

A)

IT・AI を活用した食品安全情報収集システムに関する実証

事業

A)作業報告書

(ア)効果の確認・検証等報告書

令和7年3月31日

みずほリサーチ&テクノロジーズ株式会社

情報通信研究部

目次

1. 背景と目的	5
2. プロジェクト全体管理	5
2.1. 事務局職員からの聞き取りおよび事務局内での立会	5
2.1.1. ヒアリング記録	5
2.1.2. 現行システム	7
2.2. 業務進捗報告会	8
2.3. IT・AI を活用した食品安全情報収集等検討会	8
2.4. 構築プラットフォーム	9
2.5. 開発方法	9
3. クローリング・スクレイピングによる情報収集	10
3.1. クローリング条件設定機能	10
3.1.1. 記事対象 URL のアクセス確認	10
3.1.2. クローリング実行可能性調査	10
3.1.3. クローリング条件の制約	11
3.1.4. クローリング条件 csv ファイル作成支援	12
3.2. クローリング実行機能	13
3.3. クローリング結果確認機能	14
3.4. スクレイピングの実行	15
3.4.1. スクレイピングの必要性	15
3.4.2. API 調査	15
3.4.3. スクレイピングの条件設定	16
4. AI 等による翻訳・要約翻訳・要約	19
4.1. クローリング結果文章選択機能	19
4.2. 要約機能	19
4.3. 翻訳機能	22
4.3.1. 多言語の翻訳実行調査	22
4.3.2. 記事作成画面からの翻訳の実行	23
4.3.3. お試し翻訳画面作成	24
4.4. 専門用語・定訳辞書機能	25
4.4.1. 辞書の作成	25
4.4.2. 辞書機能を加えた翻訳方法	26
4.4.3. 辞書登録画面	28
4.4.4. 辞書機能を有効にした翻訳	29
4.5. 記事編集機	30
5. 情報の保存と活用	33
5.1. クローリング結果 DB	33
5.2. 専門用語・定型辞書 DB	33
5.3. 翻訳・要約結果 DB	33
5.4. 記事 DB	33

5.5.	ラベル付与機能.....	35
5.5.1.	LLM によるラベル抽出の検討.....	35
5.5.2.	LLM によるキーワード抽出の検討.....	36
5.6.	ラベル DB.....	37
5.7.	記事情報参照機能.....	37
5.8.	クロス表表示機能.....	37
5.9.	印刷機能.....	38
6.	連携自動化.....	38
6.1.	パブリッククラウド上でのシステム構築.....	38
7.	保守.....	39
8.	IT・AI を活用した食品安全情報収集システムにおけるまとめ.....	39
8.1.	スクレイピングの必要性.....	39
8.2.	スクレイピングのコスト対効果.....	39
8.3.	プログラムからのアクセス禁止サイト.....	40
8.4.	LLM による翻訳・要約.....	40
9.	ドキュメントからの構造化抄録データベースの構築.....	41
9.1.	構造化抄録データベースの目的と課題.....	41
9.2.	項目の自動抽出の検証.....	41
9.2.1.	EFSA chemical contaminants in food and feed のスクレイピングのからの抽出.....	41
9.2.2.	EFSA chemical contaminants in food and feed の full article からの抽出.....	51
9.2.3.	LLM によるデータ抽出のまとめ.....	51
9.2.4.	ハルシネーションの判定.....	51
9.3.	構造化抄録データベース構築における課題のまとめ.....	53
9.3.1.	カテゴリー化と抽出項目の検討.....	53
9.3.2.	専門用語辞書の作成.....	53
9.3.3.	運用による修正.....	53
9.3.4.	自動抽出の最適化.....	54

結果の要約

本実証事業により、下記の事項が得られた。

IT・AI を活用した食品安全情報収集システム

1. クローリングでは作業の効率化に繋がらず、スクレイピングが必要。
2. スクレイピングは対象 web サイト毎に設定をする必要があり構築に費用がかかる。
3. スクレイピングは対象 web サイトの変更に対応するため保守メンテナンスが必要となる。
4. プログラムからのアクセス禁止サイトは今後も増えていく可能性がある。
5. LLM による要約・翻訳は実用レベルで可能であるが、食品安全情報として適しているのかは別途検討が必要である。

構造化抄録データベース構築

1. 食品安全領域の論文のカテゴリー化とカテゴリー毎の抽出項目の決定作業を、食品安全情報を提供しているエキスパートが行う必要がある
2. 決定したカテゴリー毎に文献からエキスパートが項目の抽出を行い、統計的評価が出来る程度の数の正解データセットを作成する必要がある。
3. LLM による抽出における課題
 - (ア) 記事の分類の定義は様々なパターンが考えられる。
 - (イ) 項目抽出はある程度可能であるが LLM とプロンプトの組合せ最適化が必要。
 - (ウ) 技術革新が早く、最新の技術を用いた方法論を迅速に開発する必要がある。
 - ▶ 例えば、推論機能付き LLM の登場によりハルシネーションの判定がある程度可能になった。

1. 背景と目的

食品安全委員会(食安委)は、「食品安全基本法第21条第1項に規定する基本的事項」(平成24年6月29日閣議決定)において、国の内外における食品の安全性の確保に関する情報をリスク管理機関と連携しつつ一元的に収集し、当該情報について整理、分析及びデータベース化を図ることが求められている。このため、食品安全委員会事務局では、情報・勧告広報課情報班が海外情報を収集、翻訳し、食品の安全性に関する情報について関係各省庁に提供するとともに、食品安全総合情報システムを構築し、ホームページを通じて、食品安全関係情報(食品安全情報)として広く国民へ食の安全に関する情報提供を行ってきた。食品安全情報として収集する情報源は、主として海外の食品関係行政機関や研究論文、報道であり、元情報は翻訳された後に技術的観点から校閲され、修正され、食品安全委員会委員による確認の後に公開されている。現在、情報の収集、翻訳、データの蓄積を含むこれらの作業はすべて人の作業によるものである。

このような中、近年、通常のリスク管理・評価等に関する情報に加えて、これまでの既成概念を超える食品が開発・提案されるなど新たな食品安全リスクの出現により、日々世界各国から発信される食品安全に係る情報の範囲・量が益々拡大しており、将来にわたって日本の食品安全を確保するためには、今まで以上に広範な情報の収集と日本の食品安全に影響する情報を早期に抽出することが求められており、これらに対応するためには、情報収集業務、翻訳業務、及び付随するデータの蓄積業務の効率化による情報収集範囲の拡大及び更なる分析を推進する必要がある。このため、情報収集とその効果的な活用及び翻訳・要約の自動化のためのシステムを用い業務の抜本的改革と効率化を図ることが必要となり、本実証事業では「IT・AIを活用した食品安全情報収集システム」のプロトタイプシステムを構築し、その効果の検証を目的とする。

2. プロジェクト全体管理

2.1. 事務局職員からの聞き取りおよび事務局内での立会

食品安全情報の収集、翻訳、記事作成を行っている担当者にインタビューを行い、作業工程の共有化を行った。

2.1.1. ヒアリング記録

2.1.1.1. 微生物系担当参与様

- クローリング作業と原稿情報の収集
 - European Food Safety Authority (EFSA)の確認
 - ◇ 与えられたトピック(複数)に対して、Publication on the topic の更新を確認し、必要な情報を手動で選定している。
 - European Centre for Disease Prevention and Control (ECDC)の確認
 - ◇ ニュースやパブリケーションの更新を確認している。必要な情報を手動で選定している。
 - 韓国のサイト(Chronic Wasting Disease (CWD)の確認)
 - ◇ 更新されたか確認している。
 - ◇ 韓国で定常的に見ている唯一のサイトである。
 - ◇ 統一的なデータベースがないため、このサイトで確認している。他の国のCWDについては別の担当者が実施している。
 - The University of Edinburgh の確認
 - ◇ Creutzfeldt-Jakob Disease (CJD)の発症数について確認している。

- ◇ ファイルが更新されており、かつ、中身を見て数値に特定の変更あった場合に記事化を行う。
- ◇ 記事化をしない場合でも、更新があった場合はファイルを保存している。
- ◇ 記事化を自動化することは難しいケースである。
- ◇ 同大のサイトは2年間更新されていないが、更新されたら対応するものもある。
- World Health Organization (WHO)のサイトの確認
 - ◇ Monthly risk assessment が更新されたら記事としている。
- ノルウェーのサイトの確認
 - ◇ CWD の記事が web page で出ることが確認できたら記事化し、ページ情報を付加していく。
 - ◇ 複数のデータから記事の一つにまとめている。
- フィンランドのサイトの確認
 - ◇ 地図に CWD 等の発生の更新があった場合に記事化する予定(実施実績はない)。
- スウェーデンのサイトの確認
 - ◇ 地図に CWD 等の発生の更新があったら記事化を行う。
- The Center for Infectious Disease Research and Policy (CIDRAP)の確認
 - ◇ 感染症情報が更新されたら記事化の検討を行い、場合によっては別担当に記事化を依頼する。
- Morbidity and Mortality Weekly Report (MMWR)の確認
 - ◇ 米国 Centers for Disease Control and Prevention (CDC)が出している論文サイトであり、週1回更新される。
 - ◇ 5つほど対象領域が割り当てられており、更新が1~2カ所あることが多い。
 - ◇ 記事化に関しては、タイトル、アブストラクトを確認して判断している。
 - ◇ 論文情報のため完全に網羅できていないわけではないが、可能な範囲で記事化している。
 - ◇ CC BY (クリエイティブ・コモンズ)ライセンスであり、翻訳公開が可能。
- Eurosurveillance の確認
 - ◇ ECDC が出している論文サイト
 - ◇ アブストラクトを翻訳することも多いが、全文を確認し、情報を追加することもある。
 - ◇ 翻訳公開が可能な CC BY ライセンスである。
- PubMed
 - ◇ 担当領域が割り振られており、それに関する情報収集を行っている。
 - ◇ キーワードは固定だが、ヒット数次第であり、適当なキーワードだと適度に絞りこむことが難しい。
 - ◇ Feedly (Rich Site Summary (RSS)リーダー)を用いて更新情報を取得している。
 - ・ PubMed 以外にも ECDC 等のサイトの RSS フィードを登録している。
 - ・ Feedly のフィルターを利用している。
- 翻訳・要約
 - 基本的には、アブストラクトやサマリーを翻訳機にかけて、翻訳前文章を確認しながら修正している。意味が良く分からない場合は、本文があれば本文を見て補足を行うこともある。
 - 短い PDF (1~2枚でアブストラクトがない場合)は全訳したものを記事化することがある。
 - 食安委で翻訳サイトと法人契約することが難しいため、個人依存となっている。

- ◇ DeepL を使用している(有料版は個人契約)。
- ◇ Google 翻訳を使用している人もいる。
- 記事は Access に格納している。
 - ◇ ワードで記載して Access にコピペ入力している。
 - ◇ 更新日などは、Access のインターフェースでプルダウン選択する形式。
 - ◇ システムの都合上コロン「:」の使用は禁止である。
- 英語担当であるが、韓国語、ノルウェー語の記事化も行っている。
 - ◇ 韓国語はテンプレートに当てはめている。
 - ◇ ノルウェー語から日本語への翻訳は精度が低いため、英語に翻訳して確認している。

2.1.1.2. アジア地域担当参与様

- 中国は翻訳公開の許可がとれない記事が多いため、公開データを収集し、そのデータを用いて新規に記事を書きおろすという形で対応している。
- 形式だった文章が多く、必要な部分に分かっているため基本はテンプレートに当てはめて作成している。
- 香港、マカオ、台湾、中国本土の食品安全センターのサイトを確認している。
 - トップページを更新一覧を確認する。
 - 更新があったら、記事の中身を確認する。
 - 香港は、英語情報と中国語情報を両方みながら確認している。
- テンプレート外の記事については個別対応が必要になる。
 - サイトの中で確認する場所は大体決まっている。例えば以下のような場所を確認している。
 - ◇ 公文書
 - ◇ プレスリリース
 - ◇ 調査報告
 - URL の変更が多いため「情報収集対象 HP 一覧.xlsx」にはトップページの URL のみを記載しており、そこから 3、4 回クリックして確認する必要がある。
 - 香港、マカオのサイトの情報から本土の情報を探しに行くことがあり、これを自動化することは困難だろう。
- 中国では輸出入の税関の情報も収集している。
 - 食品安全の情報ではないが、影響が大きい。
- 辞書情報は個人的なメモがあるが整理しておらず、提供は難しい。

2.1.2. 現行システム

現在のシステムについて、食品安全委員会(食案委)様から情報提供を受けた。

- 日報及び週報作成の業務フロー
 - 担当者が日報を日報データベース(Access)に登録する
 - ◇ 日報は、主にタイトルやテンプレートにあてはめた情報で作成された概要記事。
 - 日報を勧告広報課情報班が意見を出し合ってチェック・選定を行い、選定された記事に対して内容の修正・追加等ブラッシュアップにより作成された記事(週報)を週報データベース(Access)に登録する。その後食品安全委員の先生方の会合にて公開非公開の判断等が行われる。
 - 週報は公開用非公開に分けられ、公開用は公開用データベース登録する。

- 日報データベースは、広報課、評価課、食品安全委員の先生から閲覧可能。
- 週報データベースは、上記に加えて、広報課、評価課、食品安全委員の先生、各省庁に公開される。
- 現状、日報と週報で記事が2段階に分かれているが、これは速報性、マンパワーの問題があるため、自動的に週報レベルの記事を作成できるならば、2段階にしなくてよい。
 - ◇ 食品安全委員の先生が全部の記事を確認できるようにする。
 - ◇ 膨大になることも予想され、結局要約版も必要になることも考えられる。

