて、緯度経度の値が粗すぎる。
11	**追跡可能性 (Traceability)**	データに疑義が生じたりした時に、データの原典などを参照する場合がある。データの追跡可能性について以下の点に着目して評価する。 ● 外部データが明確になっているか。 ● データ変更の際に、変更者、変更日などを記録しているか。	● 外部データの出所が明確になっていない。 ● いつ、誰が変更したかが分からない。 ● データの変更箇所や変更方法（例：機械処理なのか人手なのか）が不明

ISO/IEC25012（つづき）

#	データ品質の指標	評価項目	問題となる例
12	**理解性 (Understandability)**	データを活用する時には、データの項目を正しく理解して活用する必要がある。利用者がデータについて理解できるかについて以下の点に着目して評価する。 ● データ全体及びその各項目が意味するものを利用者が理解できるようになっているか ● データ全体や必要に応じてその各項目にメタデータが提供されているか。 ● 共通語彙基盤1のような意味を定めたものに関連付けがされているか。	● データの説明がなく、データが意味するものを正確に理解できない。 ● 住所が本店所在地なのか、事業所所在地なのか判断できない。 ● 記述されているコードや略称の意味する内容が不明
13	**可用性 (Availability)**	データは必要な時に使えるようになっている必要がある。データが利用可能な状態になっているかを以下の点に着目して評価する。 ● 必要な時にいつでもデータにアクセスできるようになっているか。 ● データを公開するシステムは常時稼働しているか。	● 頻繁にシステムが停止する。 ● データ公開システムにアクセス可能な時間帯が限定されている。
14	**移植性 (Portability)**	システムの入替えやシステム間の連携を行う際には、データを簡易に移行できる必要がある。データの移植のしやすさについて以下の点に着目して評価する。 ● 標準的なフォーマットで出力できないソフトウェアに依存していないか。 ● データを管理するシステムから標準的な形式によりデータをエクスポートすることができるか。	● ソフトウェア固有のフォーマットでしか出力できない。 ● PDF や画像データであるため再利用できない。 ● システムからデータをエクスポートできない。
15	**回復性 (Recoverability)**	データセンターなどで事故が起こった時に、そのデータが早急に復元され る必要がある。データの回復性について以下の点に着目して評価する。 ● データのバックアップが保存されているか。 ● システム障害が発生した場合であっても、継続してデータを提供するバックアップシステムが存在するか。	● バックアップされないディスク上にのみデータが保存されている。 ● 特定のサーバーのみからしかデータが取得できない。

The FAIR principles（FAIR原則）とは

国際的な学術コミュニケーションのための国際会議 FORCE11が、策定・公開する研究データの公開時に関するガイドラインです。

FAIRとは、

Findable　　：見つけられる
Accessible　　：アクセスできる
Interoperable　：相互運用できる
Reusable　　：再利用できる

のそれぞれの頭文字を取った略語です。

データ公開の適切な実施方法を表現しており、データ共有の原則として広まっています。
欧米をはじめ、日本でも、いくつかの研究資金を提供する機関が、研究データを扱う際の基準として「FAIR原則に従うこと」をガイドラインで推奨しています。

参考：https://www.force11.org/group/fairgroup/fairprinciples
https://www.nature.com/articles/sdata201618

The FAIR principles （その1）

分類	#	英文	和訳
Findable	F1	(meta)data are assigned a globally unique and eternally persistent identifier.	(メタ)データには、グローバルでユニークな永続的識別子が割り当てられていること
	F2	data are described with rich metadata. (defined by R1 below)	豊富なメタデータによってデータが説明されていること(以下R1に定義)
	F3	metadata clearly and explicitly include the identifier of the data it describes	メタデータには、そのメタデータが説明するデータの識別子を明示的に含むこと
	F4	(meta)data are registered or indexed in a searchable resource	検索可能なリソースに(メタ)データが登録または索引付けされている
Accessible	A1	(meta)data are retrievable by their identifier using a standardized communications protocol	(メタ)データは、標準的な通信プロトコルを使った識別子にて再取出/回収可能なこと
	A1.1	the protocol is open, free, and universally implementable	プロトコルはオープン、フリー、かつ普遍的に実装可能であること
	A1.2	the protocol allows for an authentication and authorization procedure, where necessary	プロトコルは、必要に応じて、認証および認可手続が可能なこと
	A2	metadata are accessible, even when the data are no longer available.	データが利用できない場合でも、メタデータにアクセスできること

