

農林水産省 御中

令和元年度AI-OCRを活用した 輸出貿易統計分析委託事業

実施内容とりまとめ報告書

PwCコンサルティング合同会社
2020年3月19日



目次

1. 本事業の概要
2. AI-OCR事業者の選定支援等
 1. AI-OCR事業者へのRFP
 2. AI-OCR事業者リスト(ロングリスト40社、ショートリスト4社)
 3. AI-OCR事業者評価基準及び評価結果
 4. AI-OCR事業者への発注
 5. AI-OCR実施方法及び必要となるスキル・作業時間
3. AI-OCR用いた輸出貿易統計のデータセット化
(1970～1987年、1988～2019年)
4. AIコンペの開催支援
 1. AIコンペ募集方法
 2. AIコンペ参加候補者情報
 3. AIコンペの企画案(コンペの課題、審査基準、スケジュール等)
 4. AIコンペ募集サイト
 5. AI/アナリティクスを活用した輸出分析及び政策立案に有効な情報リスト

別添1. AI-OCR事業者リスト:ロングリスト40社

別添2. 事業者情報(AI-OCRショートリスト4社、AIコンペ2社)

別添3. 輸出貿易統計のデータセット(1970～1987年、1987年～2019年)

別添4. 再委託先作業実施報告書(pluszero, SIGNATE)

1

本事業の概要

本業務の背景・目的の理解

背景

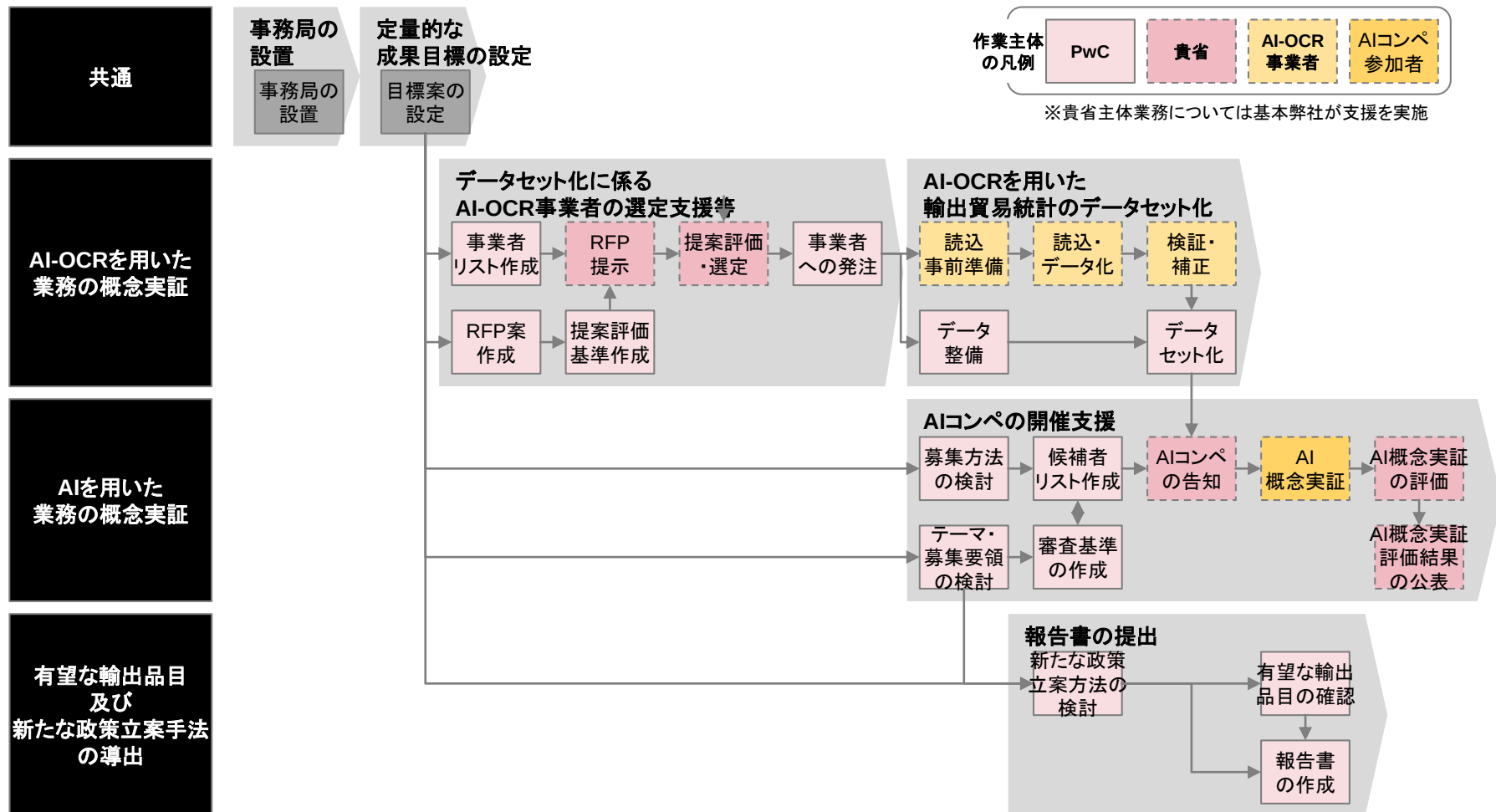
- 農林水産業・農林水産行政の現場では、少子高齢化の影響を強く受け、農業の担い手の不足が顕在化し、既存業務の在り方の見直しが必要
- 目まぐるしく変化する国境を越えた経済活動に伴い、需要を見越した生産活動体制の整備が必要
- 農林水産省職員自らが先進技術を理解し、適切に概念実証等の発注やオープンイノベーションの場の創出ができる能力を習得する必要がある