2.2. 業務進捗報告会

開催した業務進捗報告会等の一覧を表 2-1 に示した。これらの議事録は「会議資料及び議事録」フォルダーにまとめた。

表 2-1 開催した進捗報告会議

開催日	時間	会議名
2024年10月3日	13:00～16:00	食品安全情報収集システムの資料提供と記事作成者へのインタビュー
2024年10月18日	14:00～15:00	キックオフミーティング
2024年10月31日	15:10～16:40	第1回業務進捗報告会
2024年11月14日	18:00～19:10	第2回業務進捗報告会
2024年11月29日	9:00～10:00	第3回業務進捗報告会
2024年12月11日	15:00～16:00	業務説明と進捗報告
2024年12月13日	16:00～17:30	第4回業務進捗報告会
2024年12月26日	10:00～11:00	第5回業務進捗報告会
2025年2月6日	16:00～17:00	第6回業務進捗報告会
2025年2月10日	17:00～18:30	EPA IRIS 記事構造化妙録例報告
2025年2月21日	15:00～16:00	第7回業務進捗報告会
2025年3月5日	13:00～14:00	構造化項目抽出結果報告
2025年3月12日	16:00～16:50	第8回業務進捗報告会

2.3. IT・AI を活用した食品安全情報収集等検討会

開催した IT・AI を活用した食品安全情報収集等検討会の一覧を表 2-2 に示した。これらの議事録は「会議資料及び議事録」フォルダーにまとめた。

表 2-2 IT・AI を活用した食品安全情報収集等検討会の開催一覧

開催日	時間	会議名
2025年1月22日	10:00～11:30	第1回 IT・AI を活用した食品安全情報収集等検討会
2025年3月28日	16:00～17:15	第2回 IT・AI を活用した食品安全情報収集等検討会

主管課担当者の求めに応じて「IT・AI を活用した食品安全情報収集等検討会」を開催する。本検討会では、本システム構築に直接関係する業務計画及び報告といった進捗管理のほか、効率的な業務実施のためのユーザーインターフェース作成に係る意見交換、試用中の不具合や改善事項等に係る連絡調整のほか、事務局内の個別システムや他課業務との効率的な連携、収集した情報の有効活用、その他 IT・AI

を活用することによる事務局業務の改善に関する事項を協議する。

2.4. 構築プラットフォーム

本実証事業では、アマゾンウェブ サービス¹(AWS)上にシステムを構築し、効果の確認・検証等を行った。AWS上にシステムを構築するメリットは下記のとおりである。

- ① 食安委様の環境にサーバーを設置する必要がない
- ② 食安委様に設定されたサーバーにシステムのインストール作業を行わなくてよい
- ③ アジャイル開発で更新されたシステムを直ぐに反映できる
- ④ 食安委様のネットワークから web ブラウザーを経由してアクセスするため、食安委様のネットワークにセキュリティ上のリスクはない
- ⑤ 必要なファイルは csv 形式のテキストファイルでダウンロードするためウイルス感染の懸念がない
- ⑥ クローリングでは大量のファイルをダウンロードするが、食安委様のネットワーク負荷にならない

2.5. 開発方法

本事業は「IT・AI を活用した食品安全情報収集システム」のプロトタイプシステムを構築し、その効果の検証が目的である。そのため要求仕様が明確ではなく、ウォーターフォール開発は困難である(図 2-1 参照)。そのためアジャイル開発を採用し、試行を繰り返しながら本番システムに求められる要件及び課題の収集を行った。

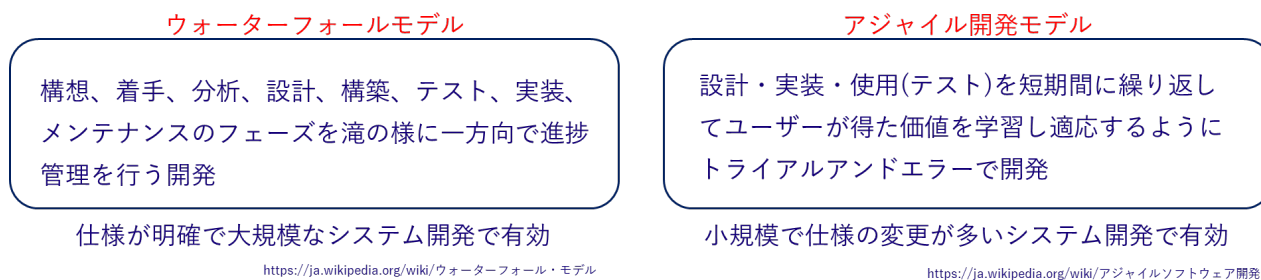


図 2-1 システム開発におけるウォーターフォールモデルとアジャイル開発モデル

¹ <https://aws.amazon.com/jp/>

3. クローリング・スクレイピングによる情報収集

3.1. クローリング条件設定機能

3.1.1. 記事対象 URL のアクセス確認

食案委様よりご提供を頂いた「情報収集対象 HP 一覧.xlsx」(以降、記事対象 URL リスト)には、「URL(又はアクセス先)」として 236 レコードの記載があった。これらについてプログラムからのアクセスが可能かどうかの確認を行った。ただし「具体的なアクセス先」として更に指定されているサイトもあるが、これは確認していない。この結果を表 3-1 に示した。「アクセス禁止」や「Web ページなし」には URL が存在しなくなっている場合が 20 サイトあり、他は web ブラウザーを使って手動ではアクセス出来た。「アクセス過多」については、本検証時の結果であり、異なるタイミングで実行すれば「アクセス成功」から発生する可能性もあるし、「アクセス過多」のサイトが「アクセス成功」に変わる可能性もある。またプログラムからアクセス出来ても規約を確かめるとプログラムからのアクセスを禁止している場合があるので注意が必要である。

表 3-1 記事対象 URL リストのプログラムからのアクセス結果

総レコード数	URL 記載なし	アクセス成功	アクセス禁止 (403)	Web ページなし (404)	アクセス過多 (429)	その他
236	3	158	16	24	32	3

記事対象 URL リストにおけるロボットによる収集が禁止されていたサイトを調査したところ、55 URL はロボットからの探索を禁止もしくは制限があった。この代表例を表 3-2 に示した。また欧州食品安全機関(EFSA)はアクセス頻度制限があり、プログラムからの取得には工夫が必要であることが分かった。

表 3-2 ロボットによるアクセスが禁止もしくは制限があったサイトの例

米国食品医薬品庁 (FDA)	中国疾病予防コントロールセンター
米国食品安全検査局 (FSIS)	中国国家市場監督管理総局
フランス競争・消費・不正抑止総局 (DGCCRF)	中国海関総署
ドイツ連邦消費者保護・食品安全庁(BVL)	中国国家食品安全リスク評価センター
オーストラリア農薬・動物用医薬品局(APVMA)	

3.1.2. クローリング実行可能性調査

Web ページからの情報収集では、クローリング(crawling)とスクレイピング(scraping)がある。この概要を表 3-3 に示した。クローリングは網羅的収集を目的としており、探索深度を深くすると大量のデータをダウンロードすることになる。一方でスクレイピングは必要な情報をピンポイントで収集する場合であり、対象とする web サイトごとにスクレイピングの設定を作りこむ必要が生じる。

表 3-3 クローリングとスクレイピングの比較

	クローリング (crawling)	スクレイピング (scraping)
概要	Web ページのリンクをたどりながら網羅的にコンテンツもしくは URL のリストを保存する。	Web ページから特定の情報を抽出してダウンロードして保存。

実用例	・ 検索エンジンの作成(Googlebot) ・ 盗作検出	・ ホテルの価格まとめサイト ・ 求人情報のモニタリング
備考	Web サイトの負荷軽減のためプログラムからのアクセスを禁止しているサイトも多い。	Google の検索結果を利用してスクレイピングすることは Google の規約違反

記事対象 URL リストにおいて、毎日確認と記されている 42 URL について明らかに記事ではない header、footer、navigation の部分を除いて 3 階層(深度)まで探索してファイルを取得した場合の実行時間と取得ファイル数、ファイル容量を表 3-4 に示した。ただし 3 階層までクローリングを実行すると 1 プロセスでは現実的な時間で終わらなかったため 6 URL だけ 3 階層探索を行った結果を示した。この結果からユーザーに深度を指定させてクローリングを行うことは大きなリスクが発生することが分かった。

表 3-4 3 階層(深度)まで探索して取得したファイル数とファイル容量。深度 3 については時間がかかりすぎたため 6 URL までクローリングを行った結果を示している。

	実行時間	取得ファイル数	取得ファイル容量 (MB)
深度 1	3 時間程度	3,007	214.4
深度 2	9 時間程度	12,596	4417.9
深度 3	6 URL で 20 時間程度	15,517	6540.1

3.1.3. クローリング条件の制約

クローリングの階層が深くなると探索数が非現実的になる。そこで下記の制約を加えることにした。

- ・ Header・footer・navigation 部分について、タグ名だけでなく、class 名での部分一致も含めて、探索対象外とすることで、単一ページ内のクローリング探索量を削減。
- ・ 基本設定は 1 階層までとする(ただし条件指定により変更可)。
- ・ 1 URL ごとに最大 1000 ページまでの探索とする。
- ・ 1 URL ごとに最大 10 分間の探索とする。
- ・ 記事一覧のページ割りについては日次で収集することでページ割りを無視する方針とする。
- ・ URL の差を見て差分取得(収集済みのページは新たに収集しない)。
- ・ URL が変化せず内容のみが修正されるケースについては、ハッシュ値を計算して同じならば消す。

Header・footer・navigation 部分の例として Food and Agriculture Organization of the United Nations (FAO)のページを図 3-1 に示した。ページ割りとは、記事の一覧表示などで複数ページに及ぶ場合の分割表示のことであり、図 3-1 の右側中段の① ② ③と続く表示方法である。記事対象 URL リストにおいては「具体的なアクセス先」の URL を一つ一つ登録してもらうことになる。

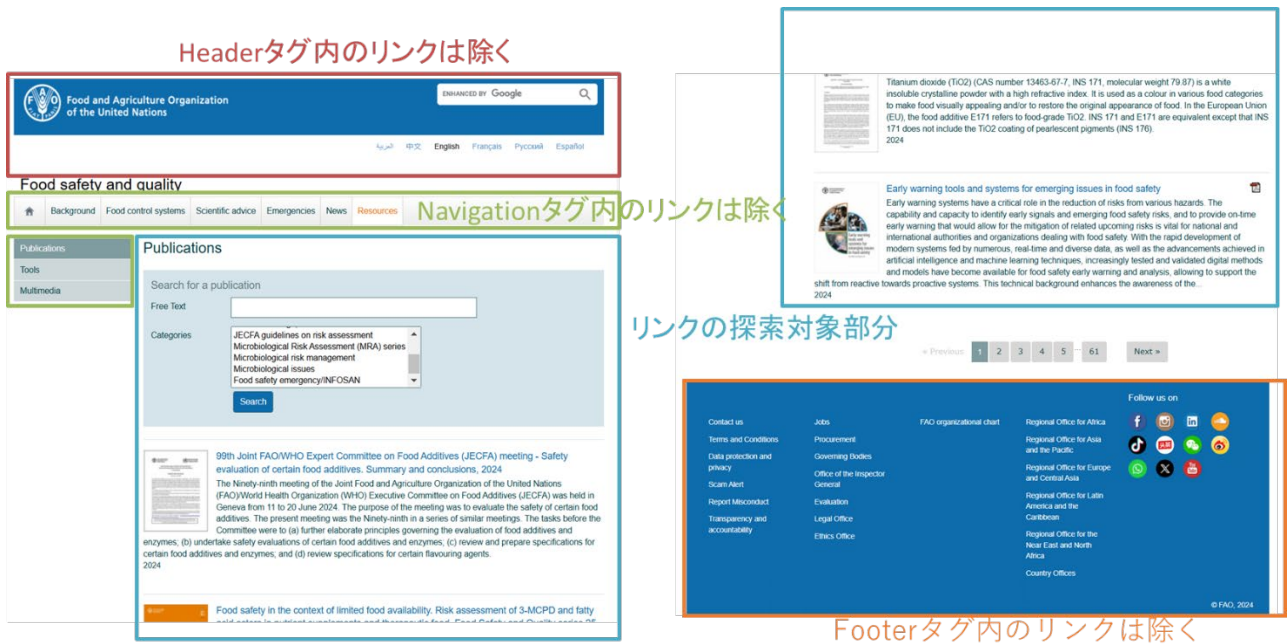


図 3-1 Food and Agriculture Organization of the United Nations におけるクローリング除外箇所とクローリング対処箇所の例

この対策の結果、記事対象 233 URL のクローリングにおいても実行時間が削減され、現実的な実行時間、収集ファイル数・容量となった。この結果を表 3-5 にまとめた。ただしこの中には 50 URL 程度はプログラムからのアクセスが禁止されている web サイト(表 3-2)が含まれている。

表 3-5 記事対象 233 URL のクローリングの実行

URL リスト数	親ページ(0 階層目)の取得数	合計取得ページ数	実行時間(h)	合計取得容量(GB)
233	134	10,563	5.45	2.54 GB

3.1.4. クローリング条件 csv ファイル作成支援

クローリング先を指定する場合、地域、国・地方、情報源などを指定し、この内部コードを付与する必要がある。これら選択して入力できるエクセルのマクロを作成し、クローリング条件 csv ファイルを出力するエクセルファイルを作成した。ただし探索する階層は 1 と設定している。このエクセルファイルの画面ダンプを図 3-2 に示した。地域、国・地方、情報源をプルダウンから選んでクローリングする URL を入力後「CSV 出力」ボタンを押すとクローリング条件 CSV ファイルが出力される。