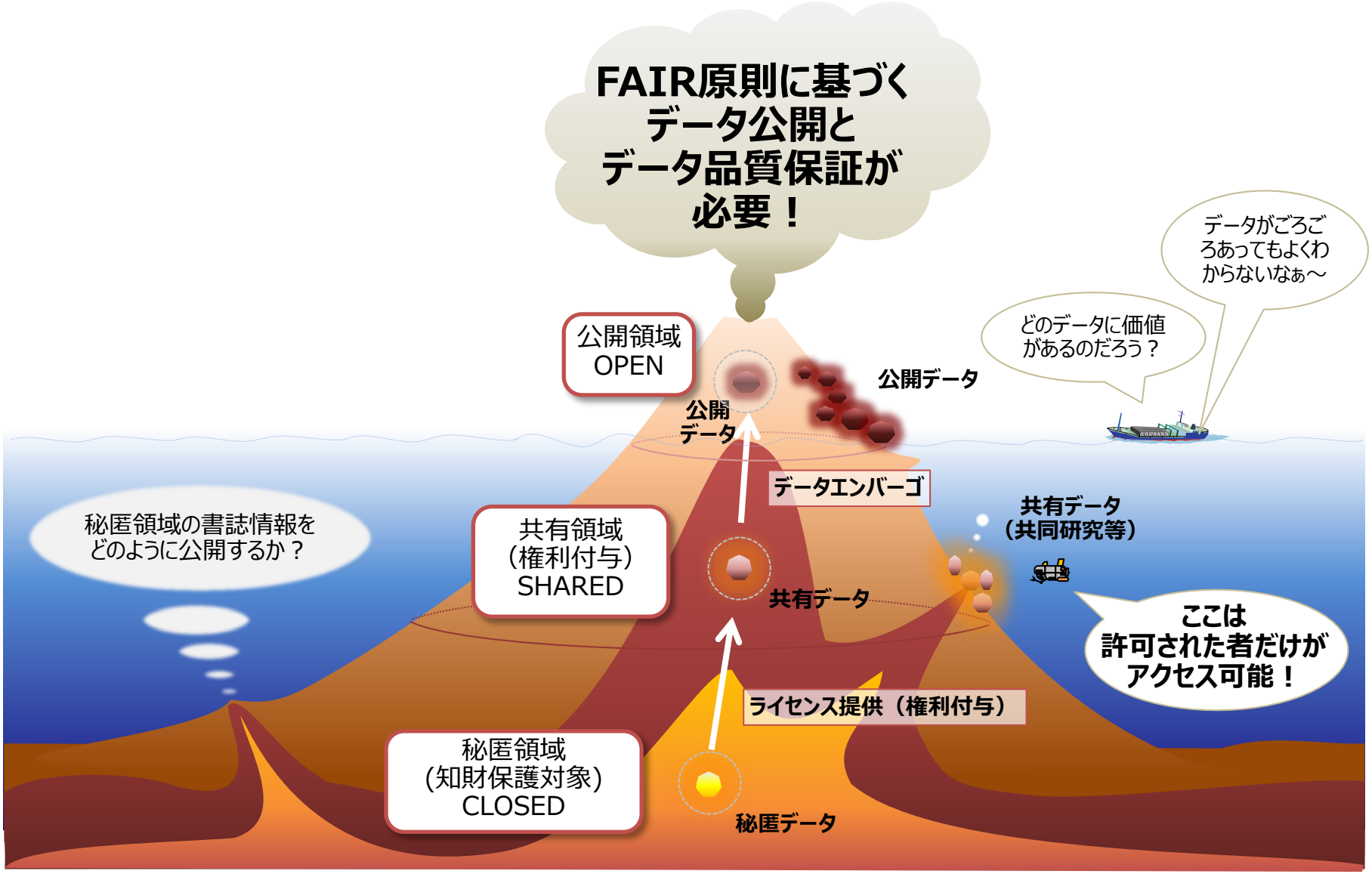
参考：https://www.force11.org/group/fairgroup/fairprinciples
https://www.nature.com/articles/sdata201618