目的

- AI技術による紙データのデジタルデータ化や、大量のデータによる需要予測等を分析する業務の概念実証、輸出戦略等を検討するための新たな政策手法の導出(本事業では、品目・数値が標準化され、公表資料である輸出貿易統計を題材として実証)
- 農林水産省におけるAI 技術の知見を有した職員の人材育成

本事業の実施アプローチ

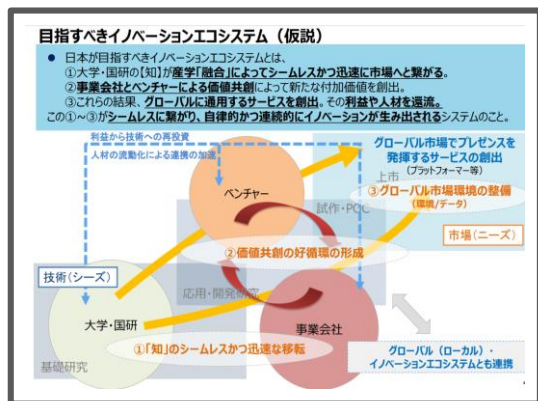
3か月弱にてAI-OCRを用いたデータセットを作成しAIコンペを実施するために、AIコンペ告知前にデータセットを用意することを想定し、AI-OCRに係る業務及びAIコンペに係る業務を並行して推進します



本事業を通じて目指されている姿(仮説)

貴省では、本事業を通じて、貴省職員と技術力のある日本発のベンチャー企業とのリレーションを構築し、今後、農林水産業や農林水産行政における課題の解決に向けて、概念実証等の発注やオープンイノベーションの場の創出を適切に実施できるようになることを目指されていると理解しております。弊社では、これを「農林水産省イノベーションエコシステム」として認識し、農林水産省を中心としたベンチャー等との連続的なイノベーション創出に寄与したいと考えています。

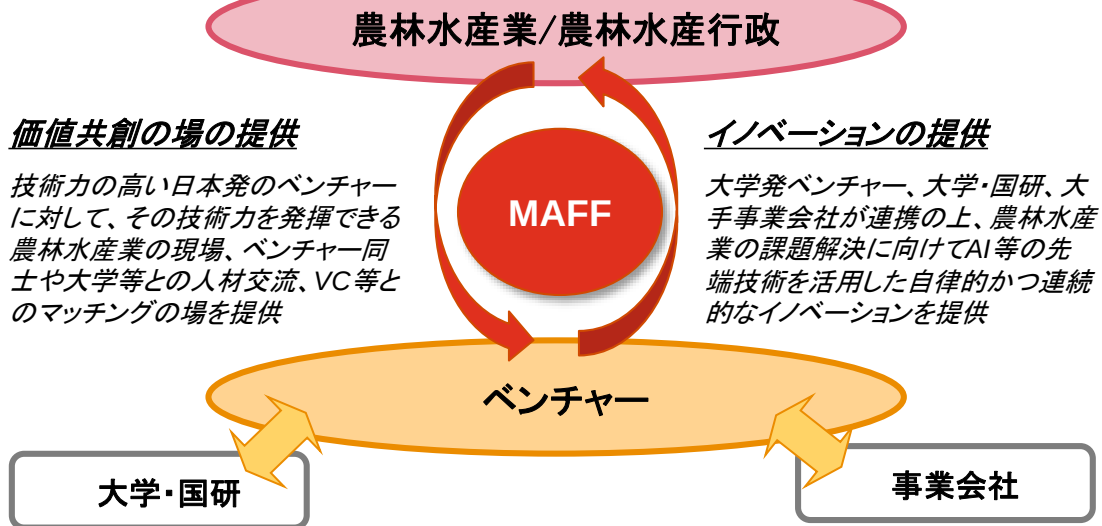
経済産業省「新たなイノベーションエコシステムの構築に向けて」を参照



農林水産省として…

農林水産省イノベーションエコシステム(仮)

農林水産省を中心としたベンチャー等との連続したイノベーション創出



本事業の実施スケジュール



目的	実施内容	タスク	1/6	1/13	1/20	1/27	2/3	2/10	2/17	2/24	3/2	3/9	3/16	3/23	3/30-
共通	マイルストーン														
	事務局の設置	事務局の設置													
	定量的な成果目標の設定	定量的な成果目標の設定													
AI-OCRを用いた業務の概念実証	データセット化に係るAI-OCR事業者の選定支援等	事業者リスト作成													
		RFP案作成													
		提案評価基準作成													
		RFP提示													
		提案評価・選定													
		事業者への発注													
	AI-OCRを用いた輸出貿易統計のデータセット化	読込事前準備													
		読込・データ化													
		検証・補正													
		データ整備													
AIを用いた業務の概念実証	AIコンペの開催支援	データセット化													
		募集方法の検討													
		テーマ・募集要領の検討													
		審査基準の作成													
		AIコンペの告知													
		AI概念実証													
		AI概念実証の評価													
		AI概念実証評価結果の公表													
有望な輸出品目及び新たな政策立案手法の導出	報告書の提出	新たな政策立案方法の検討													
		有望な輸出品目の確認													
		報告書の作成													