図 3-2 地域、国・地方、情報源をプルダウンから選択して入力を支援するエクセルの「クローリング条件 csv ファイル出力マクロ」

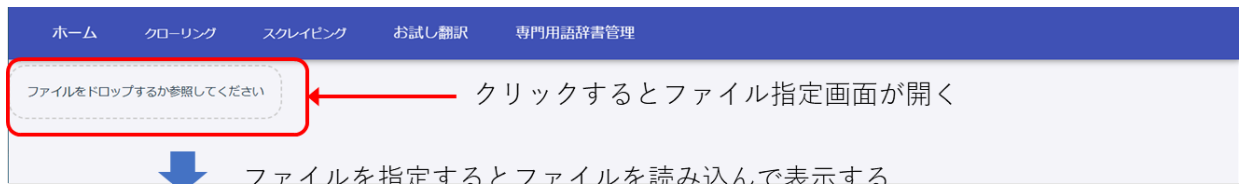
3.2. クローリング実行機能

3.1.4 節で作成したクローリング条件 CSV ファイルをアップロードし、クローリング対象を設定する画面を図 3-3 に示した。アップロードしたリストに従い、3.1.3 節で示した制約の下でクローリングを実行する。

クローリングでは、HTML ファイル、PDF ファイルの他に EPUB 形式及び Tableau 形式にも対応した。これらの形式からテキストファイルに変換するために使用したライブラリーを表 3-6 に示した。

表 3-6 EPUB 形式及び Tableau 形式に対応するために使用したライブラリー

形式	ライブラリー名	説明
EPUB	PyMuPDF (https://github.com/pymupdf/PyMuPDF)	EPUB 形式から PDF 形式への変換、及び PDF 形式から MD 形式への変換を行った。
Tableau	zipfile	標準ライブラリー。拡張子が.twbx のファイルは zip ファイルであり、この中に XML 形式のワークブックファイル(.twb)とデータファイルが含まれている。
	xml.etree.ElementTree	標準ライブラリー。XML 形式の読込とタグの削除を行った。



No	地域	AREACODE (自動入力)	国・ 地方	LOCALCODE (自動入力)	情報源 (公的機 関)	BULLETINCODE (自動入力)	深 度	頻 度	URL
1	その他 (国際機 関)	9	その他	250	国際連合食糧農 業機関 (FAO)	295	2	日	http://www.fao.org/
2	その他 (国際機 関)	9	その他	250	国際連合食糧農 業機関 (FAO)	295	2	日	http://www.fao.org/news/en/
3	その他 (国際機 関)	9	その他	250	国際連合食糧農 業機関 (FAO)	295	2	日	http://www.fao.org/food-safety/resources/publications/en/
4	欧州	5	オー スト リア	123	国際原子力機関 (IAEA)	521	2	日	https://www.iaea.org/
5	その他 (国際機 関)	9	その他	250	国際がん研究機 関 (IARC)	298	2	日	https://www.iarc.who.int/
6	その他 (国際機 関)	9	その他	250	世界保健機関 (WHO)	294	2	日	http://www.who.int/en/
7	その他 (国	9	その	250	世界保健機関	294	2	日	https://www.who.int/news-room/

図 3-3 クローリング条件 CSV ファイルアップロード画面とファイルを指定してアップロードした画面

3.3. クローリング結果確認機能

クローリングの結果確認画面を図 3-4 に示した。クローリング条件 CSV に記載の URL に対して、日時、データ形式、探索深度、ステータス、タイトルの情報が表示される。クローリングのため記事以外の情報も多数ダウンロードされる。この画面から記事となる原稿を選び、記事作成のボックスをクリックして記事作成画面を表示させて記事データベースに保存することで記事化する。

ホーム スクレイピング お試し翻訳 専門用語辞書管理

結果表示
記事一覧
記事作成

検索対象 クローリング設定 検索タイプ キーワード 期間絞り込み 機関名絞り込み

タイトル 記事本文 作成記事本文

AND あいまい検索 OR

2025/03/11 ~ 2025/03/18

地域: -- 地域を選択 --
国・地域: -- 国・地域を選択 --
機関: -- 機関を選択 --

検索

<< < > >>

1 ページに移動 Page1 of 95 表示件数 10 件

記事作成	クローリング条件URL	日時	データ形式	探索深度	ステータス	対象URLタイトル
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	0	成功	Homepage - ECHA
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	1	成功	All news - ECHA
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	1	成功	All news - ECHA
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	1	成功	All news - ECHA
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	1	成功	Maintenance messages - ECHA
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	1	成功	News archive - ECHA
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	1	成功	SCIP - ECHA
<input checked="" type="checkbox"/>	https://www.efsa.europa.eu/en/topics/topic/pesticides	2024/12/18	html	1	成功	Loading...
<input checked="" type="checkbox"/>	https://echa.europa.eu	2024/12/18	html	1	成功	Home - European Observatory for Nanomaterials
<input checked="" type="checkbox"/>	https://www.efsa.europa.eu/en/topics/topic/pesticides	2024/12/18	html	1	成功	Regulation - 396/2005 - EN - EUR-Lex

図 3-4 クローリング結果表示画面

3.4. スクレイピングの実行

3.4.1. スクレイピングの必要性

クローリングの結果を確認すると記事の原稿となるページ以外にも様々なページがダウンロードされていることが分かった。例えば表 3-4 では食安委で毎日確認している 42 URL のクローリング結果であるが、深度 1 でも 3,007 ファイルがダウンロードされている。記事化では、この中から目視で記事の原稿となるレコードを探し出さなければならない。この作業は結局元のページにアクセスして目視確認することになり、作業が効率化されているとは言えない。この対策として、

1. API 利用による記事原稿の収集
2. スクレイピングによる記事原稿の収集

の 2 つが考えられる。そこで記事対象における API 提供の状況を調査した。

3.4.2. API 調査

クローリングにおいて、何らかの理由によりアクセスが失敗する機関の中から、特に重要と思われる機関として選定した FDA、EFSA、USDA の 3 機関について API の提供状況を調査した。この結果を表 3-7 に纏めた。詳細な調査結果は「B)システム開発(ア)アジャイル開発で生じるドキュメント.docx」

の 1.1.2.4 項を参照されたい。結論として記事を収集するための API は各機関作成しておらず、API を通じた記事収集を行うことは現状では困難であることが分かった。

表 3-7 FDA、EFSA、USDA の API 提供状況

	FDA	EFSA	USDA
取得できる情報	・有害事象 ・製品リコール	なし	製品リコール
データの更新頻度	・(有害)四半期ごと ・(リコ)毎週	-	逐次
API によるアクセス判定	成功	-	失敗
API を用いないクローリングによるアクセス判定	成功/失敗が半々	成功/失敗が半々	失敗
サイトポリシーによる自動クローリングの禁止の有無	無	無	無

3.4.3. スクレイピングの条件設定

スクレイピングでは記事対象 URL の HTML ファイルの構造を解析し、個別に記事原稿個所を収集するようにプログラムを設定する。プログラム(ロボット)によるアクセス禁止のサイト(表 3-2)を除いても対象 URL は 179 URL あり、全てのスクレイピングを本調査で実行することは現実的ではない。そこで本調査では、記事原稿が多い米国 EPA と欧州食品安全機関(EFSA)を対象としてスクレイピングを実行することになった。

記事対象 URL 中の具体的なアクセス先を全て URL に展開すると米国 EPA は 8 URL、欧州食品安全機関(EFSA)は重複を含めて 44 URL があつた(図 3-5)。この一つ一つを食安委様と記事対象領域の確認作業を行い、スクレイピング対象を決定した。この確認作業の例を図 3-6 に示した(全確認資料は「第 1 回 IT・AI を活用した食品安全情報収集等検討会」の「EPA_EFSA スクレイピング説明及び構造化検討.pptx」参照)。

確認作業の結果をスクレイピングスクリプトに落とし込み、このスクリプトを毎日実行するように設定を行った。なお、EFSA はアクセス制限があるため毎日複数回実行している。2 ヶ月間の実行により、スクレイピングでは下記の設定となった。

- ① PDF ファイルはテキスト化して原文をデータベースに保存。
- ② スクレイピング実行時に英語要約、日本語要約を行い、記事データとしてデータベースに登録する。
- ③ 大きな PDF ファイルの記事を一度に多数公開することがあるため、時間がかかる日本語翻訳をスクレイピング時には行わない。
- ④ 日付情報のある記事については、取得失敗に対応するため過去 3 日間取得対象とする。
- ⑤ 一度取得した URL を保存しておき、この URL には次回以降アクセスしない。
- ⑥ EFSA のアクセス制限に対応するため、(クローリングで実行していた)ダウンロードしてからのハッシュ値による記事の更新チェックは行わない。

EFSA からのスクレイピングでは、EFSA Journal の full article のリンクへの対応も協議したが、提供元の Wiley の利用規約(<https://onlinelibrary.wiley.com/terms-and-conditions>)にクローリングやスクレイピングが禁止であることが明記されていたため、取得対象から除いた。

図 3-5 米国 EPA と欧州食品安全機関(EFSA)の記事対象 URL リスト

図 3-6 スクレイピング確認の例。この確認作業を重複、リンク切れを除いた 32 URL について行った。

スクレイピングでは、クロールングと異なり記事原稿を手動で選ぶ必要が無くなる。そのためスクレイピングスクリプトから直接記事データベースに登録することが可能になった。スクレイピングの結果確認は、図 3-7 に示したスクレイピング記事一覧画面から行う。

ホーム クローリング スクレイピング お試し翻訳 専門用語辞書管理 トレンド表示

検索対象
 タイトル
 記事原文
 作成記事本文

検索条件
 完全一致
 あいまい検索

検索タイプ
 AND
 OR

キーワード

期間絞り込み
 ~

機関名絞り込み
 地域:
 国・地域:
 機関:

検索

検索結果: 10件

ページに移動 Page 1 of 1 表示件数 10件

記事作成	スクレイピング条件URL	記事作成日時	データ形式	探索深度	ステータス	対象URLタイトル
🔗	https://www.efsa.europa.eu/en/publications?items_per_page=30	2025/03/27	html	1	成功	Pest survey card on Venturia nashicola EFSA
🔗	https://www.efsa.europa.eu/en/publications?items_per_page=30	2025/03/27	html	1	成功	Pest survey card on Grapholita inopinata, G. packardii and G. prunivora EFSA
🔗	https://www.efsa.europa.eu/en/publications?items_per_page=30	2025/03/27	html	1	成功	Pest survey card on Aplosporina morbosa EFSA
🔗	https://www.efsa.europa.eu/en/publications?items_per_page=30	2025/03/27	html	1	成功	Pest survey card on Saperda candida EFSA
🔗	https://www.efsa.europa.eu/en/publications?items_per_page=30	2025/03/27	html	1	成功	Pest survey card on Gremmeniella abietina EFSA

図 3-7 スクレイピング記事一覧表示画面

4. AI 等による翻訳・要約翻訳・要約

4.1. クローリング結果文章選択機能

クローリング結果表示画面は図 3-4 に示したが、クローリングでは作業の効率化にならないためスクレイピングを実行した。スクレイピングでは直接記事データベースに保存のため、図 4-1 に示した様にスクレイピング記事一覧画面から「記事作成」ボタンを押すことにより文章を選択する。

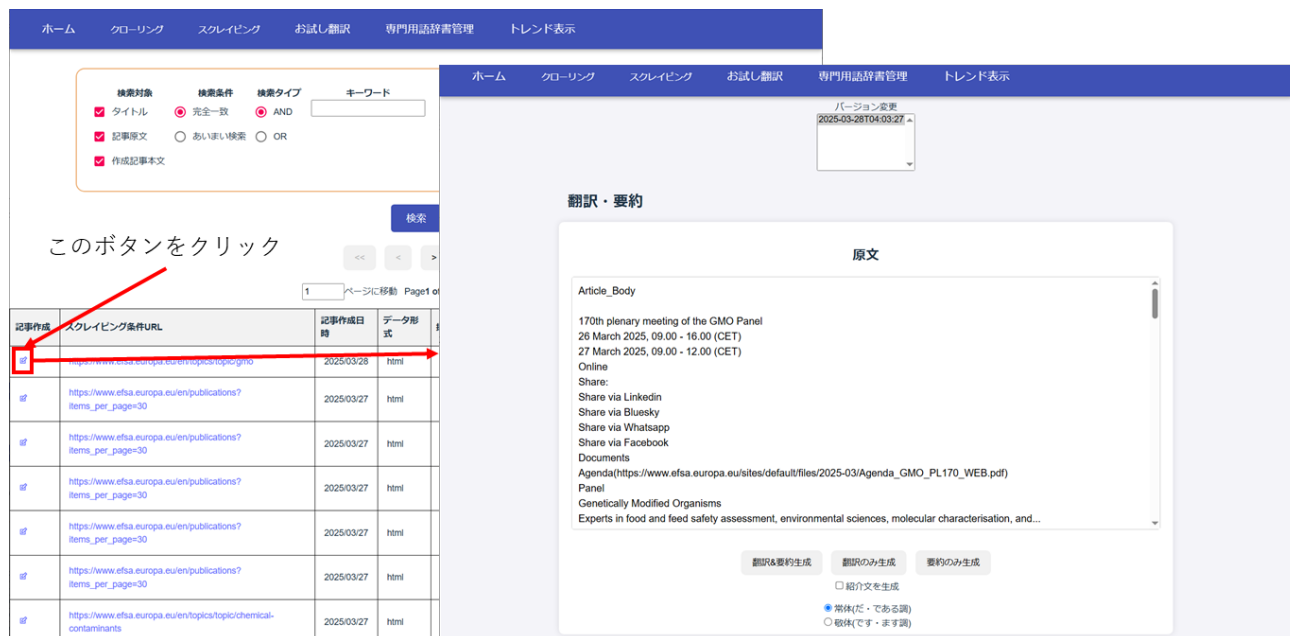


図 4-1 クローリング結果からの記事選択機能

4.2. 要約機能

AWS から利用できる claude-3-5-sonnet v1 を用いて要約機能の検討を行った。AWS 上で claude-3-5-sonnet v1 を利用する場合、PDF ファイルを直接アップロードして要約を実行することが出来るが、ファイルサイズ 4.5MB の制限があるため PDF ファイルをマークダウン形式のテキストファイルにして claude-3-5-sonnet v1 を利用することにした。

