データ連携基盤サブワーキンググループ（第2回）

資料1

データ連携基盤の技術的検討事項

平成30年3月1日（木）

内閣府
政策統括官（科学技術・イノベーション担当）

1. 主な検討課題

方針

- 複数分野のデータを組合せ、付加価値の高いアプリケーションの創出
- 各分野プロジェクトで類似機能の重複整備を極力防止
- 提供データの品質、相互運用性の向上

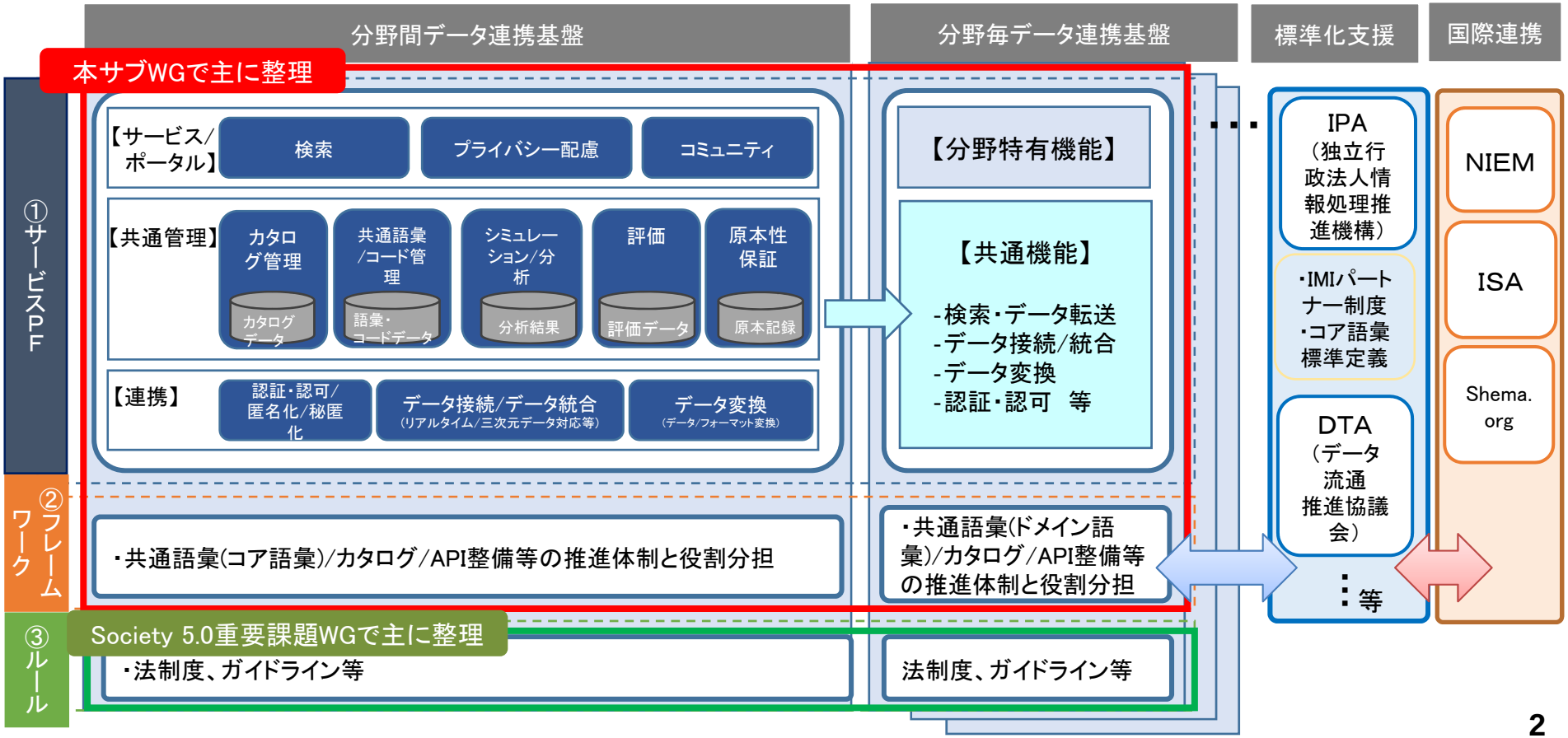
例：データの機械可読性向上（PDF形式の廃止等）、データ毎の信頼度、粒度、取得頻度、語彙等の差異に起因するデータ活用の困難性を取り除く等

検討課題

- プラットフォームに実装する機能について、協調領域・競争領域の設定
- 持続的な運営が可能となる体制構築と機能の検討
- 国際間連携も意識し、相互運用性を確保するための共通語彙、コード、カタログ、API等の整備
- 分野横断サービス/アプリケーションでの検証・評価を反映するなど、PDCAサイクルによる発展

2. データ連携基盤の構成要素

- データ連携基盤（分野間／分野毎）は、以下の3つで構成
  - ① **サービスプラットフォーム（サービスPF）**
    クラウド等によるデータ連携に必要なプログラム・サービス各機能で構成されるプラットフォーム
  - ② **フレームワーク**
    共通語彙(コア/ドメイン語彙)、カタログ、APIの定義等の推進体制や役割分担
  - ③ **ルール**
    上記を実現する上で必要な法制度、ガイドライン等



3. 主な論点

① プラットフォームに実装すべき機能の検討

- 競争領域、協調領域の設定
- 継続的に運営、発展するための機能群を実装

例）データカタログの管理
　カタログデータ（名称、作成者など）を登録、更新、共有。
例）共通語彙・コード管理
　共通語彙基盤（IMI）の語彙やコードを用いることで、異なる表記や構造をもつ複数のデータに共通の表記や構造を与えることができ、データの共有/機械処理が容易にさせる。
例）機械可読性の乏しいデータの変換
　EXCELデータにおけるセルの結合解除、キャプションの削除等、自動化支援ツールをデータ連携基盤で提供。