3/19: 報告書提出
(契約完了日)

AIコンペ告知

定量的な成果目標の設定

事務局は、本事業の事業目的を踏まえ、3つの目的ごとに目的を実現できたかを検証する定量的な成果目標を着手前に設定します。

	本事業の目的	定量的な成果目標(案)
AI-OCRを用いた業務の概念実証	✓ 農林水産省職員が、複数のAI-OCR エンジン、前処理・後処理手法、データセット化等について理解を深め、今後、業務の中でAI-OCR を適切な技術を持つ者に、適切な価格で発注できるようにすること	✓ 事務局から2者以上のAI-OCR 事業者分散してデータセット化の業務を発注する(仕様書記載) ✓ 今後のAI-OCRの実証実験、本格展開を見据え、ベンチャー・大手事業者含めAI-OCR事業者20社以上のリストを作成する ✓ AI-OCRの事前作業・読込/データ化・事後作業の3つの分類にて作業内容・必要スキル・価格帯を整理する
AIを用いた業務の概念実証	✓ AI コンペの開催を通じて、農林水産省職員が上記特について理解を深め、今後、業務の中でAI を適切な技術を持つ者に、適切な価格で発注できるようにすること	✓ 今後のAIの実証実験、本格展開を見据え、ベンチャー・大手事業者含めたAI事業者10社以上のリストを作成する
有望な輸出品目及び新たな政策立案手法の導出	✓ AI そのものに対する知見から、今後、農林水産省が輸出戦略、輸出のための生産基盤強化戦略等を検討するための新たな手法を導出すること	✓ AI/アナリティクスを活用した輸出の今後の予測及び生産計画の検討に有効な情報ソース10以上のリストを作成する

2

AI-OCR事業者の 選支援等

1. AI-OCR事業者へのRFP
2. AI-OCR事業者リスト
(ロングリスト40社、ショートリスト4社)
3. AI-OCR事業者評価基準及び評価結果
4. AI-OCR事業者への発注
5. AI-OCR実施方法及び必要となるスキル・作業時間

1. AI-OCR事業者へのRFP

AI-OCR事業者への委託内容RFP

AI-OCR事業者への委託は下記内容にて実施する

読取り対象	✓ 貿易統計過去データ(1970年～1987年) https://www.customs.go.jp/toukei/suui/index.htm 上記リンク>表の左列「品別国別表」>「品別国別月表_輸出品別国別表」の最初の1つor2つの食品(たばこまで)
読取り枚数	✓ 35ページ/年×18年＝630枚 ※1年当り30～35ページのダウンロードデータを弊社から提供
データ化の対象	✓ 品目CD・品目名称、国名、単位、累計数量、累計価格、の5つ ※上記読取り年・枚数分です
AI-OCR業者 委託作業	① 読取りの事前作業 ・ 精度をあげるための枠・文字種類設定 ・ 読込させるための解像度や傾きの調整 ・ その他必要な事前作業 ② 読みとり・データ化作業 ③ 読みとった結果の検証作業・再データ化 ・ 読取り結果の検証・補正 ※各年度最初と最後の1, 2ページ程度のみを対象に全件チェックを想定 ・ 補正結果を教師データとしての再読取り・データ化
納品物	1. 630枚分をエクセルに出力した出力結果(Excel or CSV) 想定項目:年、品目CD・品目名称、国名、単位、累計数量、累計価格 2. OCRのやり方(事前準備・事後検証・補正等含む) 3. 上記2の各作業の実施に必要なスキル・知識・習熟度との方が作業に要した時間

参考)AI-OCR事業者への発注の読取り対象

品名・国名 COMMODITY & COUNTRY	単位 UNIT	当月 CURRENT MONTH		累計 CUMULATIVE YEAR TO DATE	
		数量 QUANTITY	価額 VALUE	数量 QUANTITY	価額 VALUE
112-430 リキュール Liqueurs					
TAIWAN L	-	-	-	660	189
HG KONG L	9840	2556	21048	10366	
U KING L	-	-	129	131	
NETHLD L	-	-	14330	7152	
FRANCE L	-	-	518	200	
FR GERM L	849	403	7410	2810	
ITALY L	259	161	2147	1178	
AUSTRIA L	-	-	3492	1763	
CANADA L	-	-	702	629	
USA L	91835	26313	477625	154433	
AUSTRAL L	990	301	2205	854	
GUAM L	-	-	990	278	
TOTAL L	103773	29734	531256	179983	
(E . C) L	1108	564	24534	11471	