多くの LLM では毎回同じ回答をしないよう揺らぎを持たせている。このため「Document の内容を要約してください。」の指示文で要約させてところ、記事の要約の他に記事の紹介文の様な内容があった(「B)システム開発(ア)アジャイル開発で生じるドキュメント.docx」の 1.2.2.2 節参照)。また記事が長い場合、英語で要約を返す場合が散見された。そこでまず英語で要約を実施させ、これを日本語に翻訳することで安定して日本語要約が得られるようにした。英語要約では「要約の作成」もしくは「紹介文の作成」をする様に指示することにより選択可能にし、語調は日本語翻訳時に指定することを選択可能にした(図 4-2)。これらのプロンプトを表 4-1 に示した。

表 4-1 要約機能のプロンプト

機能	指示文
英語要約プロンプト	対象の文章の詳細な要約を英語で作成してください。要約以外の説明文は不要です。
英語紹介文プロンプト	対象の文章の紹介文を英語で作成してください。紹介文以外の説明文は不要です。

常体(だ・である調) 日本語訳プロンプト	翻訳対象の文章を日本語に翻訳してください。 翻訳以外の説明文は不要で、略語はそのまま翻訳してください。 語調は「だ・である調」としてください。
敬体(です・ます調) 日本語訳プロンプト	翻訳対象の文章を日本語に翻訳してください。 翻訳以外の説明文は不要で、略語はそのまま翻訳してください。 語調は「です、ます調」としてください。

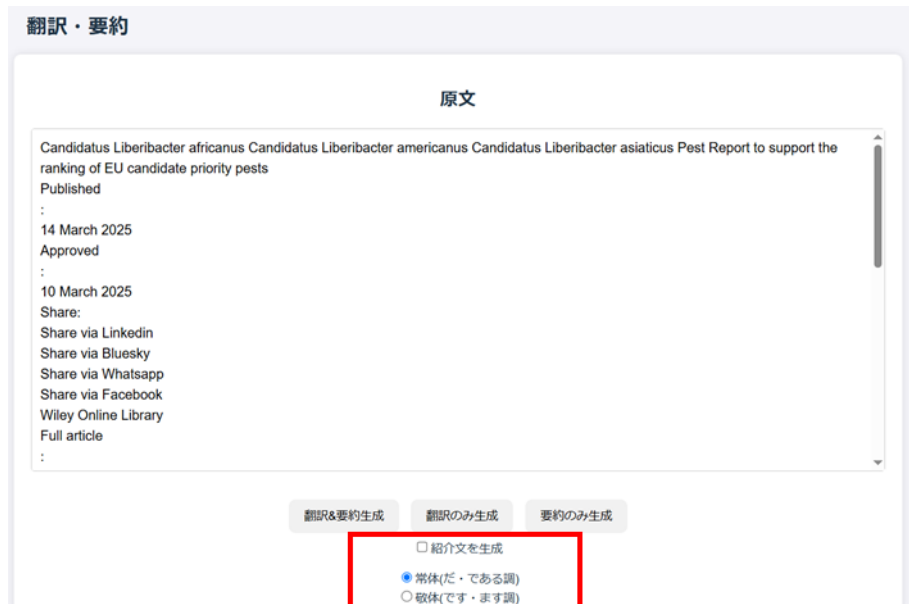


図 4-2 記事作成画面からの要約形式の選択

LLM から作成された要約及び紹介文の例を表 4-2 に示した。文章に違和感はなく、実用レベルになっていると思われる。ただし食品安全情報の記事として適しているのかどうかについては検討を行う必要があるだろう。

表 4-2 EFSA 記事における要約と紹介文の例

記事	要約	紹介文
presentation-cvmp-workplan-2024-beyond-johanschefferlie_en.summary Answer Translation	タイトル：2024 年以降の CVMP 作業計画 イベント：EMA 獣医薬品情報デー2024 主要ポイント： 1. CVMP 作業計画 2024： - EMA ウェブサイトで公開 - 以下を含む様々な横断的活動をカバー： - 抗菌薬耐性 - 抗寄生虫薬耐性 - 抗菌薬の代替品 - ワクチンと免疫学的製剤の利用可能性 - 食用動物種向け VMP の利用可能性 - 規制試験における 3Rs の適用 - 科学的能力/対応力の強化 - 国際協力と調和 - ステークホルダーの関与 2. CVMP 作業計画 2024 のハイライト： - 製品開発のサポート - 科学的助言	この文書は、EMA 獣医薬品情報デー2024 で発表された、2024 年以降の獣医薬品委員会 (CVMP) の作業計画の概要を示しています。今後数年間の CVMP の計画された活動と戦略的目標の概要を提供しています。 主な内容は以下の通りです： 1. EMA のウェブサイトで公開されている 2024 年の CVMP 作業計画 2. 現在作成中で、利害関係者との協議の対象となっている獣医学分野の 3 年戦略作業計画 (2025-2027 年) 文書では、抗菌薬耐性への対応、ワクチンの利用可能性の向上、国際協力の促進などの横断的活動が強調されています。また、環境リスク評価や最大残留基準値などの具体的な重点分野についても詳述されています。

	<ul style="list-style-type: none"> - 専門分野： - 環境リスク評価：パロア、抗がん剤、BE、抗コクシジウム薬のガイドライン改訂 - 最大残留基準値 - 抗菌薬耐性： - リスク評価と承認後調査に関するガイドライン - 診断テストに関するリフレクションペーパー - 既存の動物用抗生物質の用量調整 - 免疫学的製剤： - mRNA ワクチンの品質面に関するガイダンスのコンセプトペーパー - プラスミド DNA、生組換えベクターワクチン、複合ワクチンのガイドライン改訂 <p>3. 獣医領域 3年戦略作業計画：</p> <ul style="list-style-type: none"> - 2025年1月から2027年12月までをカバー - 現在構築中 - ステークホルダー協議を予定 - 戦略目標、主要目的、および基本活動を明記 - 7つの戦略目標： <ol style="list-style-type: none"> 1. イノベーション 2. 利用可能性 3. 承認後 4. AMR（抗菌薬耐性） 5. ERA（環境リスク評価） 6. 3Rs（代替、削減、改善） 7. 科学的プロセスとステークホルダー <p>4. 2025-2027年の主要目的と活動：</p> <ul style="list-style-type: none"> - 2024年の多くの活動が継続 - 新しい活動には以下が含まれる可能性： - AMRに関するCVMP戦略2026-2030 - 遺伝子治療に関するガイダンス - コンパニオンアニマルにおけるAMRのリスク評価に関するガイダンス <p>5. ステークホルダーの関与：</p> <ul style="list-style-type: none"> - 22-25週（暫定）に協議を予定 - 26週（暫定）にステークホルダー会議を予定 <p>この文書は、獣医学における抗菌薬耐性への対応、ワクチンの利用可能性の向上、環境リスク評価の強化の重要性を強調しています。また、将来の戦略とガイドラインの形成における国際協力とステークホルダーの関与へのコミットメントも強調しています。</p>	<p>さらに、2025年1月から2027年12月までをカバーする今後の3年戦略作業計画を紹介しています。この計画では、イノベーション、利用可能性、承認後活動、抗菌薬耐性、環境リスク評価、3R原則の実施、利害関係者の関与という7つの戦略目標が概説されます。</p> <p>全体として、この文書は獣医薬品分野における近い将来のCVMPの優先事項と計画された活動に関する貴重な洞察を提供しています。</p>
<p>PARC_AD6.3.summary Answer Translation</p>	<p>この文書は、Partnership for the Assessment of Risks from Chemicals (PARC)プログラムのActivity A6.2.1の下で開発中の戦略の初期段階を示しており、規制の縦割りを超えて化学物質のリスク評価と管理のより統合的なアプローチを提案することを目的としています。主なポイントは以下の通りです：</p> <ol style="list-style-type: none"> 1. 目的： - 一般環境と職業環境からの曝露を組み合わせる知識を進展させる - 複数の源、経路、生活環境からの化学物質曝露を集計する方法を開発する - 効果的なリスク管理に情報を提供するために、異なる曝露源の寄与を決定する <p>2. 主要構成要素：</p>	<p>要旨</p> <p>リスク評価の区分化された見方の文脈において、活動A6.2.1「一般人口と労働者の複数の源と経路からの集計曝露評価」は、一般環境と職業環境からの曝露を組み合わせることに関する知識を進展させることを目的としています。これは、欧州機関および欧州委員会による最近のイニシアチブに沿って、規制の縦割りを超えたより統合的なリスク評価と管理を提案するのに役立ちます。</p> <p>このデリバラブルは、異なる生活環境、源、および経路を通じた集計曝露</p>

<ul style="list-style-type: none"> - モデルとデータのインベントリ - 集計曝露評価に関するガイダンス文書のレビュー - 集計方法の開発 - 方法論を適用し検証するためのケーススタディ <p>3. モデルとデータのインベントリ：</p> <ul style="list-style-type: none"> - 曝露モデル、集計モデル、関連データセットのインベントリを編集 - 一般環境、職業、ソースから用量までの曝露モデルを分類 - 化学物質濃度、曝露因子、生物学的モニタリングなどのデータソースを特定 <p>4. ガイダンスのレビュー：</p> <ul style="list-style-type: none"> - ECHA、EFSA、EMA、SCCS からの規制ガイダンスを集計曝露について分析 - 機関間で集計評価の調和された方法が一般的に欠如していることを発見 - ガイダンスで有用なデータソースとモデリング推奨事項を特定 <p>5. 集計戦略：</p> <ul style="list-style-type: none"> - 確率論的方法を用いて単一ソース曝露モデルの出力を組み合わせることを提案 - モンテカルロシミュレーションを使用して変動性と不確実性を考慮 - 一般環境と職業曝露をリンクする計画 <p>6. ケーススタディ：</p> <ul style="list-style-type: none"> - PFAS、金属、農薬などの優先化学物質グループに方法論を適用 - ケーススタディは一般環境、職業、または組み合わせとして分類 - 方法論の改良と曝露プロファイルの生成を支援 <p>7. 次のステップ：</p> <ul style="list-style-type: none"> - 関連する曝露モデルを選択し接続 - モデル統合のための機能設計を開発 - 2年目にケーススタディを開始してアプローチをテストし改良 - モデル化された集計曝露を生物学的モニタリングデータと比較 <p>この作業は、曝露経路/シナリオ間の比較を容易にし、予防努力の優先順位付けをするためのツールを提供することを目指しています。また、優先化学物質の効果的なリスク管理をサポートするために、ソース寄与の理解を向上させます。</p>	<p>を評価するために開発された戦略とロードマップの最初のステップを提示します。セクション1では、この作業が対処しようとする科学的、社会的、および規制上の質問と、提案されたロードマップの概要を示します。セクション2では、ISES と OECD による以前の作業を基に、一般および職業人口の曝露、源から用量までのモデル、および集計モデルの目録を提示します。曝露モデルに供給するための異なるデータソースが特定されます。セクション3では、様々な生活環境からの集計曝露を評価するための戦略の概要を示します。セクション4では、2年目から始まる方法論を適用し洗練するための提案されたケーススタディを紹介し、さらなる発展、推奨事項、および結果は、プロジェクトの後続の年に出てくる予定です。</p> <p>キーワード：集計曝露、人間曝露、多源曝露、多経路曝露、リスク緩和、職業曝露、生涯曝露、消費者曝露、環境曝露</p>
--	--

4.3. 翻訳機能

4.3.1. 多言語の翻訳実行調査

AWS から提供されている大規模原モデル(LLM)の内、Claude 3.5 Sonnet v1 を利用して、アラビア語、中国語、ロシア語、英語のドキュメントの日本語翻訳を行った。

アラビア語、中国語、ロシア語、英語の PDF ファイルをマークダウン形式(MD)に変換し、この日本語翻訳を行った。実際に翻訳を行ったドキュメントは「241213 第4回業務進捗報告会」フォルダーの「3_翻訳サンプル.zip」の中にある PDF ファイルから変換した md ファイルである。Web インターフェースからの環境では AWS の API Gateway の制限により LLM からのレスポンス時間が長い場合にタイムアウトが発生するため、元のドキュメントを 2048 文字以内の文章に区切って(chunk)翻訳を行った

上で結合した。この chunk 毎の実行時間とドキュメント全体の翻訳に要した時間を測定した。ただし要約と同様に英語以外の言語から直接日本語訳を指示しても英訳することが散見されたため、翻訳でも①英訳、②日本語訳の順番に行った。原稿が英語のドキュメントの場合は、そのままの文章を返す様に LLM に指示を行った。

具体的なプロンプトは下記の通りである。

- ① 以下のドキュメントを英語に翻訳してください。翻訳以外の説明文は不要で、略語はそのままで翻訳してください。英語の文章の場合はそのまま出力してください。
- ② 以下のドキュメントを日本語に翻訳してください。翻訳以外の説明文は不要で、略語はそのままで翻訳してください。日本語の文章の場合はそのまま出力してください。