② データ連携で先行する欧米との連携

- 2014年からは、EU、米国、日本が参加する各国語彙（NIEM、SEMIC、IMI）の連携会議が開始。各語彙体系の相互運用性の確保を目指し、現在活動中。共通語彙に基づいた分野毎、分野間データ連携基盤の整備を推進し、G7 や G20 で世界に発信することを意識しながら、米、EUとも繋がるデータ連携基盤とする。
- 国際標準化戦略としては、デジュール標準とデファクト標準を総合的に取組むことが重要。

③ 自立的、持続的発展を担う運営体制

- 厳密な運用が必要となる政府系データ連携基盤上では、政府ガバナンスの下での運営体制が望ましい。
- 技術革新の激しい産業分野では、民間主導の運営体制が効果的なため、分野毎に対応すべき。
- 分野（ドメイン）毎に政府（各府省庁）、民間の担う役割分担し責任を明確化すべき。

④ その他

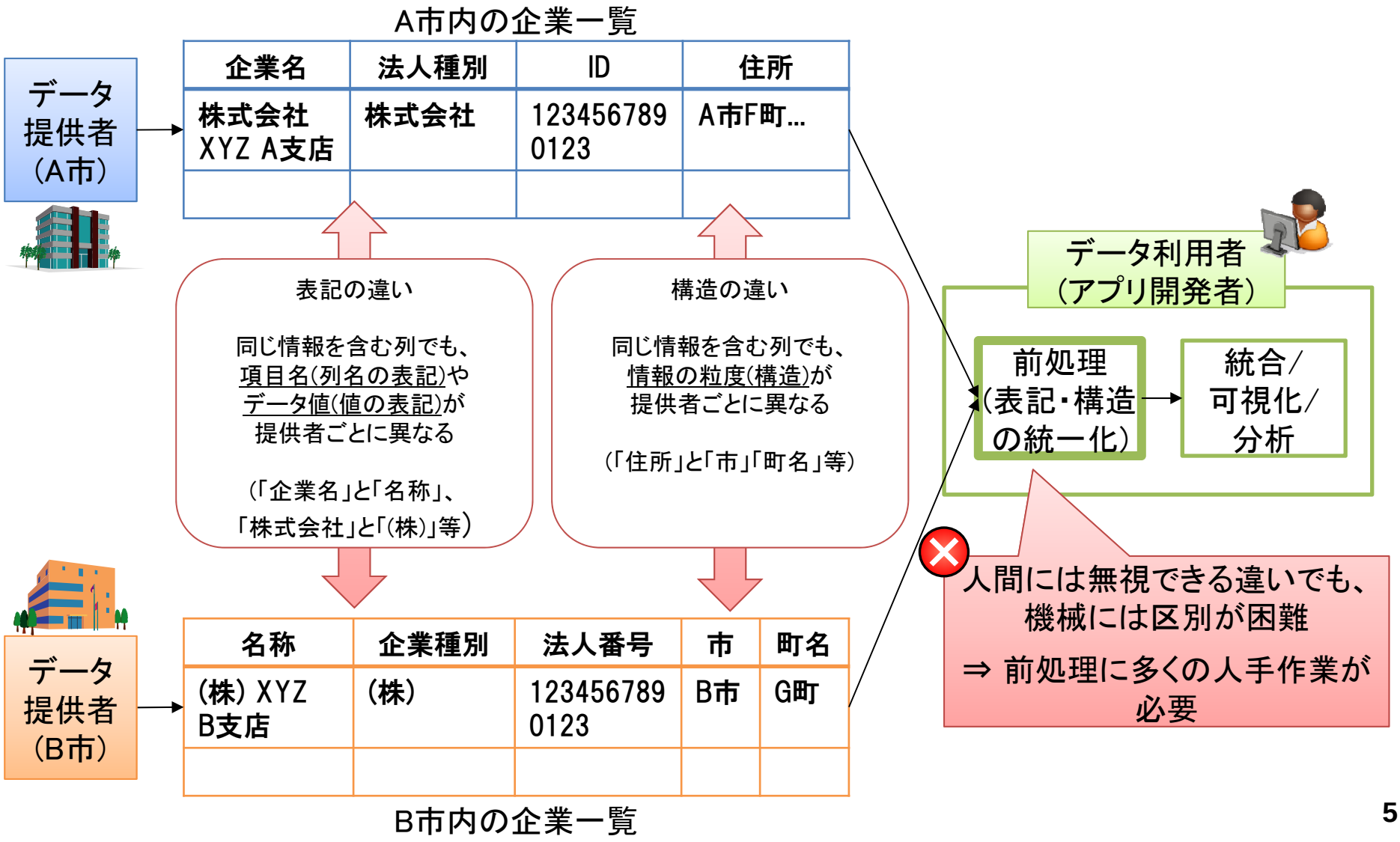
- メタデータ項目の検討（政府系オープンデータに加え、IoTデータへの対応も今後重要な課題）
- 様々な利用規約を有するデータの組み合わせ利用における権限の整理・類型化（政府標準利用規約 Creative Commons License（CC BY等）、Open Database License（ODbL等）、独自規約等）