The FAIR principles （その2）

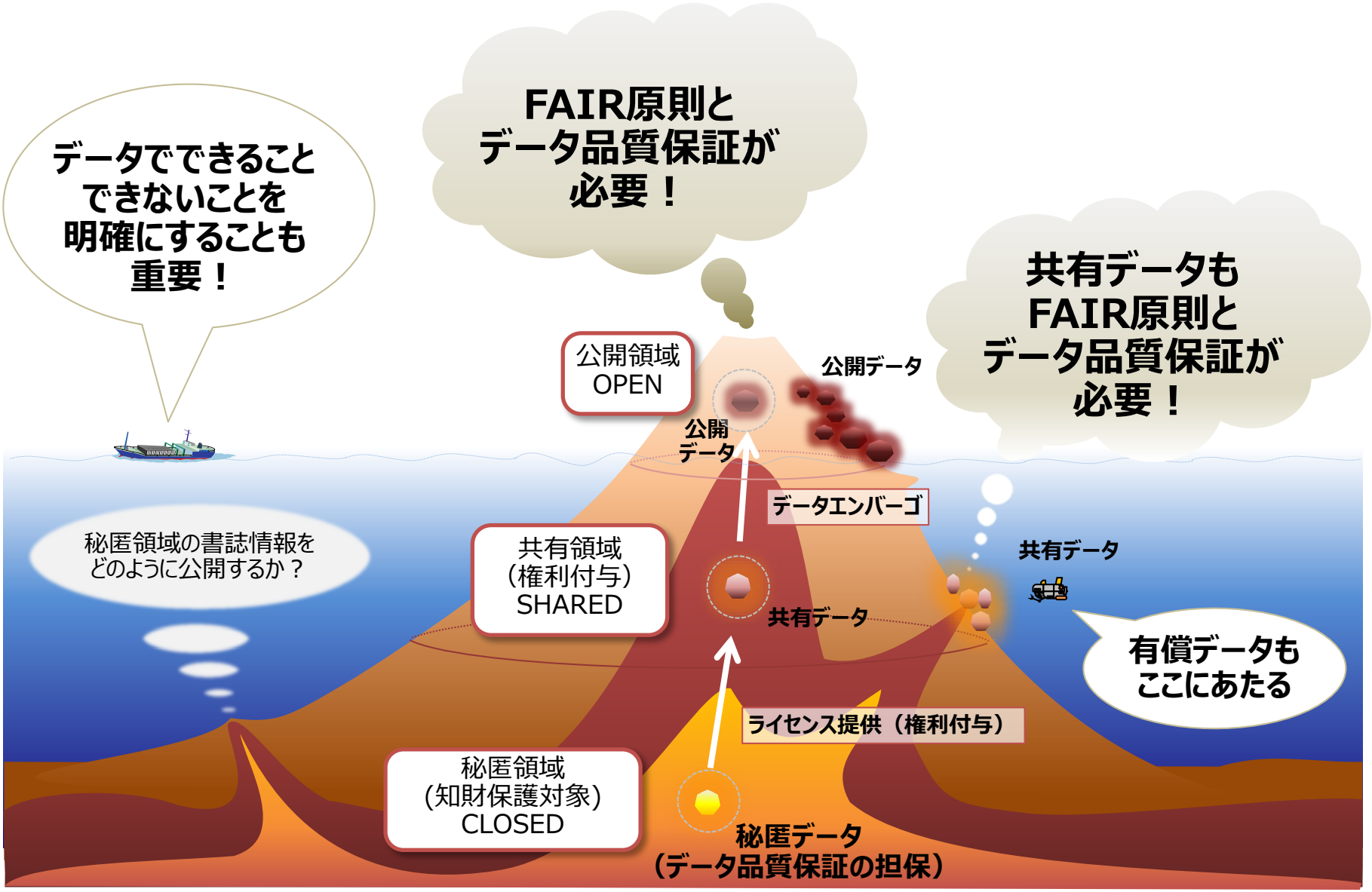
分類	#	英文	和訳
Interoperable	I1	(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation	(メタ)データでは、知識表現に広く適した正式かつ、利用・共有しやすい言語を使用されていること
	I2	(meta)data use vocabularies that follow FAIR principles	(メタ)データでは、FAIRの原則に従う語彙が使用されていること
	I3	(meta)data include qualified references to other (meta)data	(メタ)データは、他の(メタ)データへの適正な参照・出典を含むこと
Re-usable	R1	(meta)data are richly described with a plurality of accurate and relevant attributes	(メタ)データは、正確かつ関連する複数の属性で十分に説明されていること
	R1.1	(meta)data are released with a clear and accessible data usage license	(メタ)データは、明確で分かり易いデータ利用ライセンスで公開されること
	R1.2	(meta)data are associated with detailed provenance	(メタ)データは、詳細な来歴/出典と関連付けされていること
	R1.3	(meta)data meet domain-relevant community standards	(メタ)データは、当該領域/分野関連のコミュニティの基準を満たしていること