ドキュメントの文字数と実行(英訳+日本語訳の合計時間)にかかった時間を表 4-3 に示した。英訳、日本語訳結果は「3_翻訳サンプル.zip」の中の末尾がそれぞれ_en.txt と_jp.txt のファイルである。Chunk 毎の実行時間では最大でアラビア語から英語への翻訳で 516 秒かかっている場合があったが、次に時間がかかったケースは 27 秒のため、たまたまの可能性はある(LLM が混んでいた、アメリカとのネットワークに問題が発生していたなど)。中国語の場合は英訳から日本語訳に時間がかかっており、表意文字のため他言語よりも情報量が多いのかもしれない。英語からの日本語訳よりも英語以外の言語の原文からの英訳の方が通常速いが、ロシア語の場合は英訳にやや時間がかかっていた。

表 4-3 ドキュメントの大きさと翻訳実行時間

ファイル名	行数	文字数	実行時間	最大断片実行時間	2 番目実行時間
Arabic.md	4,681	37,638	65m53.755s	25.800s	22.640s
Arabic_m.md	14,107	41,078	92m56.815s	515.887s	27.163s
Chinese.md	3,512	5,249	39m30.230s	68.031s	65.564s
Chinese_s.md	1,516	6,247	17m42.558s	62.250s	60.764s
Russian.md	4,635	33,869	75m9.661s	63.910s	52.251s
Russian_s.md	224	1,043	2m3.668s	18.920s	16.923s
English.md	5,908	23,042	30m48.683s	22.212s	19.400s
English_s.md	64	222	0m27.202s	18.579s	8.266s

アメリカとの通信や LLM の混雑状況など、実行時間の揺らぎが大きく API Gateway のタイムアウト時間が課題となる。そのため 2048 文字の半分の 1024 文字に分割して翻訳を実行することにした。

100 ページ程度のドキュメントの翻訳に 30 分～90 分程度かかっており、web インターフェース上の実行はやや難がある。そこで web インターフェースからは chunk を並列に LLM に投げることで翻訳時間を削減する機能を実装することにした。

4.3.2. 記事作成画面からの翻訳の実行

記事作成画面からの翻訳の実行を図 4-3 に示した。スクレイピングでは記事の翻訳を実行しないことにしたため、元記事の日本語訳が必要な場合はこの画面から翻訳を行い、記事データを保存する必要がある。翻訳は一度英語に訳し、これを日本語訳している。元記事が英語の場合も「翻訳後の文章(英語)」に原文が表示されるが、4.4 節の辞書に登録されている単語・文章には鍵かっこ「」が付与される。この時のプロンプトを表 4-4 に示したが、このプロンプトとなった経緯は 4.4.2 節を参照されたい。

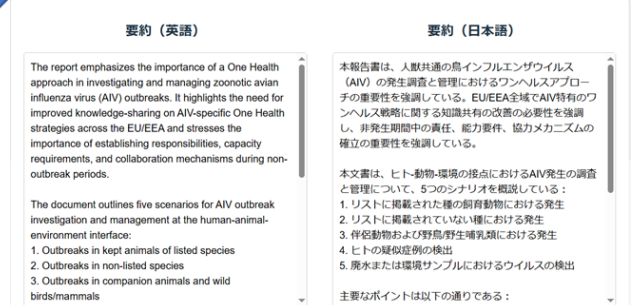
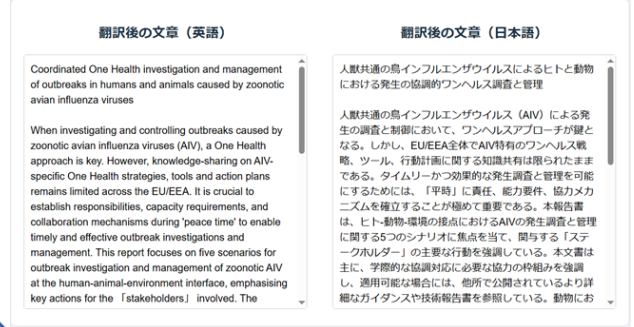


図 4-3 記事作成画面からの翻訳・要約機能の実行

表 4-4 翻訳機能のプロンプト一覧(4.4.2 節参照)

機能	指示文
英語翻訳プロンプト	<p>翻訳対象の文章を英語に翻訳してください。</p> <p>翻訳以外の説明文は不要で、略語はそのまま翻訳してください。英語の文章の場合はそのまま出力してください。</p>
日本語訳プロンプト	<p>翻訳対象の文章を日本語に翻訳してください。</p> <p>ただし「」内は既に日本語に翻訳しているため、その場合のみ「」を含めてそのまま出力してください。</p> <p>翻訳以外の説明文は不要で、略語はそのまま翻訳してください。</p>
お試し翻訳プロンプト	<p>翻訳対象の文章を日本語に翻訳してください。</p> <p>ただし「」内は既に日本語に翻訳しているため、その場合のみ「」を含めてそのまま出力してください。</p> <p>翻訳以外の説明文は不要で、略語はそのまま翻訳してください。</p> <p>語調は「だ・である調」としてください。</p>

4.3.3. お試し翻訳画面作成

記事作成支援機能として、日本語訳のみを行う機能を「お試し翻訳」として提供した。作成した画面と翻訳実行の様子を図 4-4 に示した。尚、「お試し翻訳」のプロンプトは表 4-4 に示した通りであり、

辞書機能は有効になっていない。

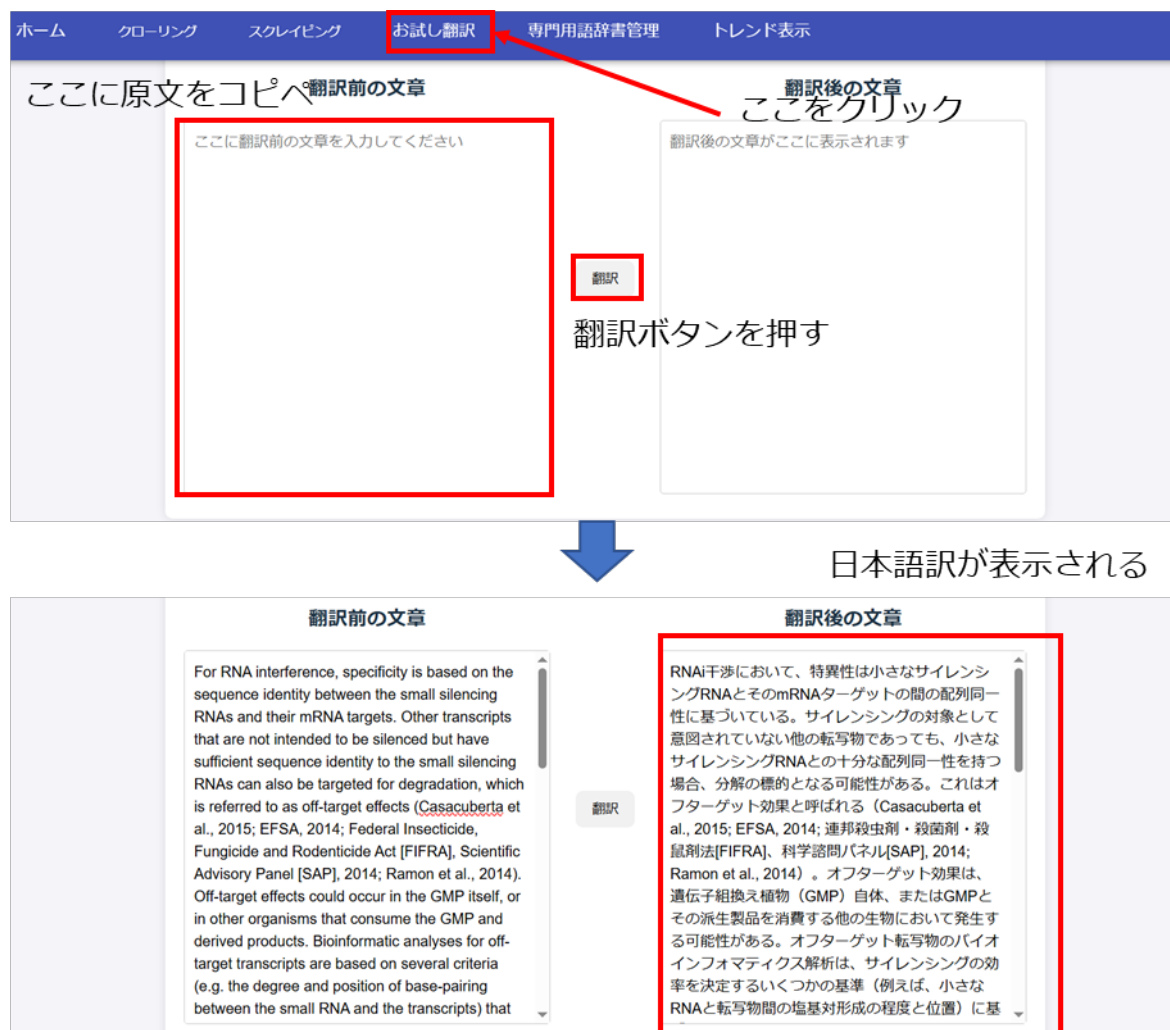


図 4-4 お試し翻訳画面

4.4. 専門用語・定訳辞書機能

4.4.1. 辞書の作成

専門用語辞書として、食品安委員会の web page からダウンロードした「食品の安全性に関する用語集」² (食安委用語集)から辞書の原稿を作成した。食安委用語集は辞書ではないため必ずしも辞書として良い英語と日本語のペアとはならない(例えば「Hot」と「ホット」のペアがある)が、機能確認のみを目的として使用した。

辞書機能は英語の原文に完全一致した単語・文字列を置き換えることで実装したため、先頭文字の小文字化や複数形も置換対象に加えるために、python の複数形を作成するライブラリー-Inflector³ (3.1.1)を組み込んで複数形を自動的に追加することにした。この対応を行った日本語と英語のペアの例を表 4-5 に示した。英語と日本語のペアは 1 対多となる。

表 4-5 大文字・小文字の文字列の追加、及び複数形を加えた後の英語と日本語ペアの例

日本語	英語
-----	----

² <https://www.fsc.go.jp/yougoshu.data/yougoshu.pdf>

³ <https://pypi.org/project/Inflector/>

IARC の発がん性分類	IARC Monographs on the Identification of Carcinogenic Hazards to Humans
IARC の発がん性分類	iarc monographs on the identification of carcinogenic hazards to humans
IARC の発がん性分類	iarc monographs on the identification of carcinogenic hazards to humans
体内動態試験	Absorption, Distribution, Metabolism, Excretions
体内動態試験	absorption, distribution, metabolism, excretions
体内動態試験	Absorption, distribution, metabolism, excretions
体内動態試験	Absorption, Distribution, Metabolism, Excretion
体内動態試験	absorption, distribution, metabolism, excretion
体内動態試験	Absorption, distribution, metabolism, excretion
OECD テストガイドライン	OECD Guidelines for the Testing of Chemicals
OECD テストガイドライン	oecd guidelines for the testing of chemicals
OECD テストガイドライン	Oecd guidelines for the testing of chemicals
介入における防護の最適化	Optimization of Protection in Interventions
介入における防護の最適化	optimization of protection in interventions
介入における防護の最適化	Optimization of protection in interventions

4.4.2. 辞書機能を加えた翻訳方法

専門用語・定訳辞書機能の実装として、辞書ファイルに登録されている単語を置き換えて LLM に翻訳させる機能を試した。登録されている文字列を日本語で「」付きで置き換えることにより、LLM に翻訳対象外であることを知らせている。この時下記のプロンプトで試した。

- ①以下のドキュメントを英語に翻訳してください。翻訳以外の説明文は不要で、略語はそのまま翻訳してください。英語の文章の場合はそのまま出力してください。
- ②以下のドキュメントを日本語に翻訳してください。ただし「」内は既に日本語に翻訳しているため、「」を含めてそのまま出力してください。翻訳以外の説明文は不要で、略語はそのまま翻訳してください。

単語・文章の置き換えは、英語の文字列が長い順にソートし、この順番でマッチした場合に「」付き日本語で置き換えた。これは例えば「Risk」と「Risk Assessment」の両方が辞書にある場合の対応である。この変換前後のドキュメントの例を図 4-5 及び図 4-6 に示した。

3. Risk assessment template and examples

Template

You can use a **risk assessment** template to help you keep a simple record of:

- who might be harmed and how
- what you're already doing to control the **risks**
- what further action you need to take to control the **risks**
- who needs to carry out the action
- when the action is needed by

Risk assessment template (Word Document Format) (.docx)

Risk assessment template (Open Document Format) (.odt)

Example risk assessments

These typical examples show how other businesses have managed **risks**. You can use them as a guide to think about:

- some of the **hazards** in your business
- the steps you need to take to manage the **risks**

Do not just copy an example and put your company name to it as that would not satisfy the law and would not protect your employees. You must think about the specific **hazards** and controls your business needs.

図 4-5 翻訳対象英語原稿の例

3. 「リスク評価」 template and examples

Template

You can use a 「リスク評価」 template to help you keep a simple record of:

- who might be harmed and how
- what you're already doing to control the 「リスク」
- what further action you need to take to control the 「リスク」
- who needs to carry out the action
- when the action is needed by

「リスク評価」 template (Word Document Format) (.docx)
「リスク評価」 template (Open Document Format) (.odt)

Example 「リスク評価」

These typical examples show how other businesses have managed 「リスク」. You can use them as a guide to think about:

- some of the 「ハザード」 in your business
- the steps you need to take to manage the 「リスク」

Do not just copy an example and put your company name to it as that would not satisfy the law and would not protect your employees. You must think about the specific 「ハザード」 and controls your business needs.