① プラットフォームに実装すべき機能イメージ

#	区分	機能	機能概要
1	サービス/ポータル		利用者がデータを利用し易くサポートする機能
2		検索/データ統合	あいまいなキーワードにて検索し、複数分野のデータを時空等で統合。今後、増大が予想されるIoTデータへの対応も必要
3		プライバシー配慮	プライバシーにも配慮し、オプトイン・オプトアウト・データ利用目的の追加等の利用者・提供者とのやり取りを実装
4		コミュニティ	利用者同士の意見交換・イノベーション協創（共同研究等）の場を提供
5	共通管理		データ検索や連携のためにデータ連携基盤が管理する機能
6		カタログ管理	カタログデータ（名称、作成者など）を登録、更新
7		共通語彙/コード管理	データ変換にて共通項目名等に揃えるための語彙/コード情報を登録、保管
8		シミュレーション/分析	分野共通のシミュレーションや相関等分析（EBPM向け等を含む）
9		法人・データ評価	法人・データの品質・実績等をランク付け。認証と連携しアクセスを制御
10		原本性保証	各分野で発生するデータの原本性を保証。データ流通品質を担保
11	連携		データ利用者の要求に応じて、データ提供者のデータを応答する機能
12		認証・認可/匿名化/秘匿化	提供者、利用者の認証、データ匿名化、暗号化等
13		データ接続	データ提供者、利用者との接続を行う機能(センサデータ等のリアルタイム接続、三次元データ接続も含む)
14		データ変換	データの単位、座標系、項目名等及びデータフォーマットを揃える機能

（参考）共通のデータ構造、語彙を用いたデータ変換

- 現状、表記や構造が異なる複数のデータを統合/可視化/分析する際には、データ利用者は、それらの表記や構造を統一するためのデータ修正に多くの時間/手間を要している。