✓ 品目CD・品目名称、
国名、単位を読取り

✓ 数量・価額ともに
年間累計のみを読取り

✓ 農林水産物の最初から
飲料の最後まで(=たばこの
まで)の品目をすべて読取り

品名・国名 COMMODITY & COUNTRY	単位 UNIT	当月 CURRENT MONTH		累計 CUMULATIVE YEAR TO DATE	
		数量 QUANTITY	価額 VALUE	数量 QUANTITY	価額 VALUE
24.01-210 その他の葉たばこ (骨を除いてないもの) Other leaf tobacco, unstripped					
FR GERM K G	-	-	16807	38168	
TOTAL K G	-	-	16807	38168	
(E . C) K G	-	-	16807	38168	
24.02-200 紙巻たばこ Cigarettes					
TAIWAN T H	-	-	6008	14944	
HG KONG T H	1000	2287	12900	26212	
THAILND T H	-	-	608	1721	
SINGAPOR T H	500	917	2458	4892	
MALAYA T H	-	-	1211	2641	
INDONESIA T H	250	357	898	1504	
SRI LANK T H	-	-	279	590	
MALDIVE T H	430	909	590	1360	
BANGLA T H	-	-	20	120	
IRAN T H	-	-	1500	3771	
P YEMEN T H	-	-	283	724	
KUWAIT T H	-	-	1080	2239	
U ARB E T H	-	-	484	1350	
NETHLD T H	-	-	3230	7234	
BELGIUM T H	1000	2210	4000	9295	
USSR T H	-	-	2800	4557	
GREECE T H	60	184	60	164	
BULGAR T H	-	-	7180	12353	
USA T H	-	-	158	459	
MALI T H	-	-	158	419	
AUSTRAL T H	-	-	400	891	
NE HEHR T H	-	-	400	798	
SOLOMON T H	-	-	380	604	
R NAURU T H	-	-	280	538	
GUAM T H	-	-	170	335	
TOTAL T H	3240	6744	49232	99575	
(E . C) T H	1000	2210	7250	16529	

25.01-000 塩(岩塩、海塩及び食卓塩を含む。)純塩化ナトリウム、
かん水及び海水
Common salt (including rock salt, sea salt and table salt);
pure sodium chloride; salt liquors; sea water

R KOREA	MT	11	2362	66	7211
N KOREA	MT	-	-	10	238
TAIWAN	MT	-	-	10	1773
HG KONG	MT	0	157	0	989
THAILND	MT	-	-	8	1309
SINGAPOR	MT	-	-	2	203

想定データセット

OCRで読みとったデータは、今後の分析活用を見据えて、下記フォーマットにてデータセット化して頂くことを想定

※1988年以降のデータと統合したうえでの分析が容易であれば若干異なるフォーマットでも可能です
適宜、ご相談ください

年	品名コード	品名	単位	国名	累計数量	累計金額
1976	0101-000	馬、ろ馬、ら馬及びヒニー	NO	R KOREA	90	41124
1976	0101-000	馬、ろ馬、ら馬及びヒニー	NO	TAIWAN	10	2799
1976	0101-000	馬、ろ馬、ら馬及びヒニー	NO	USA	13	120485
1976	0101-000	馬、ろ馬、ら馬及びヒニー	NO	TOTAL	113	164408
1976	0101-100	牛(増殖用のもので純血種のもの)	NO	CHINA	1	2025
1976	0101-100	牛(増殖用のもので純血種のもの)	NO	TAIWAN	1	184
...	...	...	...	...	...	...

英語名はなしでもOK

品名ごとのTotal行は識別なしでOK
ある場合は、識別し削除しやすくなっていること

2. AI-OCR事業者リスト (ロングリスト40社、ショートリスト4社)

AI-OCR事業者リスト:ロングリスト40社

自社にてAI-OCRを開発・提供している事業者を中心に40社リストアップ

※詳細は「AI-OCR事業者リスト_v1.xlsx」を参照

#	事業者名	事前コンタクト 事業者
1	ABBYジャパン株式会社	✓
2	AI inside株式会社	✓
3	AnyTech株式会社	
4	Automagi株式会社	✓
5	株式会社Cogent Labs	✓
6	株式会社EduLab	✓
7	Kofax Japan株式会社	
8	LINE株式会社	
9	NECネクサソリューションズ株式会社	
10	東日本電信電話株式会社(NTT東日本)	
11	株式会社NTTデータNJK (ブランド名:メディアドライブ)	✓
12	株式会社PFU	
13	株式会社アジラ	
14	アライズイノベーション株式会社	
15	株式会社インフォディオ	○:再委託候補
16	沖電気工業株式会社	
17	オートメーションラボ株式会社	✓
18	キャノンITソリューションズ株式会社	
19	グローリー株式会社	
20	サンネット株式会社	

#	事業者名	事前コンタクト 事業者
21	株式会社シナモン	
22	スターティアレイズ株式会社	
23	ソースネクスト株式会社	
24	ネットスマイル株式会社	○:再委託候補
25	東芝デジタルソリューションズ株式会社	
26	株式会社ハイパーギア	
27	パナソニックソリューションテクノロジー株式会社	✓
28	株式会社ハンモック	
29	株式会社日立情報通信エンジニアリング	
30	株式会社日立ソリューションズ	
31	ファーストアカウンティング株式会社	
32	富士ゼロックス株式会社	
33	富士通フロンテック	
34	株式会社プリマジェスト	
35	三菱電機インフォメーションシステムズ株式会社	○:再委託候補
36	株式会社ユニメディア	
37	リコージャパン株式会社	
38	ロゴヴィスタ株式会社	
39	株式会社pluszero	◎:再委託者
40	株式会社LABOT	○:再委託候補

AI-OCR事業者リスト：ショートリスト4社

本事業にて実施するAI-OCRの特性及び事業期間を考慮し、汎用的製品だけでの対応ではなく、個別対応も含め実施可能な事業者を4社選定し、貿易統計資料のAI-OCRのサンプル実施を依頼

再委託先 選定方針・理由	✓ 既存コンタクト先より、後述する評価項目に応じて選定 【理由】 - 読取り精度の確保(本当に優秀かの確認)には、委託先にて一定レベル以上の工数をかけた工夫が必要 - 既に工数をかけた/かけようとしている事業者が存在 - 新たにコンタクトとする事業者が一定の工数をかけて検証する時間的余裕なし
-----------------	--