図 4-6 英語原稿に専門用語・定訳辞書の置き換えを行った場合

4.3.1 節のドキュメントの一部に加えて英語記事の翻訳を、英語辞書機能を加えて実行した場合の英訳+日本語訳にかかった時間を表 4-6 に示した。表 4-3 と同じドキュメントについては実行時間を比較することが出来るが、大きな違いは無い様である。

図 4-6 の LLM よるに翻訳結果を図 4-7 に示した。専門用語・定訳辞書の置き換えを行った場合の翻訳は指示文通りに機能している様である。

表 4-6 ドキュメントの大きさと辞書機能付き翻訳実行時間

ファイル名	行数	文字数	実行時間	最大断片実行時間	2 番目実行時間
Arabic.md	4,681	37,638	62m45.877s	30.254s	25.848s
Chinese_s.md	1,516	6,247	15m27.679s	53.363s	51.762s
Russian_s.md	224	1,043	1m28.585s	13.458s	11.709s
English.md	5,908	23,042	30m15.395s	28.886s	21.731s
PARC_AD6.3.md	4,413	40,369	52m49.688s	19.018s	17.458s
bambermycin.md	50	219	0m17.068s	10.016s	6.698s

3. 「リスク評価」 テンプレートと例

テンプレート

「リスク評価」テンプレートを使用して、以下の簡単な記録を保持できます：

誰がどのように危害を受ける可能性があるか

「リスク」を制御するために既に行っていること

「リスク」を制御するためにさらに必要な対策

誰が対策を実行する必要があるか

対策がいつまでに必要か

「リスク評価」テンプレート (Word 文書形式) (.docx)

「リスク評価」テンプレート (OpenDocument 形式) (.odt)

「リスク評価」の例

これらの典型的な例は、他の企業が「リスク」をどのように管理しているかを示しています。以下について考える際のガイドとして使用できます：

あなたの事業における「ハザード」の一部

「リスク」を管理するために必要な手順

例をそのままコピーして会社名を記入するだけでは法律を満たさず、従業員を保護することもできません。あなたの事業に特有の「ハザード」と対策について考える必要があります。

図 4-7 専門用語・定訳辞書の置き換えを行ったドキュメントの日本語翻訳例

4.4.3. 辞書登録画面

辞書データ登録・表示画面を図 4-8 に示した。登録項目は削除ボタンをクリックすることで行う。日本語と英語のペアの CSV ファイルをアップロードして辞書の上書きをする様子を図 4-9 に示した。アップロードするファイルは日本語、英語の順で且つ UTF-8 の文字コードである必要がある。

The screenshot shows a web interface for dictionary management. At the top, there are navigation links: ホーム, クローリング, スクレイピング, お試し翻訳, 専門用語辞書管理, and トレンド表示. The main content area is titled '辞書ファイルアップロード' and includes a 'ファイルを選択' button with the text '選択されていません'. Below this is a 'プレビュー' section with a '辞書登録' button. The bottom section is titled '登録済み辞書一覧' and contains a table with three columns: '日本語', '英語', and '削除'.

日本語	英語	削除
IARCの発がん性分類	IARC Monographs on the Identification of Carcinogenic Hazards to Humans	削除
IARCの発がん性分類	IARC Monographs on the Identification of Carcinogenic Hazards to Humans	削除
IARCの発がん性分類	IARC Monographs on the Identification of Carcinogenic Hazards to Humans	削除
体内動態試験	Absorption, Distribution, Metabolism, Excretions	削除
体内動態試験	absorption, distribution, metabolism, excretions	削除
体内動態試験	Absorption, distribution, metabolism, excretions	削除
体内動態試験	Absorption, Distribution, Metabolism, Excretion	削除

図 4-8 辞書データ登録・表示画面

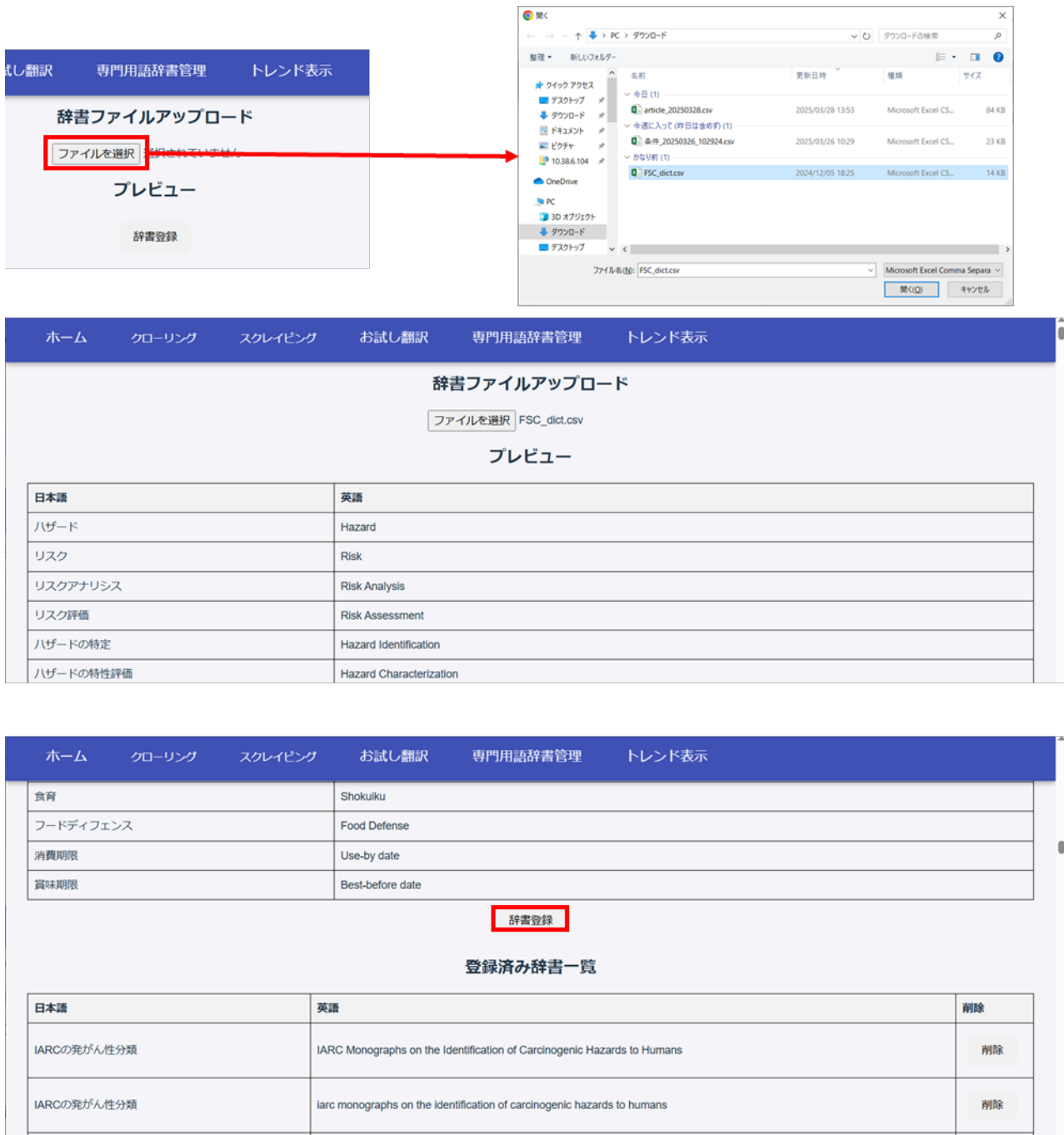


図 4-9 辞書データ CSV ファイルのアップロードと辞書登録

4.4.4. 辞書機能を有効にした翻訳

辞書機能は記事作成画面から翻訳を行った場合に有効になる様にした。この時の翻訳結果の例を図 4-10 に示した。辞書機能により置換対象となった英語単語・文章は鍵かっこ付きで表示され、対応する日本語訳も鍵かっこ付きで表示される。

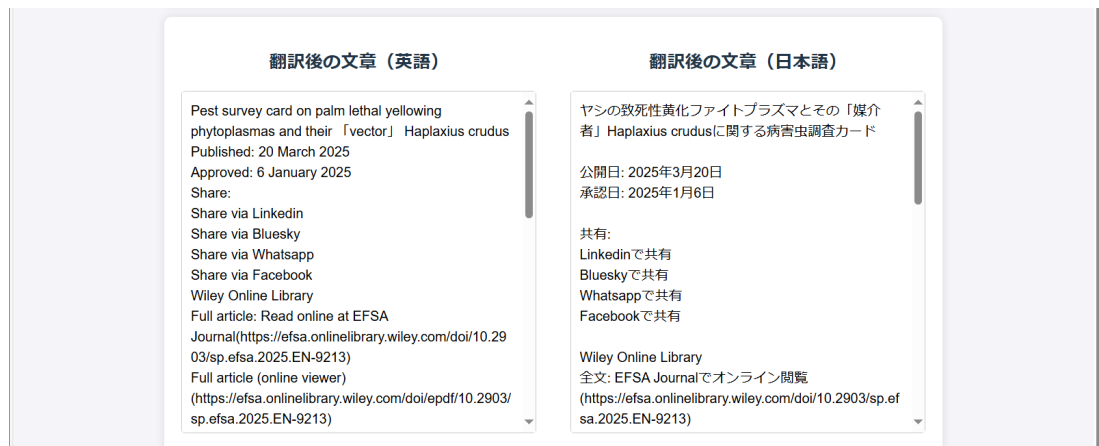


図 4-10 記事作成画面から辞書機能を有効にした時の翻訳結果表示例。辞書機能により「vector」が「媒介者」と訳されている。

4.5. 記事編集機

記事作成を行う画面を図 4-11 記事作成画面に示し、この項目の説明を表 4-7 に示した。この画面からの入力必須項目は「ソース URL」であり、記事データベース上の ID に利用される。他の項目への入力は任意であり、下部の記事が空でもデータベースに保存することが出来る。

原文の欄に記事原稿をコピー&ペーストすると LLM による翻訳、要約を実行することが出来るが、直接翻訳及び要約を入力しても良い。

スクレイピングから記事データベースに登録した記事情報を表示させた場合の記事編集画面を図 4-12 に示した。スクレイピングでは様々な情報が LLM により自動で付与されており、この説明を表 4-8 に示した。必要であればこの画面から編集して記事情報を保存する。

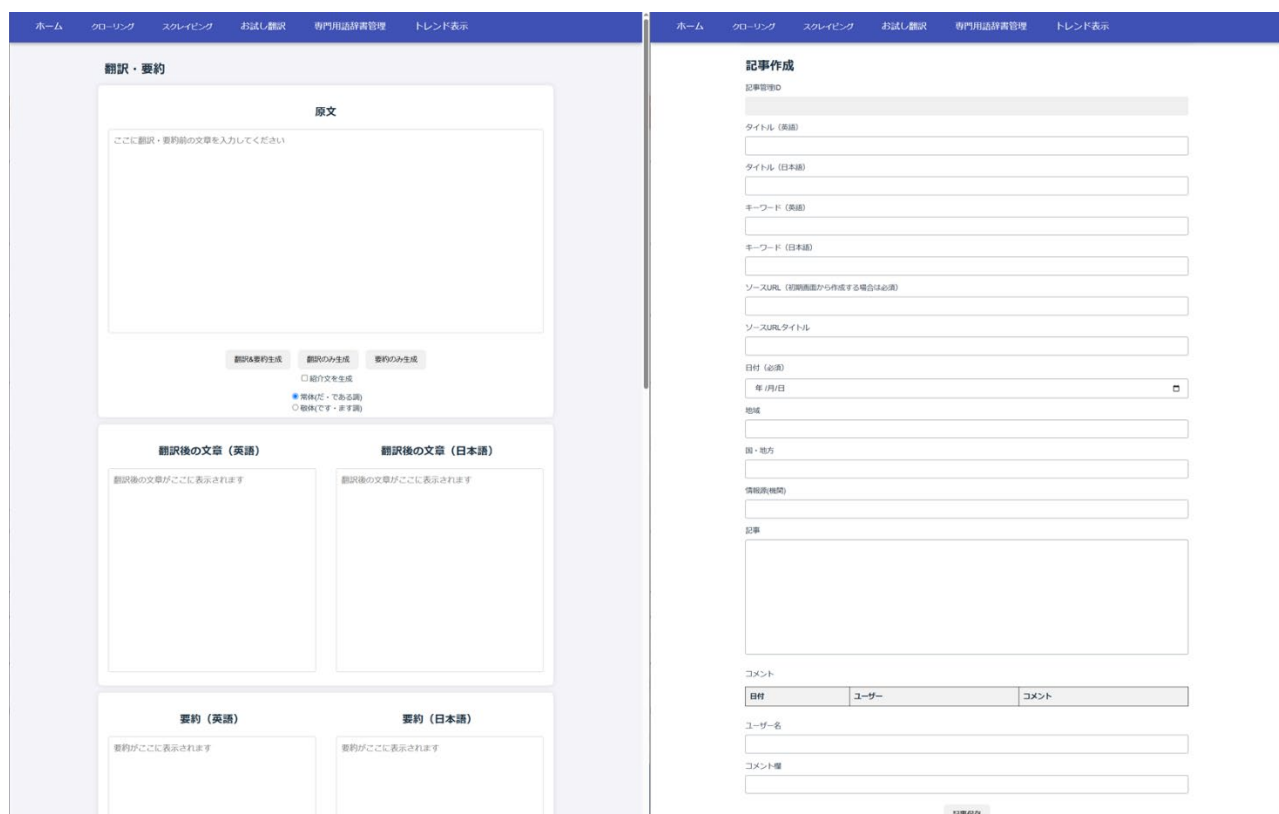


図 4-11 記事作成画面

表 4-7 記事作成画面における項目の説明

項目名	説明
原文	テキスト化した元記事。コピペで入力すると翻訳、要約を実行することが出来る。
翻訳後の文章(英語)	元記事の英訳。辞書による置換対象は鍵かっこ付きで表示される。
翻訳後の文章(日本語)	「翻訳後の文章(英語)」の日本語訳
要約(英語)	英語要約を入力。もしくは LLM に要約させる。
要約(日本語)	「要約(英語)」の日本語訳
記事管理 ID	記事データベースで使用される ID。記事保存後に自動作成されるため入力不可。
タイトル(英語)	英語の記事のタイトル
タイトル(日本語)	日本語の記事のタイトル
キーワード(英語)	英語のキーワード。カンマ区切りで入力。
キーワード(日本語)	日本語のキーワード。カンマ区切りで入力。
ソース URL	記事の URL を入力。必須項目であり、この URL から記事管理 ID が作成される。
ファイル名	手入力の記事を作成した場合は使用しない
資料日付	日付情報。元記事が公開された日付を想定。
地域	食安委で定義している地域情報
国・地方	食安委で定義している国・地方情報
情報源(機関)	食安委で定義している情報源(機関)情報
記事	記事の内容
コメント	保存されたコメントを表示
ユーザー名	コメントに表示されるユーザー名を入力する。
コメント欄	コメント内容を入力する。