（参考）データ構造、語彙、コードの整備によって得られる効果

- 共通語彙基盤の語彙やコードを用いることで、異なる表記や構造をもつ複数のデータに共通の表記や構造を与えることができ、データの共有/機械処理が容易になる。

データ提供者（A市）

A市内の企業一覧

企業名	法人種別	ID	住所
株式会社 XYZ A支店	株式会社 [301（国税庁コード）]	1234567890123	A市F町…

語彙による表記・構造の共通化

法人

名称

ID

組織種別

住所

表記

識別値

識別値

表記

名称

ID

コード

住所

表記

識別値

識別値

市区町村

町名

データ提供者（B市）

名称	企業種別	法人番号	市	町名
（株）XYZ B支店	（株） [301（国税庁コード）]	1234567890123	B市	G町

B市内の企業一覧

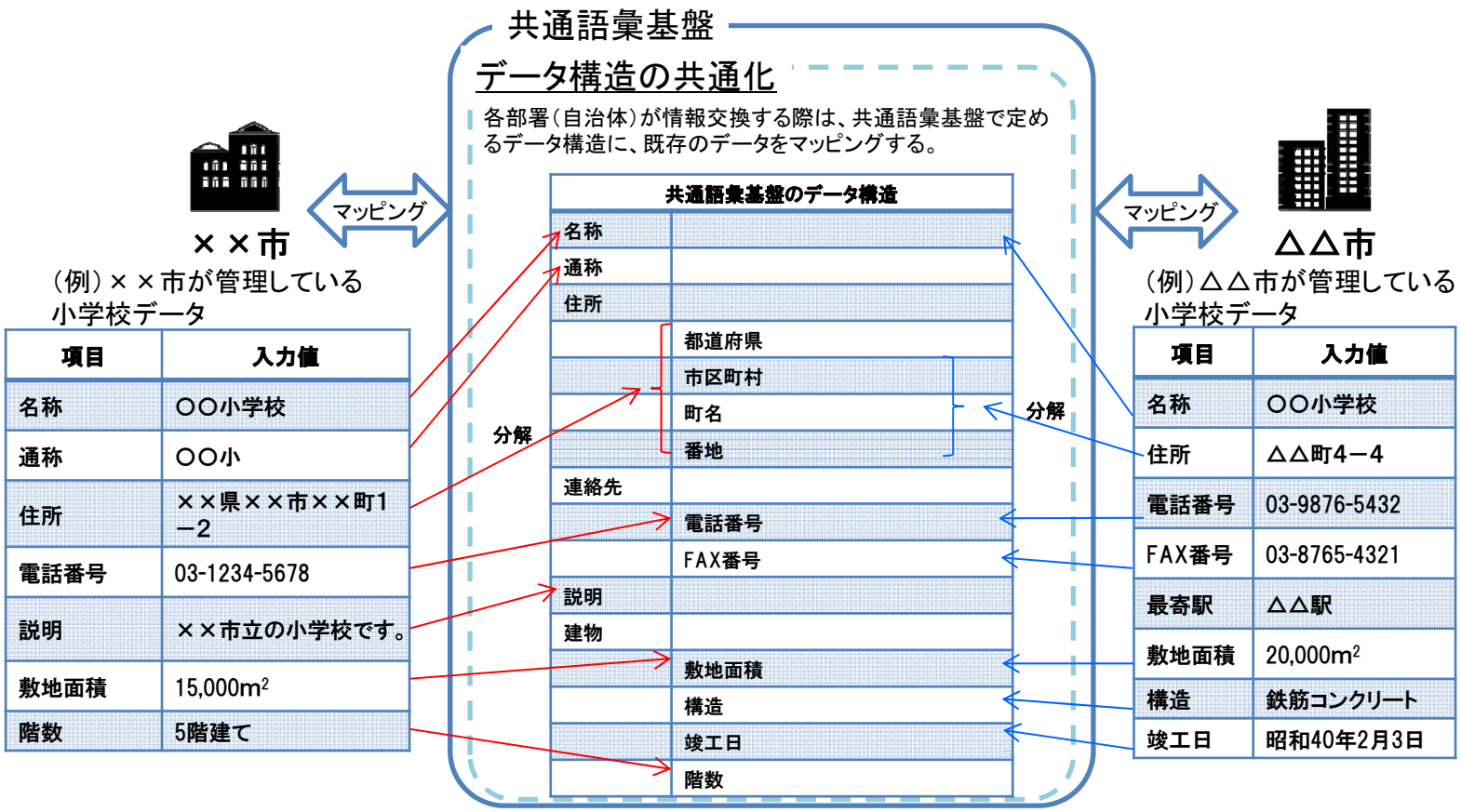
データ利用者（アプリ開発者）

前処理（表記・構造の統一化）→ 統合/可視化/分析

各列の意味を共通の語彙で記述することで、伝達/共有/機械処理が容易に

（参考）共通語彙基盤によるデータ構造の共通化イメージ

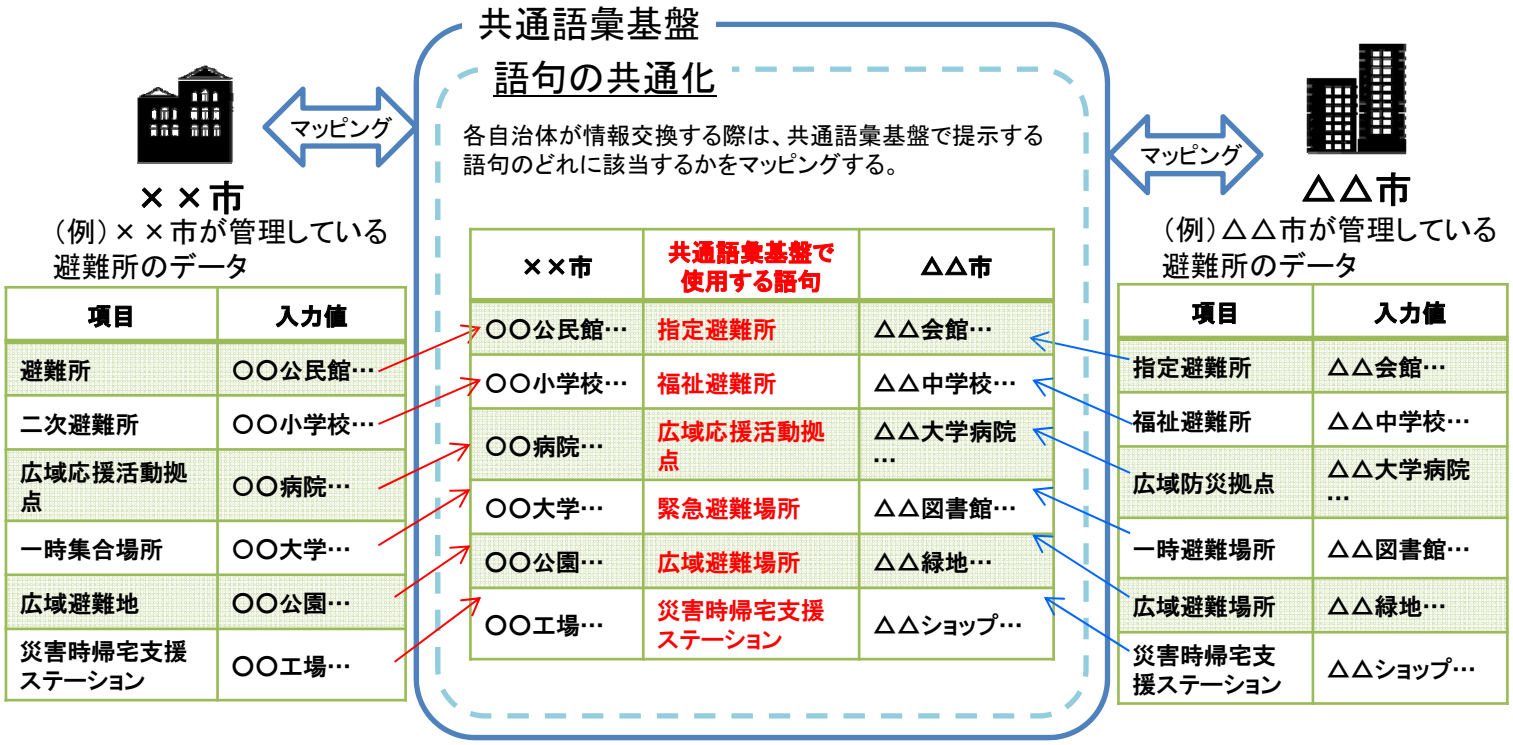
- 共通語彙基盤により、組織間の情報交換の時にデータの構造・形式の違いを埋めるイメージ



※1）経済産業省（平成26年9月）「自治体が保有する情報の可能性～情報が利活用しやすい環境の整備～」より
http://www.kantei.go.jp/jp/singi/it2/densi/jichibukai/dai2/siryou4.ppt

（参考）共通語彙基盤による語彙（語句）の共通化イメージ

- 共通語彙基盤により、組織間で異なる語句を用いている場合における情報をつなぐイメージ



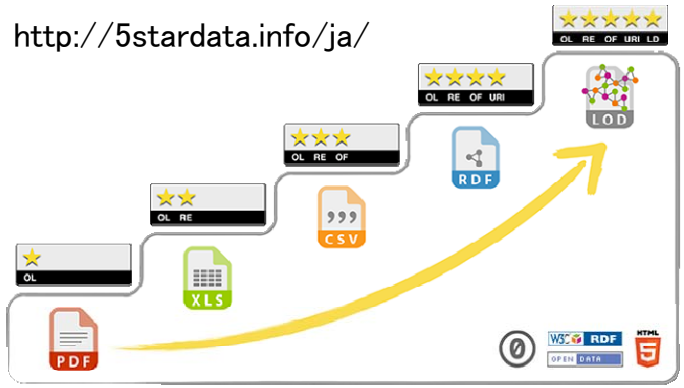
共通語彙基盤は自治体ごとに異なるデータ構造及び語句の使い方（意味）を吸収し、
既存のシステム等を変えることなく、情報交換が可能となる。