会社名	株式会社インフォディオ	ネットスマイル株式会社	株式会社pluszero	株式会社LABOT
AI-OCRの 特性	✓ 製品: Smart OCR ✓ 読取り精度を上げるための事前処理の多彩さ、独自の自然言語処理が強味	✓ 製品: Aiスキャンロボ ✓ 読取り精度を上げるための事前処理の多彩さ、設定の容易さが強味	✓ 製品: なし(個別対応) ✓ 自然言語処理に強味を持つ独自のAI-OCRエンジンをベースに本事業の特性を踏まえた個別対応を実施可能	✓ 製品: なし(個別対応) ✓ Googleが提供するAI-OCRエンジンをベースに本事業の特性を踏まえた個別対応を実施可能

3. AI-OCR事業者評価基準及び評価結果

AI-OCR機能評価のポイント

AI-OCRの機能を評価するうえでは、OCRエンジンとしての文字認識率だけでなく、認識率を高めるための前処理や、後処理（ベリファイ・後工程連携）の効率性等も含めて審査基準を設ける必要がある

文字認識率	対応文字	✓ 手書き／印刷文字 ✓ 文字種類 (ひらがな／カタカナ／漢字／英語／数字／記号)
	機械学習	✓ 過去修正履歴の当てはめ ✓ 画像認識
前処理	画像補正	✓ スポット(白点)埋め ✓ 傾き補正 ✓ 歪み補正 ✓ 解像度補正 ※今回5080dpi ✓ ノイズ(黒点)除去 等
	事前設定	✓ 読取り範囲指定の容易性((ドラッグ&ドロップ等) ✓ 読取り範囲内の行・列等の項目数の指定 ✓ 項目に対する文字種類指定(英語・数値) ✓ 項目に対するマスタ設定(ホワイト／ブラックリスト)
後処理	ベリファイ	✓ 確認対象(元画像・OCR読取り結果)の視認性 ✓ 確認・修正の操作性(マウスレス等) 等
	後工程連携	✓ アウトプット形式の指定 ✓ RPA等との連携(API) 等

品名・国名 COMMODITY & COUNTRY	単位 UNIT	当月 CURRENT MONTH		累計 CUMULATIVE YEAR TC DATE	
		数量 QUANTITY	価額 VALUE	数量 QUANTITY	価額 VALUE
CYPRUS	KG	-	-	13952	1523
CANADA	KG	-	-	1584	364
USA	KG	3835	1046	38310	9098
MAD CPV	KG	-	-	800	72
CAMROUN	KG	-	-	244	66
PAPUA	KG	-	-	1243	259
Y NEWGN	KG	-	-	1982	429
NEWCALD	KG	-	-	265	39
GUAM	KG	-	-	5692	1533
M N CAR	KG	-	-	14027	3247
TOTAL	KG	60539	12959	2511975	501648
(E E C)	KG	-	-	88	54
032-073 あわび (気密容器入りのもの) Abalone in airtight containers					
HG KONG	KG	32944	55681	47405	79339
THAILND	KG	1020	1881	1020	1881
SNGAPOR	KG	6630	10538	7882	12840
MALAYA	KG	2040	2342	2040	2342
CANADA	KG	-	-	621	855
USA	KG	2193	4026	9972	14473
TRJNIDD	KG	-	-	32	40
NEWZELD	KG	-	-	77	133
TOTAL	KG	44827	74468	69049	111903

列数設定
文字種類指定(英語)
マスタ設定(国名等)

列数設定
文字種類指定(数字)

委託業者選定の観点・評価項目

今回のAI-OCRの委託は、今後、貴省におけるデジタル化や政策立案手法・プロセスの見直し（EBPM：証拠に基づく政策立案）を進めるうえで適切なPoCを実施するパートナーの選定を兼ねており、①技術力、②今後のPoC協力の2つの観点より選定する

観点	評価項目	種別	評価内容	参考
技術力	AI-OCRの読取り精度	基礎	✓ 項目単位（≠文字。特に国、量・金額）の読取り精度 ✓ 読取り精度をあげるための工夫や技術（製品、製品外含む）	1970年の最初数ページでOK
	データセット	基礎	✓ 複数頁、1頁複数段（左段・右段）を1つのデータセットとして出力 ✓ データセットを出力・作成するための工夫や技術（製品、製品外含む）	
	品目・国	加点	✓ 年度によって品目や国の括りが変わることへの対応方法案	
今後のPoC協力	機動力	基礎	✓ 製品・サービスの標準仕様外へのニーズへの協力的取組み	
	領域幅	基礎	✓ デジタル化や政策立案手法やプロセスを見直すうえで有用な技術領域の多様さ	