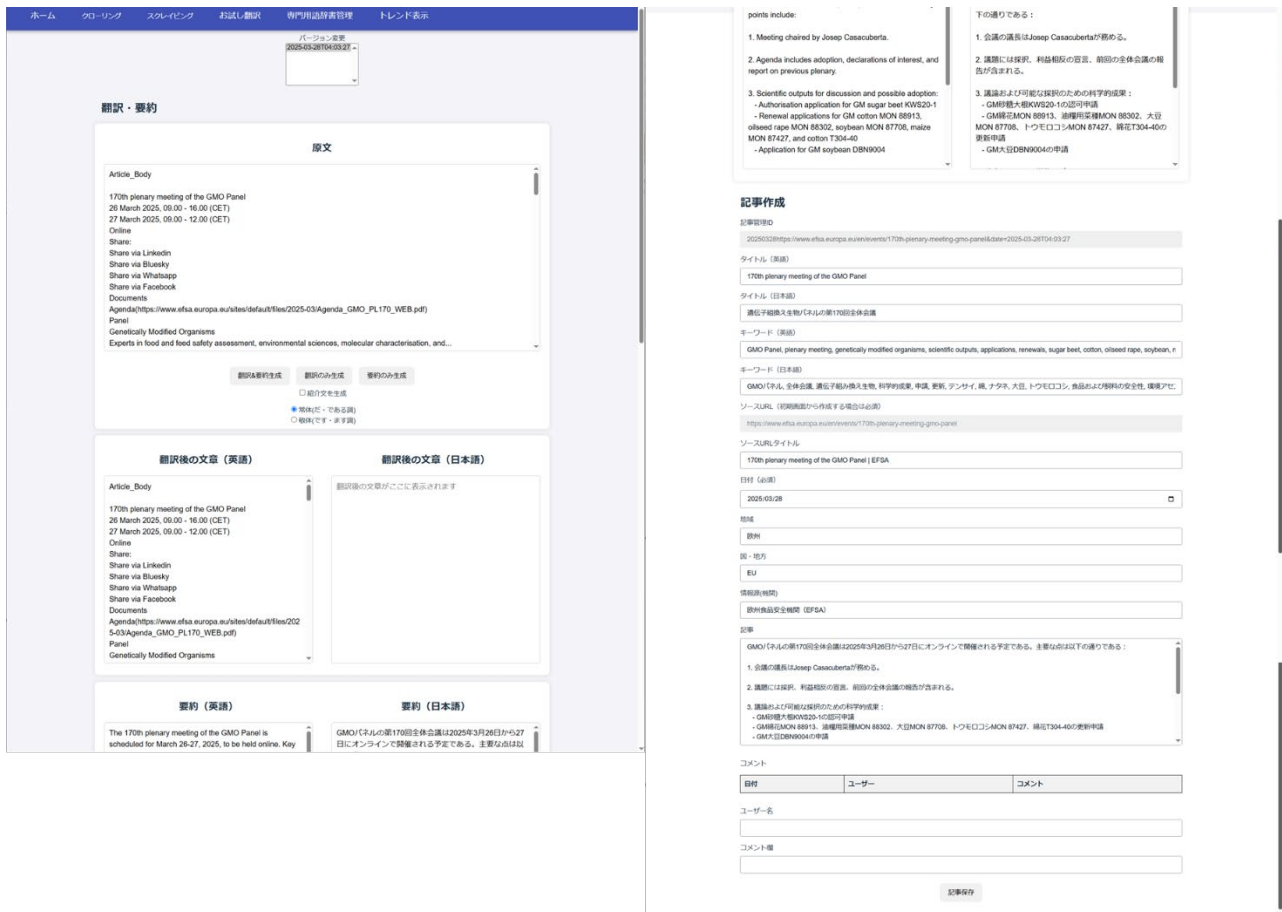


図 4-12 スクレイピングから保存された記事の表示例

表 4-8 スクレイピングで登録される記事データベースの項目内容の説明

項目名	説明
原文	元記事。PDF ファイルから抽出した文字列もしくは HTML ファイルから得られたテキスト部分。
翻訳後の文章(英語)	LLM による元記事の英訳。辞書による置換対象は鍵かっこ付きで表示される。
翻訳後の文章(日本語)	スクレイピングでは付与しない。必要であれば翻訳を実行する。
要約(英語)	LLM による英語要約
要約(日本語)	「要約(英語)」の日本語訳
記事管理 ID	記事データベースで使用される ID (編集不可)
タイトル(英語)	LLM に抽出させた記事のタイトル
タイトル(日本語)	「タイトル(英語)」の日本語訳
キーワード(英語)	LLM に抽出させたキーワード
キーワード(日本語)	「キーワード(英語)」の日本語訳
ソース URL	取得した PDF ファイルもしくは HTML ファイルの URL
ファイル名	ダウンロードした元記事を保存したファイル名
資料日付	スクレイピングで記事を取得した日付
地域	食安委で定義している地域情報
国・地方	食安委で定義している国・地方情報
情報源(機関)	食安委で定義している情報源(機関)情報
記事	初期状態は要約(日本語)の内容

5. 情報の保存と活用

5.1. クローリング結果 DB

クローリング結果はクローリング結果 DB に格納される。画像ファイルや PDF ファイルなどはストレージに保存され、このファイルのパス情報がクローリング結果 DB に保存される。

5.2. 専門用語・定型辞書 DB

4.4.3 節の辞書登録画面からアップロードされた辞書原稿は、下記の加工を行って専門用語・定型辞書 DB に保存される。

1. 複数形を追加
2. 先頭文字が小文字の場合、大文字にした文字列を追加
3. 大文字が含まれる文字列の場合、全て小文字にした文字列を追加
4. 英語の文字列長が長い順番にソート

辞書機能付き翻訳では、この処理を行った辞書情報を提供し、4.4.2 節で検討した方法で辞書機能を実現している。

5.3. 翻訳・要約結果 DB

翻訳・要約結果は記事 DB に統合した。

5.4. 記事 DB

記事 DB に格納している情報を表 5-1 に示した。「食安委既存データ項目」は互換性のために残しており、基本的には使用していない。

表 5-1 記事 DB のデータ項目

項目	説明
note_id	記事管理 ID。URL から自動的に作成される
note_datetime	資料日付
AREACODE	食安委既存データ項目
ATTACHEDFILE	食安委既存データ項目
BIBLNO	食安委既存データ項目
BULLETINCODE	食安委既存データ項目
CHECKFLG_RAISINGDATE	食安委既存データ項目
CLASS1CODE	食安委既存データ項目
CLASS2CODE	食安委既存データ項目
classification	分類情報
classification1	分類 1
classification2	分類 2
comments	コメント情報
CONTENTS	食安委既存データ項目
crawling_condition_file_path	クローリング条件ファイルのパス
crawling_condition_url	クローリング対象 URL
crawling_date	クローリング実行日
crawling_frequency	クローリング頻度。日(毎日)など。

data_format	データ形式。html や pdf など。
data_path	保存したファイルのパス
depth	クローリング深度
DISCLOSURELEVEL	食安委既存データ項目
DOCDATE	食安委既存データ項目
EDITDATE	食安委既存データ項目
EDITORID	食安委既存データ項目
encoded_url	エンコードした URL
error_message	エラーメッセージ
hash_url	URL のハッシュ値
HAZARDCODE	食安委既存データ項目
id	日付情報から作成した ID
INFOSOURCE	食安委既存データ項目
INFOURL	食安委既存データ項目
key_words	英語キーワード(スクレイピングでは LLM 抽出)
key_wordsJa	日本語キーワード(スクレイピングでは key_words の翻訳)
LLM 英語要約	LLM による英語要約(保存用)
LLM 日本語要約	LLM による英語要約英語要約の日本語訳(保存用)
LOCALCODE	食安委既存データ項目
MEDIACODE	食安委既存データ項目
noteText	記事作成画面に表示される記事
PUBLISHDATE	食安委既存データ項目
PUNCHERID	食安委既存データ項目
RECEIPTDATE	食安委既存データ項目
RECEIPTNO	食安委既存データ項目
REMARKS	食安委既存データ項目
response_status	スクレイピングによる取得の成功もしくは失敗
response_status_code	HTTP のレスポンスコード
REVISIONDATE	食安委既存データ項目
REVISIONID	食安委既存データ項目
SHELFCODE	食安委既存データ項目
status	記事のステータス。スクレイピングからは全て new になる。
summaryEn	記事作成画面に表示される要約(英語)
summaryJa	記事作成画面に表示される要約(日本語)
target_area_code	食安委指定エリアコード
target_area_name	食安委指定エリア名
target_bullentin_code	食安委指定公報コード
target_country_name	食安委指定国名
target_local_code	食安委指定地域コード
target_organization_name	食安委指定組織名
target_url	URL
target_url_page_publication_data	公開元に記載されている公開日
target_url_page_title	HTML ファイルに記載されているタイトル
TEMPRECEIPTNO	食安委既存データ項目

text_info_body	テキスト化された元記事
TITLE	元記事タイトル
translateEn	英語翻訳(スクレイピングでは辞書登録文字列は鍵かっこ付き)
translateJa	日本語翻訳
URL パス	食安委既存データ項目
WEEKLYNO	食安委既存データ項目
WEEKLYVOLNO	食安委既存データ項目
xBIBLNO	食安委既存データ項目
アイテムの種類	食安委既存データ項目
アプリの作成者	食安委既存データ項目
アプリの変更者	食安委既存データ項目
エンコードされた絶対 URL	食安委既存データ項目
コンテンツ タイプ	食安委既存データ項目
コンプライアンス資産 ID	食安委既存データ項目
タイトル	食安委既存データ項目
パス	食安委既存データ項目
ファイルの種類	食安委既存データ項目
ワークフロー インスタンス ID	食安委既存データ項目
記事 ID 検索キー	食安委既存データ項目
更新者	食安委既存データ項目
更新日時	食安委既存データ項目
色タグ	食安委既存データ項目
登録者	食安委既存データ項目
登録日時	食安委既存データ項目

5.5. ラベル付与機能

5.5.1. LLM によるラベル抽出の検討

M_BIBLIOGRAPHIC.xlsx に記載の最後の 100 件の記事について LLM を使った情報抽出方法の検討を行った。

AWS から利用することができる Claude 3.5 Sonnet を利用して

食品添加物名	物質名	医薬品	農薬名	肥料名	毒物名	食品名	材料名
農作物名	工業製品名	商品名	疾患名	動物種	植物種	細菌	ウイルス

の項目について抽出できるのかどうかの検討を行った。具体的な質問形式の例を図 5-1 に示した。

<p>[記事]以下のドキュメントから下記の[質問]に回答してください。誤字脱字、途中で切れていることがあるため、内容を予測しながら質問に回答してください。</p> <p>[質問] 食品添加物名が含まれていれば名前のみを抽出してください。含まれていなければ NA とのみ答えてください。補足情報等は回答に加えないでください。</p> <p>[記事] 欧州食品安全機関(EFSA)は 8 月 8 日、肥育用豚、肥育用のマイナー豚種及び肥育用七面鳥に使用する飼料添加物(<i>Trichoderma reesei</i> MUCL 49755 株により生産されるエンド-1,4-β-キシラナーゼ及び <i>T. reesei</i> MUCL 49754 株により生産されるエンド-1,3(4)-β-グルカナーゼからなる(AveMix(登録商標)XG 10))の認可更新の評価に関する科学的意見書(7 月 3 日採択、PDF 版 7 ページ、doi: https://doi.org/10.2903/j.efsa.2024.8951)を公表した。概要は以下のとおり。 欧州委員会の要請を受</p>
--