参考：https://www.force11.org/group/fairgroup/fairprinciples
https://www.nature.com/articles/sdata201618

オープンサイエンスにおけるデータ公開までの遷移のイメージ

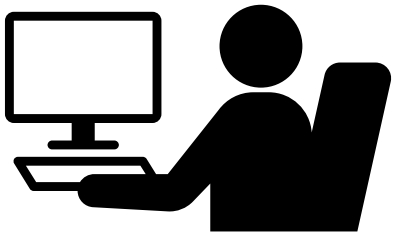


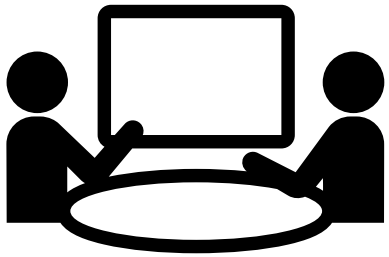
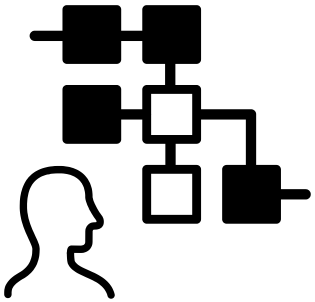
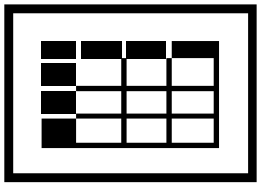
企業データの公開も同じでは