※1）経済産業省（平成26年9月）「自治体が保有する情報の可能性～情報が利活用しやすい環境の整備～」より
http://www.kantei.go.jp/jp/singi/it2/densi/jichibukai/dai2/siryou4.ppt

（参考）データの機械可読性の乏しいデータへの対応

- オープンデータに対する機械可読性については、公開度を示す指標である「5 Star Open Data」が提唱されており、それぞれの段階における機械可読性に対するその課題を示す。

http://5stardata.info/ja/



※その他、共通的に考慮が必要な項目（推奨）
①文字コードをUTF-8にする
②機種依存文字の置換
③外字置換（縮退、文字情報基盤の活用）
④データ値の標準化（日付けの記法等）
⑤標準コードの利用
⑥位置情報（緯度経度）の追加

機械判読が乏しいデータに対するデータ化支援ツールが必要な範囲

段階	公開の状態	データ形式	次の段階へ進むために必要な作業	
1段階 ★	オープンライセンスでデータを公開	PDF、JPG	①OCRや人力でデータを抽出	②Excelなどにて表形式データに出力
2段階 ★★	コンピュータで処理可能なデータを公開	XLS、DOC	①複数表を分割 ②整形用文字の削除（空白・改行・カンマ等） ③キャプション、脚注、脚注番号の削除 ④セル結合の解除	⑤省略されたセルの補完 ⑥ヘッダーを1行にまとめる ⑦単位の明確化と分離 ⑧CSVデータとして出力
3段階 ★★★	オープンに利用できるフォーマットでデータを公開	CSV、XML	①データの識別子としてURIを使用 ②データ項目の語彙対応 ③RDFデータとして出力	
4段階 ★★★★	Web標準（RDF等）のフォーマットでデータを公開	RDF	①識別子にHTTP URIを使用し、当該URIにて詳細情報提供	②他のデータのURIへのリンクの追加
5段階 ★★★★★	他へのリンクを入れたデータ（LOD）を公開	Linked-RDF	ー	

低 機械可読性 高

語彙対応によるデータ連携機能が必要な範囲

（参考）国内のデータの機械可読性の現状

- 政府オープンデータにおけるそれぞれの公開度を、5 Star Open Data（*2）の示す指標に従い、その割合を示す。

各データ形式の公開されている割合

段階	オープンデータの公開の状態	データ形式	DATA.GO.JPで公開されている割合（2017/11/24時点）	【統計データ】府省庁別の棚卸し結果（*1）の割合
1段階 ★	オープンライセンスでデータを公開	PDF、JPG、GIF、PNG、TIFF	約65%	約45%
2段階 ★★	コンピュータで処理可能なデータを公開	XLS、DOC	約30%	約44%
3段階 ★★★	オープンに利用できるフォーマットでデータを公開	XML、CSV	約5%	約11%
4段階 ★★★★	Web標準（RDF等）のフォーマットでデータを公開	RDF	約0%	約0%
5段階 ★★★★★	他へのリンクを入れたデータ（LOD）を公開	Linked-RDF	約0%	約0%

*1 内閣官房情報通信技術（IT）総合戦略室（2017年10月31日）「行政保有データの棚卸し結果及び官民ラウンドテーブルの開催等について」より
（https://www.kantei.go.jp/jp/singi/it2/senmon_bunka/data_ryutsuseibi/opendata_wg_dai4/siryou1.pdf）

*2 総務省（平成27年4月24日）「参考 5スターオープンデータについて」より
（http://www.soumu.go.jp/main_content/000353999.pdf）

（参考）機械可読性の乏しいデータに対するデータ化支援ツールの例

機械判読不可データを、機械判読可能なデータに変換するための処理例を以下に示す。（詳細は、参考資料①参照）

・キャプション、脚注、脚注番号の削除
表外のキャプション、脚注、及びセル内の脚注番号などを削除する。

表形式データの架空データサンプル（その１）

集計項目	平成23年（1000円）	24年（1000円）	差分１）
合計	55000	127768	a)232
ああ	1000	1100	110
いい	2000	2200	110
うう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

注：平成23年から平成24年のうちの増減の割合を記載している。
a)脚注番号のサンプルを示している。

集計項目	平成23年（1000円）	24年（1000円）	差分
合計	55000	127768	232
ああ	1000	1100	110
いい	2000	2200	110
うう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

（※1）

・セル結合の解除
セルが結合されている場合は、セル結合を解除し、全てのセルに元の値をコピーする。

表形式データの架空データサンプル（その４）

年度	期	A（mg）	B（mg）	C（mg）
2005	上	0.01	0.01	0.00
	下	0.01	0.01	0.00
2006	上	0.02	0.01	0.00
	下	0.01	0.01	0.00
2007	上	0.01	0.01	0.00
	下	0.02	0.01	0.01
2008	上	0.03	0.01	0.00
	下	0.02	0.02	0.00
2009	上	0.02	0.01	0.00
	下	0.02	0.01	0.00
2010	上	0.01	0.01	0.00
	下	0.01	0.01	0.00

表形式データの架空データサンプル（その４）

年度	期	A（mg）	B（mg）	C（mg）
2005	上	0.01	0.01	0.00
2005	下	0.01	0.01	0.00
2006	上	0.02	0.01	0.00
2006	下	0.01	0.01	0.00
2007	上	0.01	0.01	0.00
2007	下	0.02	0.01	0.01
2008	上	0.03	0.01	0.00
2008	下	0.02	0.02	0.00
2009	上	0.02	0.01	0.00
2009	下	0.02	0.01	0.00
2010	上	0.01	0.01	0.00
2010	下	0.01	0.01	0.00