AI-OCR委託先選定における評価結果

優秀さ・技術力と貴省における今後のPoC協力(機動力・領域幅)を中心に候補4社を評価した結果、pluszero社を委託先とする

社名 規模	優秀さ・技術力		PoC協力	
	読取り精度 Output	仕組み・工夫	機動力	領域幅
pluszero 資本金:不明 インターン中心に 100名強	99% 1頁複数表 対応あり	■ AI-OCR:独自エンジン ■ 表形式対応:個別開発	■ AI-OCRのPoCに 協力済み	広い ■ 各種予測モデル ■ ビジネス最適化 ■ 画像/動画分析 ■ データ構造分析(マイニング)
Labot 資本金:不明	98% 1頁複数表 対応あり?	■ AI-OCR:Googleエンジ ン+独自開発の学習・ ルール機能による学習 ■ 表形式対応:個別開発	■ AI-OCRのPoCに 協力済み	広い ■ 応用数学や機械学習を用いたシス テムの開発
Infordio 資本金:1億円	96% 1頁複数表 対応なし	■ AI-OCR:独自エンジン+ 独自開発の歪み文字生 成&学習 ■ 表形式対応:製品	■ 本件投資として協 力意向	狭い ■ 画像認識・文字認識エンジン開発 ■ 自然言語解析システム開発 ■ AI分類器の作成
ネットスマイル 資本金:約1億円	95% (高くはない) 1頁複数表 確認中	■ AI-OCR:独自エンジン ■ 表形式対応:製品	■ AI-OCRのPoCに 協力済み	広い ■ AI-OCR・RPA ■ 各種予測モデル ■ ビジネス最適化 ■ ビッグデータ解析

4. AI-OCR事業者への発注

pluszero社における作業概要及び実施結果

作業対象	読取り対象	✓貿易統計過去データ(1970年～1987年) https://www.customs.go.jp/toukei/suii/index.htm 上記リンク>表の左列「品別国別表」>「品別国別月表_輸出品別国別表」の最初の1つor2つの食品(たばこまで)																																			
	読取り枚数	✓35ページ/年×18年＝630枚 ※1年当り30～35ページのダウンロードデータを弊社から提供																																			
	データ化の対象	✓品目CD・品目名称、国名、単位、累計数量、累計価格、の5つ ※上記読取り年・枚数分です																																			
実施作業	実施作業	✓画像処理を行い、認識対象範囲を要素(品目/国名/単位/数値)ごとに分けて抽出。 ✓品目に対してはGoogle Cloud Visionを活用。 ✓その他の要素に対しては機械学習でモデルを作成して文字認識を実行。 ✓機械学習を用いた部分については、分類結果に対する目視での補正と再学習を実行。 ✓統計品目表を認識し、目視での補正とアルゴリズム的工夫を経て精度向上に活用。																																			
	検証作業	✓読み取り対象から無作為に抽出した4ページに対し、読み取り結果を目視で補正して正解データを作成し、照合することで精度の検証を実行。																																			
作業結果	データ化件数	✓6689個の品目数、94941個の明細数に対してデータ化を実行。																																			
	検証結果	<table><tr><th></th><th>単位</th><th>国名</th><th>数字(単語単位)</th><th>品目コード</th><th>品目</th></tr><tr><td>1972年(2)15ページ目</td><td>99.63</td><td>91.45</td><td>99.07</td><td>100.00</td><td>94.12</td></tr><tr><td>1975年(1)1ページ目</td><td>100.00</td><td>89.69</td><td>100.00</td><td>100.00</td><td>100.00</td></tr><tr><td>1978年(2)5ページ目</td><td>100.00</td><td>99.35</td><td>97.74</td><td>100.00</td><td>100.00</td></tr><tr><td>1987年(1)10ページ目</td><td>99.22</td><td>94.53</td><td>99.22</td><td>90.91</td><td>90.91</td></tr><tr><td>全体</td><td>99.69</td><td>93.68</td><td>98.92</td><td>97.22</td><td>95.83</td></tr></table>		単位	国名	数字(単語単位)	品目コード	品目	1972年(2)15ページ目	99.63	91.45	99.07	100.00	94.12	1975年(1)1ページ目	100.00	89.69	100.00	100.00	100.00	1978年(2)5ページ目	100.00	99.35	97.74	100.00	100.00	1987年(1)10ページ目	99.22	94.53	99.22	90.91	90.91	全体	99.69	93.68	98.92	97.22
	単位	国名	数字(単語単位)	品目コード	品目																																
1972年(2)15ページ目	99.63	91.45	99.07	100.00	94.12																																
1975年(1)1ページ目	100.00	89.69	100.00	100.00	100.00																																
1978年(2)5ページ目	100.00	99.35	97.74	100.00	100.00																																
1987年(1)10ページ目	99.22	94.53	99.22	90.91	90.91																																
全体	99.69	93.68	98.92	97.22	95.83																																

OCRの手順

AI-OCRによるデータセット化として、大きく下記3つの作業を実施

1 国名/単位/数値/品名 の領域を分離する

- 認識対象となるPDFファイルは、タイトル、項目名、国名、単位、数値、品名、そして罫線など、さまざまな要素を含んでいる
- そのため、今回認識したい対象である国名、単位、数値、品名の部分のみの領域をあらかじめ切り出し、文字認識のプロセスが行いやすいようにする

2 個々の領域ごとに文字認識を行う

- 国名と単位はアルファベット、数値は数字、品名は数桁のコードと日本語や英語の文字列からなるため、これらの個々に対して条件を絞り込んだうえで文字認識を行う
- 国名と単位と品名は、事前に出現するパターンが絞り込めるので、精度を高めるためにそれらの情報も活用

3 データの書き出し

- 国名、単位、数値、品名それぞれに対して個別に文字認識を行った結果をエクセルファイルにまとめて出力

1. 国名/単位/数値/品名 の領域を分離する

認識対象の画像データを白黒のピクセル単位で見たとき、英数字は一定のサイズの長方形に収まる、黒い連結した領域となる

これらをもとに、傾きや領域間の距離などを考慮したうえ、国名・単位・数値の領域を抽出する

また、抽出した国名・単位・数値の領域の位置情報をもとに、品名の領域も抽出する

N ANTIL	KG	-	-	113	100
MARITAN	KG	-	-	6736920	1244980
CANARY	KG	1100	91	3410	283
S AFRICA	KG	-	-	2630	197
FIDJI	KG	-	-	450	156
GUAM	KG	427	244	3730	2373
TOTAL	KG	95655	28560	7996209	1513545
K E E CD	KG	-	-	347338	37033

031-389 軟体動物（生きているもの及び生鮮、冷蔵又は冷凍のもの）（その他のもの）
Molluscs, live, fresh, chilled or frozen, n.e.s.