研究も含めたデータ管理の視点でのデータ品質に関する視点（1）

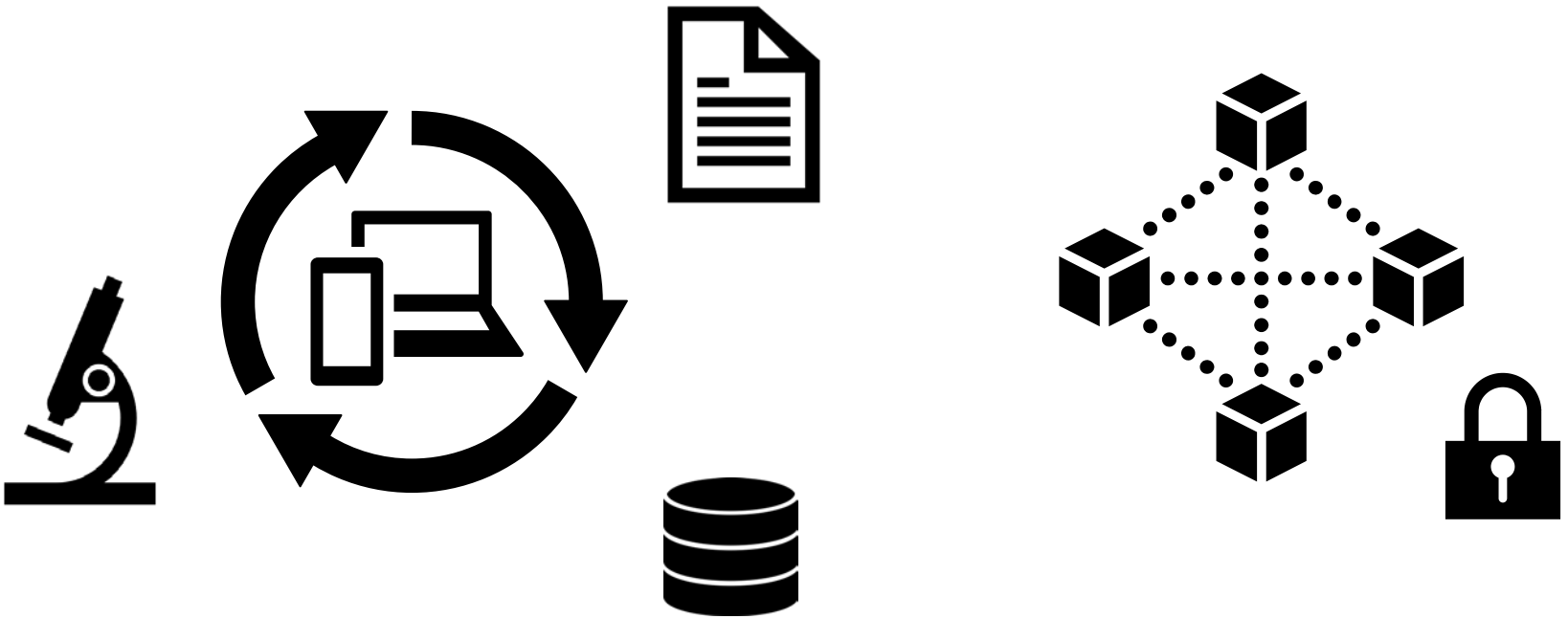
- **データ管理プロセス視点**
  - ✓ データ管理計画(DMP:Data Management Plan)の作成や承認
  - ✓ システムがデータ管理計画をどのようにサポートしているか
  - ✓ 内部/外部監査時の提示（システム上、監査時に特定データを抽出できることが求められるか）



研究も含めたデータ管理の視点でのデータ品質に関する視点（2）

- **トレーサビリティにおける品質保証視点**
  - ✓ 実験装置データの記録（実験装置PIDも含む）
  - ✓ 秘匿データから限定共有、公開に至るまでの経緯
  - ✓ 共有フェーズにおける参照、利用、加工、更新等の記録
  - ✓ 公開において研究を参照、再利用、新しいデータを作り出した記録
  - ✓ オープンデータ、共有データに付与されたPIDの有無



研究も含めたデータ管理の視点でのデータ品質に関する視点（3）

- **データキュレーションによる品質保証視点**
  - ✓ データ作成者（データオーナー）によるデータ形式のチェック
  - ✓ データキュレーターによる査読や内部矛盾のチェック
  - ✓ データレジトリーによる、過去の実績との矛盾、再現性のチェックなど
  - ✓ このデータでできること、できないことを明確化（免責事項を記載）

研究も含めたデータ管理の視点でのデータ品質に関する視点（4）

- **改ざんされないデータ記録（改ざん検知）の視点**
  - ✓ チェックサムによるデータ本体の改ざん検知
  - ✓ チェックサムに対するデジタル署名によるチェックサム自体の改ざん防止
  - ✓ 商用時刻認証局タイムスタンプ付きデジタル署名によるチェックサムの時刻認証
  - ✓ 長期署名による電子証明書の有効期限切れ対策のための証明書更新

研究も含めたデータ管理の視点でのデータ品質に関する視点（5）

- **データをどのように利用したいかの視点**
  - ✓ データの関連ツリー構造の可視化
  - ✓ データ生成者自身による過去のデータの参照、取出し、更新
  - ✓ データトレース（どのデータから派生したかなど）の可視化
  - ✓ データ管理者による業務全体の状況分析
  - ✓ 監査要件としてのデータの存在確認、改ざんの有無、再現性の有無の確認