（※1）

※1）各府省情報化統合責任者（CIO）連絡会議決定（平成25年6月25日）「二次利用促進のための府省のデータ公開に関する基本的考え方（ガイドライン）」より
http://www.kantei.go.jp/jp/singi/it2/cio/dai52/kihon.pdf

② データ連携で先行する欧米との連携

- 欧州や米国は日本に先行して語彙基盤の構築に着手。(欧州ISA / 米国NIEM)。2014年からは、EU、米国、日本が参加する各国語彙の連携会議が開始。民間では検索エンジン提供者の連合による schema.org が広く普及。
- 各語彙体系の相互運用性の確保を目指し、現在、活動中であることから、日本においては、共通語彙対応がなされたデータの流通を推進することで、海外連携を図ることができる。

EIRA: European Interoperability Reference Architecture
ISA: Interoperability Solutions for European Public Administrations
NIEM: National Information Exchange Model

③　自立的、持続的発展を担う運営体制

- 分野横断での取り組みの海外事例調査として、米国NIEM、欧州ISA、民間Schema.orgについて、各々の特徴の調査結果を示す。

	NIEM	ISA	Schema.org
概要	米国行政機関間での情報交換に用いる語彙とフレームワーク	欧州内行政機関間の相互運用性向上のための語彙やプロセス等を整備	webページの内容を検索エンジンに伝えるための語彙
類型	政府主導 （委員会ベース）	政府主導 （委員会ベース）	民間主導 （コミュニティベース）
相互運用性	厳密に確保	厳密に確保	寛容
変更頻度	低	低	高
運営体制の特徴	意思決定機関、実行組織、技術委員会、業務委員会、普及委員会、語彙最終決定から構成されており、各機関がそれぞれの権限をもち、運営されている。	欧州委員会、コミュニティ、実務組織、成果物レビュー組織、最終決定組織から構成されており、各機関がそれぞれの権限をもち、運営されている。	提案やフィードバックを行うことができる環境を提供し、利用者や開発者からの提案に対して、W3C内に設置された議論の場で議論した結果をフィードバックし、運営されている。
その他の特徴	ドメイン（分野）とその管轄省庁は1対1で対応づけされている。	欧州各国の代表で成果物のレビューを実施している。	GithubやWiki、メーリングリストを利用している
データ連携基盤の運営体制を検討する上での考慮点	厳密な運営が必要となる政府系データ連携基盤上では、政府ガバナンスの下での運営体制が望ましい。		技術革新の速い産業分野では、アプリケーション開発の活性化を図る上で、民間主導の運営体制が効果的

（参考）米国における分野間データ連携基盤の取組（NIEM）の運用組織体制

- 米国NIEM における各ドメイン（分野）とその管轄省庁の対応を以下に示す。

ドメイン(Domain)	取りまとめ省庁 (Executive Steward)
Agriculture (農業)	USDA (農務省)
Biometrics (生体認証)	DHS (国土安全保障省)
CBRN (Chemical, Biological, Radiological, Nuclear) (化学・生物・放射物質・核)	DHS (国土安全保障省)
CYFS (Children, Youth, and Family Services) (子供・若者・家族福祉)	HHS (保健福祉省)、DOJ (司法省)
Cyber (サイバー)	DHS (国土安全保障省)
Emergency Management (緊急事態管理)	DHS (国土安全保障省)
Health (保健)	HHS (保健福祉省)
Human Services (福祉)	HHS (保健福祉省)
Immigration (移民・入国)	DHS (国土安全保障省)
Infrastructure Protection (インフラ防護)	DHS (国土安全保障省)
Intelligence (諜報)	Criminal Intelligence Coordinating Council, Global Advisory Committee, DNI (国家情報長官)
International Trade (国際貿易)	US Customs & Border Protection
Justice (司法)	Global Justice Information Sharing, XSTF, DOJ(司法省), OJP
Maritime (海事)	US Navy, DHS(国土安全保障省)
Screening (監視)	DHS(国土安全保障省)

出典：“NIEM Communities” https://www.niem.gov/communities/Pages/communities.aspx
“公共情報交換標準スキームの整備に関する調査研究（2012年度）” http://datameti.go.jp/data/dataset/report-002-2012

参考資料

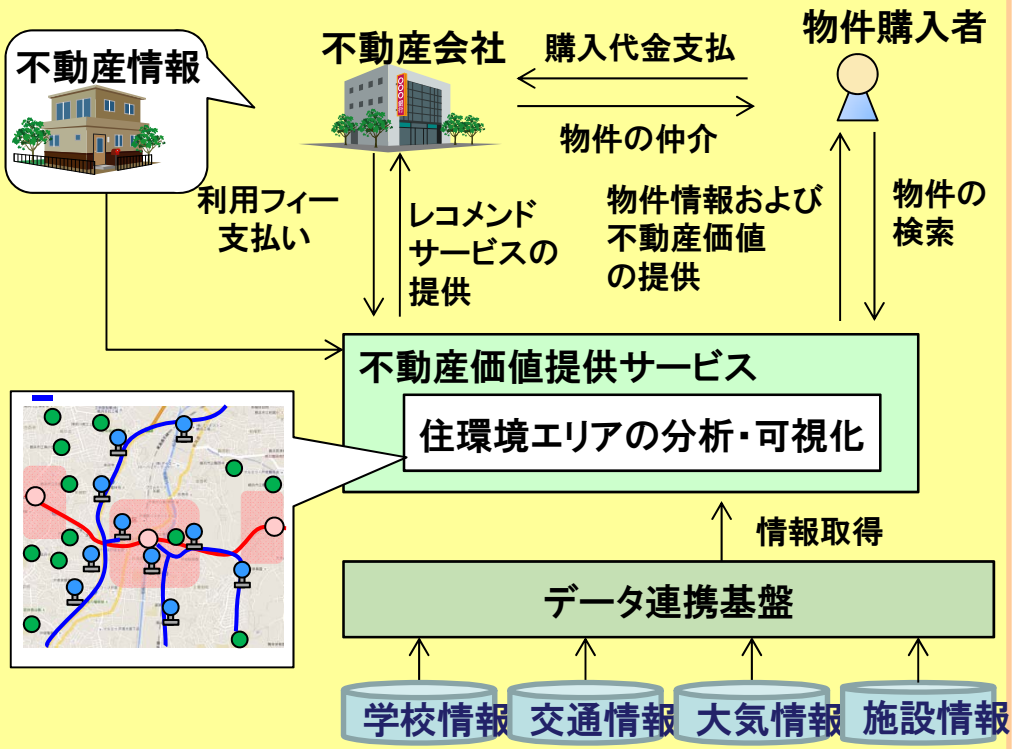
・アプリケーションイメージ
・前回(1/23実施第1回データ連携基盤サブWG)の主な意見