RYUKYU	KG	-	-	19877	2199
HG KONG	KG	138	237	899	1379
THAILND	KG	56	147	83	234
U KING	KG	-	-	82	221
FRANCE	KG	-	-	7042	896
FR GERM	KG	-	-	2480	8396
ITALY	KG	-	-	3004	1173
CANADA	KG	-	-	181	103
USA	KG	11698	16224	70886	90206
AUSTRAL	KG	-	-	333	342
NE HEBR	KG	-	-	80	110
GUAM	KG	27	46	194	325
TOTAL	KG	11919	16654	107165	105588
K E E CD	KG	-	-	14526	10467

031-399 軟体動物（塩蔵、塩水づけ又は乾燥のもの）（その他のもの）
Molluscs, salted, or dried, n.e.s.

RYUKYU	KG	-	-	1360	837
HG KONG	KG	61	114	505	3638
SU ARAB	KG	-	-	62	80
U KING	KG	-	-	122	260

2. 個々の領域ごとに文字認識を行う

数枚のPDFファイルに対して正解データを作成し、英数字を切り出して、機械学習を行う

また、1976年以降の国名・単位に対しては、1970～1975年の国名・単位と大きく形式が異なっているため、1文字単位ではなく単語単位で機械学習を行う

出現するパターンが限られているため、数字部分と英語部分を分離することで認識精度を上げる

品名・品数コードに関しては、データの出現パターンが多様なため、Google Cloud Visionを活用して文字認識を行う

その後、統計品目表などから作成したリストをもとにした類似度判定や、数値に対するTOTAL値を用いた補正を行い、精度のさらなる向上を図る



3. データの書き出し

国名、単位、数値、品名をそれぞれ個別に文字認識したため、それらを一つのエクセルファイルにまとめます。

年	品名コード	品名	単位	国名	累計数量	累計金額
1970	001-100	牛(水牛を含む。)	NO	RYUKYU	1996	236515
1970	001-100	牛(水牛を含む。)	NO	USA	32	59174
1970	001-100	牛(水牛を含む。)	NO	TOTAL	2028	295689
1970	001-200	羊及びやぎ	NO	RYUKYU	9	144
1970	001-200	羊及びやぎ	NO	TOTAL	9	144
1970	001-200	羊及びやぎ	NO	RYUKYU	84	13005
1970	001-200	羊及びやぎ	NO	TOTAL	96	13005
1970	001-400	家さん	NO	RYUKYU	685911	74615
1970	001-400	家さん	NO	R KOREA	112785	50097
1970	001-400	家さん	NO	R CHINA	138230	53951
1970	001-400	家さん	NO	HG KONG	207843	26008
1970	001-400	家さん	NO	R VTNAM	2514547	270104
1970	001-400	家さん	NO	THAILND	30950	10942
1970	001-400	家さん	NO	SARAWAK	23198	3031
1970	001-400	家さん	NO	SABAH	46126	4886
1970	001-400	家さん	NO	BRUNEI	19930	1821
1970	001-400	家さん	NO	PHILPIN	44830	25124
1970	001-400	家さん	NO	INDNSIA	24355	7304
1970	001-400	家さん	NO	INDIA	11430	821
1970	001-400	家さん	NO	PAKISTN	240	74
1970	001-400	家さん	NO	AUSTRIA	625	244
1970	001-400	家さん	NO	M M CAR	2040	352
1970	001-400	家さん	NO	TOTAL	3938450	545015
1970	001-500	馬、ろ馬、ら馬及びヒニー	NO	R CHINA	7	1008
1970	001-500	馬、ろ馬、ら馬及びヒニー	NO	TOTAL	7	1008
1970	001-900	動物(主として食用のもの)(その他のもの)	KG	RYUKYU	58	299
1970	001-900	動物(主として食用のもの)(その他のもの)	KG	R CHINA	5	108
1970	001-900	動物(主として食用のもの)(その他のもの)	KG	HG KONG	125	639
1970	001-900	動物(主として食用のもの)(その他のもの)	KG	FRANCE	0	54
1970	001-900	動物(主として食用のもの)(その他のもの)	KG	USA	42	245
1970	001-900	動物(主として食用のもの)(その他のもの)	KG	TOTAL	230	1345
1970	001-900	動物(主として食用のもの)(その他のもの)	KG	(EEC)	0	54
1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	RYUKYU	283	512
1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	HG KONG	10650	21203
1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	THAILND	2134	4712
1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	SNGAPOR	262	923
1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	PHILPIN	1920	7433
1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	INDNSIA	187	167
1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	TOTAL	15336	34970
1970	011-200	羊又はやぎの肉(生鮮、冷蔵又は冷凍のもの)	KG	RYUKYU	1000	129
1970	011-200	羊又はやぎの肉(生鮮、冷蔵又は冷凍のもの)	KG	INDNSIA	300	168