【参考】 アプリケーションイメージ①　（住環境情報提供／土地活用サービス）

アプリケーション概要

- 地理系（病院、公園、公民館・図書館、郵便局、ATM、役所等の位置）、自然環境系（年間降水日数、日射量、積雪量、花粉飛散量、騒音等）、交通情報等を活用し、住みやすい場所を見つける。
- 不動産会社は、物件購入/賃貸検討者に経済・環境・教育などの個人の好みに応じた理想的なエリアをレコメンドする。

アプリケーションイメージ



データ利用者のメリット

・物件購入/賃貸者
　希望の住環境を満たす物件購入
・不動産会社
　購入者のニーズに合った物件提供による
　他社差別化

データ提供者のメリット

・全てのデータ提供者
　データを有効活用してもらえる

使用するデータ

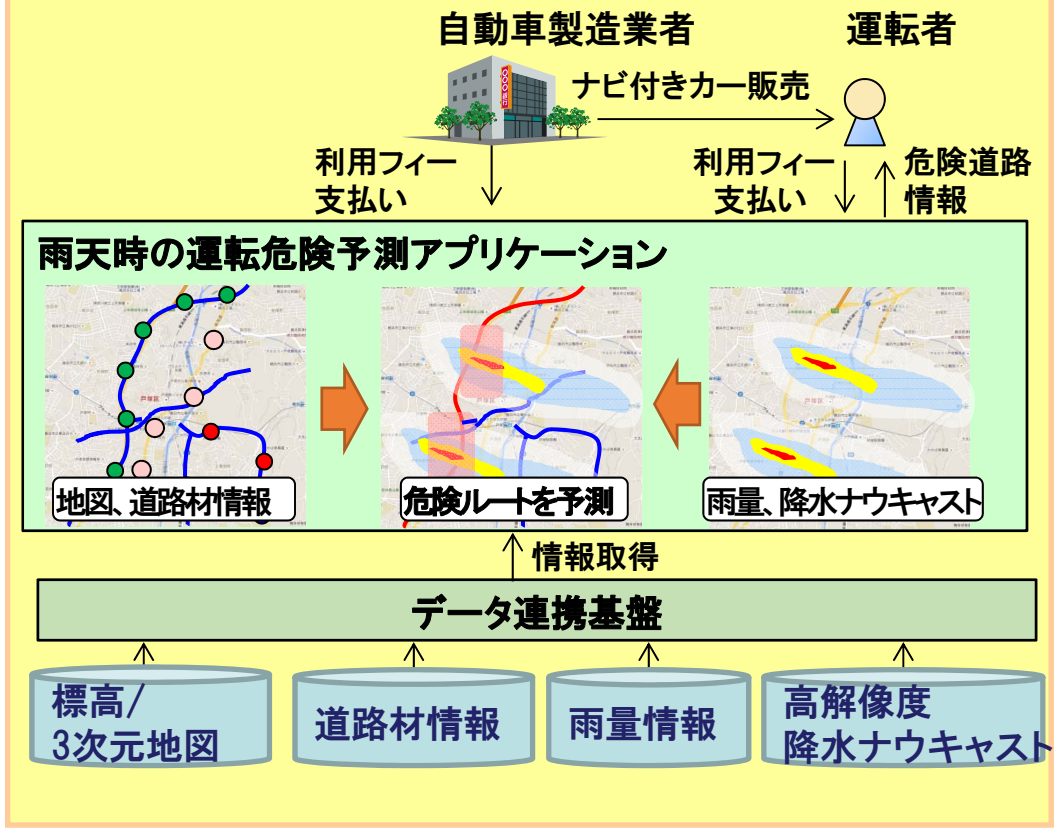
・地理系データ
　病院、公園、公民館・図書館、郵便局、
　ATM、役場等の位置情報
・自然環境系データ
　年間降水日数、積雪量、花粉飛散量、騒音等
・その他データ
　交通情報

【参考】アプリケーションイメージ②　（雨天時の運転危険予測）

アプリケーション概要

● 雨量情報や標高/3次元地図、道路の道路材(水はけ)情報を組み合わせて、道路の冠水やハイドロプレーン等の危険予測し提供することで、より安全な運転経路の運転者への推奨や自動運転の制御が可能になる。