データの書き出し

国名、単位、数値、品名をそれぞれ個別に文字認識したため、それらを一つのエクセルファイルにまとめます。

年	A	B	C	D	E	F	G
1	年	品名コード	品名	単位	国名	累計数量	累計金額
2	1970	001-100	牛(水牛を含む。)	NO	RYUKYU	1996	236515
3	1970	001-100	牛(水牛を含む。)	NO	USA	32	59174
4	1970	001-100	牛(水牛を含む。)	NO	TOTAL	2028	295689
5	1970	001-200	羊及びやぎ	NO	RYUKYU	9	144
6	1970	001-200	羊及びやぎ	NO	TOTAL	9	144
7	1970	001-200	羊及びやぎ	NO	RYUKYU	94	13005
8	1970	001-200	羊及びやぎ	NO	TOTAL	96	13005
9	1970	001-400	家さん	NO	RYUKYU	685911	74615
10	1970	001-400	家さん	NO	R KOREA	112785	50097
11	1970	001-400	家さん	NO	R CHINA	138230	53951
12	1970	001-400	家さん	NO	HG KONG	207843	26008
13	1970	001-400	家さん	NO	R VTNAM	2514547	270104
14	1970	001-400	家さん	NO	THAILND	30950	10942
15	1970	001-400	家さん	NO	SARAWAK	23198	3031
16	1970	001-400	家さん	NO	SABAH	46126	4896
17	1970	001-400	家さん	NO	BRUNEI	19930	1921
18	1970	001-400	家さん	NO	PHILPIN	44630	25124
19	1970	001-400	家さん	NO	INDONESIA	24355	7304
20	1970	001-400	家さん	NO	INDIA	11430	821
21	1970	001-400	家さん	NO	PAKISTN	240	74
22	1970	001-400	家さん	NO	AUSTRIA	625	244
23	1970	001-400	家さん	NO	M M CAR	2040	352
24	1970	001-400	家さん	NO	TOTAL	3938450	545015
25	1970	001-500	馬、ろ馬、ら馬及びヒニー	NO	R CHINA	7	1008
26	1970	001-500	馬、ろ馬、ら馬及びヒニー	NO	TOTAL	7	1008
27	1970	001-900	動物主として食用のもの(その他のもの)	KG	RYUKYU	58	299
28	1970	001-900	動物主として食用のもの(その他のもの)	KG	R CHINA	5	108
29	1970	001-900	動物主として食用のもの(その他のもの)	KG	HG KONG	125	639
30	1970	001-900	動物主として食用のもの(その他のもの)	KG	FRANCE	0	54
31	1970	001-900	動物主として食用のもの(その他のもの)	KG	USA	42	245
32	1970	001-900	動物主として食用のもの(その他のもの)	KG	TOTAL	230	1345
33	1970	001-900	動物主として食用のもの(その他のもの)	KG	(EEG)	0	54
34	1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	RYUKYU	283	512
35	1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	HG KONG	10550	21203
36	1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	THAILND	2134	4712
37	1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	SINGAPOR	262	923
38	1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	PHILPIN	1920	7433
39	1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	INDONESIA	187	167
40	1970	011-100	牛肉(生鮮、冷蔵又は冷凍のもの)	KG	TOTAL	15336	34970
41	1970	011-200	羊又はやぎの肉(生鮮、冷蔵又は冷凍のもの)	KG	RYUKYU	1000	129
42	1970	011-200	羊又はやぎの肉(生鮮、冷蔵又は冷凍のもの)	KG	INDONESIA	300	168

5. AI-OCR実施方法及び必要となる スキル・作業時間

必要なスキル・知識・習熟度・単価感

今回のAI-OCRと同様の作業を実施するためには、下記3つスキル・知識が必要
これらスキルを有する人材に作業を委託する場合、スタートアップでも月200万円程度必要

1 PythonとC++の実装ができること

- 今回の実装に用いた言語はPython及びC++
- 基本的にPython で動作しますが、処理の高速化のため、一部でC++を活用
- よって、PythonおよびC++が難なく実装できることが要求される

2 古典的なアルゴリズム及び最新のDeep Learningに精通していること

- 文字認識において、数字やアルファベット、英単語を学習させてモデルを作成
- そのため、古典的なアルゴリズムやDeep Learningに精通していることが要求される

3 アルゴリズムに関する知識と思考力

- 例えば画像に関する処理を行うにおいて、画像の特性を理解した上でこの領域をどう切り出すか、位置関係をどう取得するかななどの思考力、そしてそれらを現実的な時間で処理するためのアルゴリズムの知識が必要
- また、類似度判定やTOTAL値での補正でも、文字処理に対するアルゴリズムの知識が要求される

作業に要する時間

今回レベル・ボリュームのAI-OCRの実施には100時間程度の時間が必要

1 国名/単位/数値/品名 の領域を分離

- 画像の特性や構造を考慮した上、パラメータの修正や、様々なアプローチからの実験(失敗含む)などを行い、また1970～75年と1976～87年の形式が異なったため2通りのパターンに対応
・・・30時間程度

2 個々の領域ごとの文字認識

- 正解データの作成および再学習データの作成・・・15時間程度
- 機械学習の実装、結果の検証・・・30時間程度
- 類似度やTOTAL値での判定など、データの特徴を生かした修正・・・15時間程度

1 データの書き出し

- フォーマットの修正や、規格の修正など・・・10時間程度