アプリケーションイメージ



データ利用者のメリット

・運転者
　交通事故予防
・自動車製造/ナビゲーションソフト開発業者
　運転サポート機能の高度化による差別化

データ提供者のメリット

・地図情報提供業者
　データ提供サービスによる売り上げ
・JMBSC
　気象データの活用促進

使用するデータ

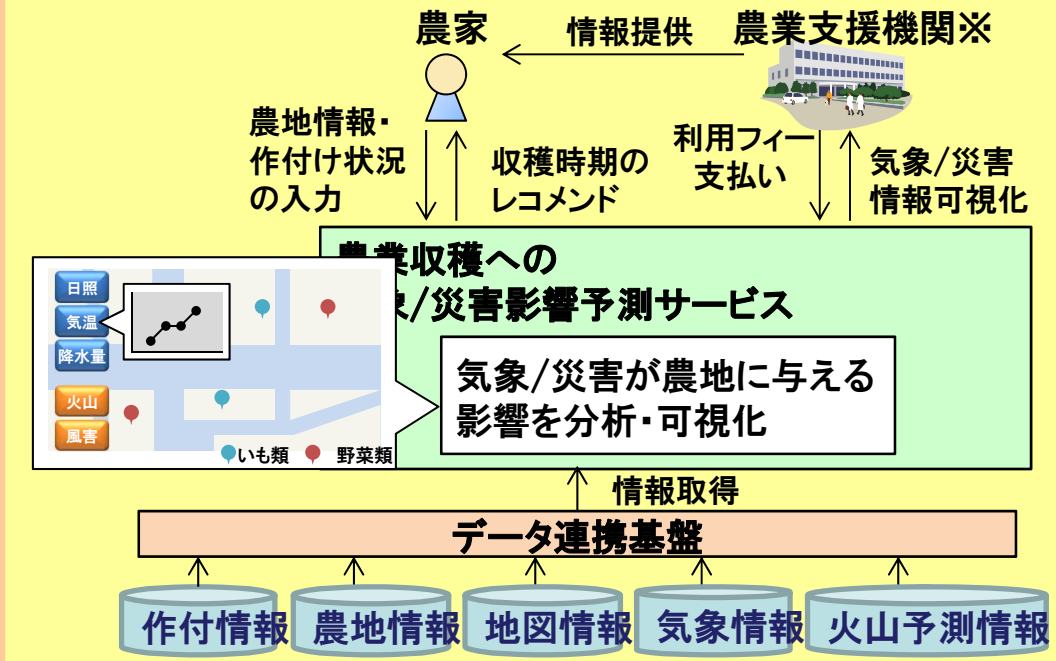
・地理系データ
　3次元地図、道路材情報、
　冠水注意箇所情報
・自然環境系データ
　雨量情報
　高解像度降水ナウキャスト(動的データ)

【参考】アプリケーションイメージ③　（農業収穫への気象/災害影響予測サービス）

アプリケーション概要

- 農地情報、作付け情報、長期/短期的な気象予測や災害予測情報をもとに、気象/災害が農業収穫に与える影響(リスク)を予測する。
- 農家は農地情報をもとに、作付け内容に応じた収穫時期のレコメンドを受ける。また、農業支援機関等が発表する地域固有の情報を参照する。

アプリケーションイメージ



※農業を支援する民間または公的な機関。
（JA、農業技術センター、農業支援センターなど）

データ利用者のメリット

・農家
作付けの最適化、気象/災害影響の最小化
・農業支援機関※
気象/災害影響の最小化

データ提供者のメリット

・全てのデータ提供者
データを有効活用してもらえる
・自治体
災害影響の最小化

使用するデータ

・地理系データ
地図情報、農地情報、作付情報
・自然環境系データ
長期/短期気象予測、災害予測情報

【参考】データ連携基盤サブワーキンググループ（第1回）における主な意見（1/3）

区分	主な課題
① サービスPF	➢ 分野毎の基盤を通さずとも、分野間連携基盤に直接データを提供できるような仕組みも必要 ➢ データフォーマット、語彙、API等の「標準化」と、「相互運用性」は別である。「相互運用性」が優先であり、既存のデータを扱う上では、語彙、フォーマットを変換するという方法もある。そうしつつも、新しいデータなどを徐々に標準化にもっていければ良い ➢ データ提供者にデータを出してもらうには、データの価値に応じた対価が必要。そのため、統一的な基準でデータ品質を正しく評価し、それを明記することが重要 ➢ どこにどんなデータがあるかを示す「データカタログ」と「データ品質」を産業界からのニーズが高く、データ流通推進協議会として特に重要と考えている。本サブWGと連携して具体化すべき

【参考】データ連携基盤サブワーキンググループ（第1回）における主な意見（2/3）

区分	主な課題
① サービスPF	➢ 欧州では官民で、FIWAREというプラットフォームを構築し、分野間データ連携の仕組みをオープンソースで構築し、欧州中心に110都市に普及拡大している。すべてゼロから開発する必要はなく、この技術資産を活用し、日本独自部分を追加して日本版のデータ利活用基盤にすべき
② フレームワーク	➢ ウォーターフォール型の開発ではなく、分野間・分野毎と同時に開発を進め、PDCAを回しながらより良いものにしていくことが大切 ➢ 各分野の分野間連携のメリットを外部に向けて発信する必要がある。効果が見えやすい分野間連携の姿を想定して、象徴的案件として先行的に進めるべき。例えば、SIP防災・減災では、実際の災害時に具体的に何ができるのかを関係者に示すことで、協力が得られやすくなった

【参考】データ連携基盤サブワーキンググループ（第1回）における主な意見（3/3）

区分	主な課題
② フレーム ワーク	➢ 分野毎のデータ基盤を立ち上げるにあたり、同じような課題を抱えているものと考える。各基盤のベストプラクティスや、データカタログ整備のロードマップ化など、情報共有するため、サブWGが終了した後も継続的な議論ができる場を設けるべき
③ ルール	➢ 機微なデータの取り扱いに対して、誰がアクセスできるか、誰がそれを判断するのか、評価や手順、ポリシーを明確にすることが必要 ➢ 公的データを積極的に基盤にのせることで、利用者が広がる。公共調達や公的資金による研究開発の成果として、1次データを公開するルールつくりが必要 ➢ データを提供するインセンティブや、提供したデータの2次利用等に対する責任問題の整理が必要