け、EFSA は、肥育用豚、肥育用のマイナー豚種、肥育用七面鳥に使用する畜産飼料添加物としての *Trichoderma reesei* MUCL 49755 株により生産されるエンド-1,4-β-キシラナーゼ及び *T. reesei* MUCL 49754 株により生産されるエンド-1,3(4)-β-グルカナーゼからなる飼料添加物(AveMix(登録商標)XG 10/AveMix(登録商標)XG 10 L)の認可更新の評価に関する科学的意見書の提出を求められた。申請者は、AveMix(登録商標)XG 10 の担体原料を大豆粕から炭酸カルシウム+小麦粉又は炭酸カルシウム+セピオライトに変更することを申告した。申請者は、炭酸カルシウム+小麦粉配合の添加物 AveMix(登録商標)XG 10 及び AveMix(登録商標)XG 10 L が認可条件に適合しているエビデンスを提出した。EFSA の動物用飼料に使用される添加物及び製品又は物質に関する科学パネル(FEEDAP パネル)は、炭酸カルシウム+セピオライト配合の AveMix(登録商標)XG 10 が認可条件に適合していることを裏付けるデータが提出されていないと指摘した。同パネルは、肥育用豚、肥育用のマイナー豚種、肥育用七面鳥、消費者、環境に対して、いずれの形態の添加物も安全であると結論した。当該認可更新の枠組みにおいては、同添加物の有効性に関する評価は不要である。

図 5-1 食品添加物名の抽出を試みた LLM への質問文の例

この結果は「241129 第 3 回業務進捗報告会」フォルダーの「3_ラベル付与検討資料.xlsx」の AW~BL 列にまとめた。グレーのセルは抽出に失敗している項目であり、分類違いや検討が必要と思われる部分を赤字で示した。分類の抽出結果の精度は高くないが、全く使えないというレベルでもなく、使用方法が難しいと思われる。100 件の範囲であるが、下記のような特徴が見受けられた。

1. 食品添加物名はほとんど取れない。
2. 肥料名は取れない。
3. 材料名には食品が入ってしまう。
4. 菌類では英語になっている場合があるが、かなり正確に取れている。
5. 今回の項目では全て NA となってしまう記事がある。

5.5.2. LLM によるキーワード抽出の検討

LLM によるラベル抽出は、項目によって精度が異なり、また LLM の仕様から毎回異なる結果となることがある。そこで比較的安定していると思われる LLM によるキーワード抽出の検討を行った。

AWS から利用することができる Claude 3.5 Sonnet を利用して記事からキーワードの抽出を行った。質問文の例を図 5-2 に示したが、図 5-1 とは[質問]が異なるだけである。

[記事]以下のドキュメントから下記の[質問]に回答してください。誤字脱字、途中で切れていることがあるため、内容を予測しながら質問に回答してください。

[質問]
キーワードをカンマ区切り形式で抽出してください。補足情報等は回答に加えなくてください。

[記事]
欧州食品安全機関(EFSA)は 8 月 8 日、肥育用豚、肥育用のマイナー豚種及び肥育用七面鳥に使用する飼料添加物(*Trichoderma reesei* MUCL 49755 株により生産されるエンド-1,4-β-キシラナーゼ及び *T. reesei* MUCL 49754 株により生産されるエンド-1,3(4)-β-グルカナーゼからなる(AveMix(登録商標)XG 10))の認可更新の評価に関する科学的意見書(7 月 3 日採択、PDF 版 7 ページ、doi: <https://doi.org/10.2903/j.efsa.2024.8951>)を公表した。概要は以下のとおり。 欧州委員会の要請を受け、EFSA は、肥育用豚、肥育用のマイナー豚種、肥育用七面鳥に使用する畜産飼料添加物としての *Trichoderma reesei* MUCL 49755 株により生産されるエンド-1,4-β-キシラナーゼ及び *T. reesei* MUCL 49754 株により生産されるエンド-1,3(4)-β-グルカナーゼからなる飼料添加物(AveMix(登録商標)XG 10/AveMix(登録商標)XG 10 L)の認可更新の評価に関する科学的意見書の提出を求められた。申請者は、AveMix(登録商標)XG 10 の担体原料を大豆粕から炭酸カルシウム+小麦粉又は炭酸カルシウム+セピオライトに変更することを申告した。申請者は、炭酸カルシウム+小麦粉配合の添加物 AveMix(登録商標)XG 10 及び AveMix(登録商標)XG 10 L が認可条件に適合しているエビデンスを提出した。EFSA の動物用飼料に使用される添加物及び製品又は物質に関する科学パネル

(FEEDAP パネル)は、炭酸カルシウム+セピオライト配合の AveMix(登録商標)XG 10 が認可条件に適合していることを裏付けるデータが提出されていないと指摘した。同パネルは、肥育用豚、肥育用のマイナー豚種、肥育用七面鳥、消費者、環境に対して、いずれの形態の添加物も安全であると結論した。当該認可更新の枠組みにおいては、同添加物の有効性に関する評価は不要である。

図 5-2 キーワード抽出の質問文の例

この結果を「241129 第3回業務進捗報告会」フォルダーの「3_ラベル付与検討資料.xlsx」の BM 列にまとめた。発信元組織や材料などが抽出されている。ただしやや一般用語である

- 小児科医
- 冷蔵庫
- 評価
- 食事,居住地,年齢,体重
- 返却プログラム
- 透明性

の様なキーワードも抽出していた。

項目ごとラベルの抽出よりもキーワード抽出の方が現実的と思われ、本事業では LLM によるキーワード抽出結果を解析機能に用いることにした。

5.6. ラベル DB

記事 DB に統合。

5.7. 記事情報参照機能

記事情報参照機能はスクレイピング記事一覧表示画面に統合した。期間、機関による絞り込みの他に、タイトル(英語、日本語両方)、元記事、記事についての検索機能がある。図 3-7 の左端のボタンをクリックすることにより、選択した1レコードの情報を表示、編集することが出来る記事作成画面に遷移する。

5.8. クロス表表示機能

記事からの LLM を使った項目抽出の精度の低さから、本機能はキーワードの頻度表示機能を構築することになった。この画面を図 5-3 に示した。

ホーム クローリング スクレイピング お試し翻訳 専門用語辞書管理 トレンド表示					
トレンド表示					
キーワード					回数
EFSA					228
EPA					134
risk assessment					95
European Commission					70
European Food Safety Authority					53
food safety					51
environmental protection					46
public health					40
dietary exposure					36
PFAS					36
plant health					34
food enzyme					29
pesticide					29
pesticides					27
scientific opinion					25
pest report					24
EU candidate priority pests					24
working group					23
quarantine pests					23
pest prioritisation					23

図 5-3 キーワード頻度解析を行った画面

5.9. 印刷機能

本システムは web ブラウザー上で動作するため、ブラウザーの印刷機能を利用して印刷を行うことが出来る。

6. 連携自動化

6.1. パブリッククラウド上でのシステム構築

本業務における全てのシステム基盤にはパブリッククラウドサービスである AWS 上に構築した。このためクローリングからの結果登録、スクレイピングからの記事 DB 登録が連携されている。また翻訳・要約についても AWS から提供している LLM を用いることにより、データベースへの保存も AWS 内で全て行っている。

ユーザーインターフェースを記述したファイルは AWS の s3 上に置かれており、web ブラウザーからアクセスすることが出来る。本システムの AWS 上の構成図を図 6-1 に示した。

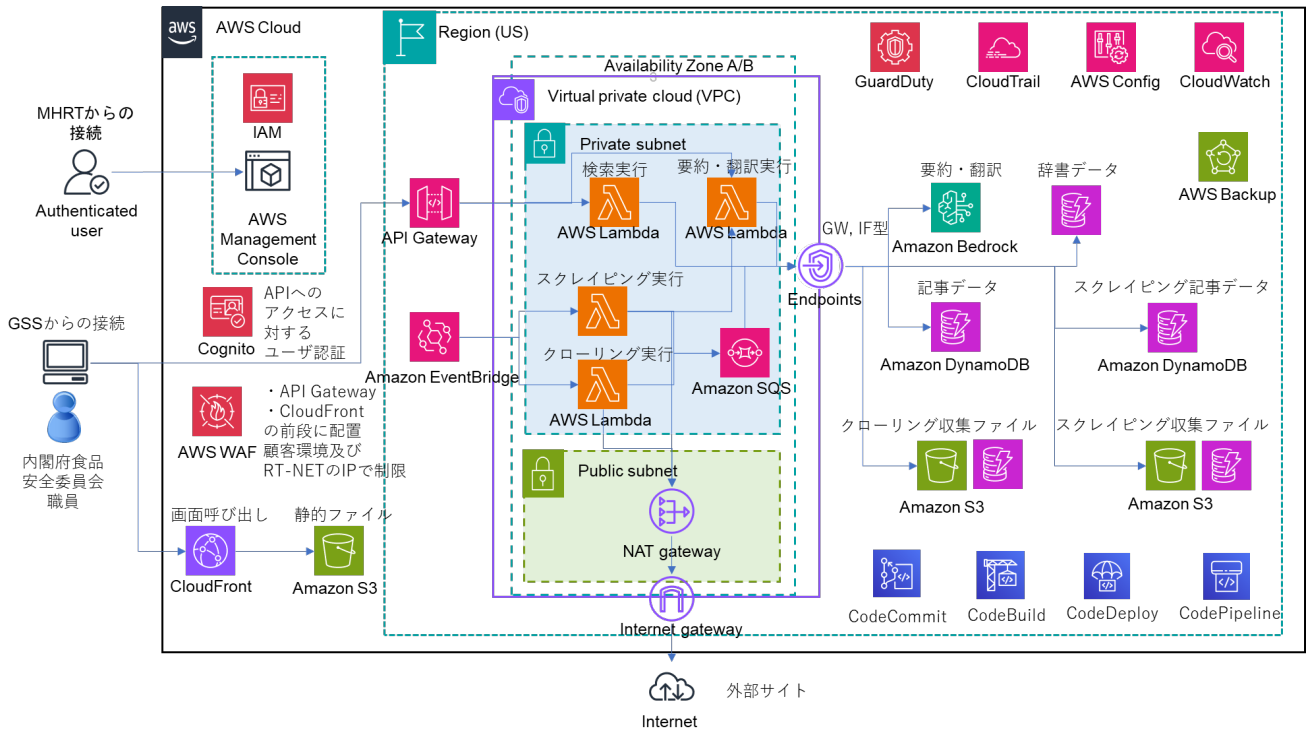


図 6-1 本システムの AWS 上の構成図

7. 保守

アジャイル開発のため保守と開発は一体化している。アジャイル開発のドキュメントは「B)システム開発(ア)アジャイル開発で生じるドキュメント.docx」にまとめた。

8. IT・AI を活用した食品安全情報収集システムにおけるまとめ

クローリングによる情報収集から記事データの解析まで、一連のプロトタイプシステムを構築し、課題の抽出を行った。これらの課題を下記にまとめた

8.1. スクレイピングの必要性

クローリングを実装し実際に情報収集を行ったが、雑多な情報を数多く含んでしまいこれを人が目視で選択する作業が生じてしまった。この選択作業を行うくらいであれば、人が見やすい様に構成されている元の web ページから探索した方がやりやすい。そのためクローリングでは作業の効率化に繋がらず、スクレイピングが必要であることが分かった。

8.2. スクレイピングのコスト対効果

スクレイピングを実行するには、スクレイピング対象 web サイト毎にどの部分の情報を抽出するのかが設定する必要があり、構築には費用がかかる。またスクレイピングの状況をモニタリングし、失敗しているケースの解析を行うなど、保守メンテナンスが必要となる。本事業期間中にも web ページの構成が変わりスクレイピングプログラムを修正したり、想定外に大量の大きなファイルが公開されてスクレイピングで収集漏れが発生したりするなどの事象が発生した。

スクレイピングの費用対効果を考えると、ある程度の数を記事化するサイトに限定した開発・運用が求められると思われる。

8.3. プログラムからのアクセス禁止サイト

Distributed denial of service (DDoS)攻撃への対策や、LLM 等を用いた 2 次サイトの構築対策としてロボットによるアクセス禁止サイトは今後も増えていく可能性がある。コストをかけてスクレイピングを構築しても、プログラムからのアクセスが禁止されてしまうと記事を収集出来なくなってしまう。

8.4. LLM による翻訳・要約

本事業では AWS から提供されている Claude 3.5 Sonnet v1 を主に用いたが、GPT-4o や Gemini 1.5 Pro でも要約・翻訳は実用レベルで可能であると思われる。ただし食品安全情報の記事として、どのような文体、要約が適しているのかを協議するまでは検討できなかった。本事業中にも前述の LLM は全てバージョンアップがリリースされており、LLM の選択やプロンプトの最適化は今後の課題である。

9. ドキュメントからの構造化抄録データベースの構築

9.1. 構造化抄録データベースの目的と課題

記事から指定した項目の情報をデータベース化することにより、今後高度な解析作業を容易にすることが出来ると考えられる。例えばヒトの臨床研究における論文の構成については、

- Introduction, Methods, Results And Discussion (IMRAD)
- Structured Abstracts (構造化抄録)の8項目: 目的、研究デザイン、セッティング、対象患者、介入、主なアウトカム評価、結果、結論

などの指針があり、システムティックレビューやメタ解析などを行いやすい構成が求められている。これらは論文誌ごとに詳細は異なるが研究者が論文に記載すべき項目であり、論文の評価をしやすくするために項目が整理されている。逆に言えば論文がアクセプトされるにはこれらの構造化がされている必要がある。

一方で食品安全に関する領域(食品安全領域)では、化学物質の毒性研究、化学物質のばく露研究、感染症、規制など純粋なサイエンスからルール策定まで幅が広い。そのため食品安全領域で構造化抄録データベースを構築するには、

A) 記事カテゴリの洗い出し

B) 記事カテゴリ毎の抽出項目設定

が必要になる。これらの情報を整理した後に、技術的な課題として

① 記事カテゴリの自動設定

② 項目の自動抽出

が挙げられる。ここでは記事カテゴリが同一であり、抽出項目が設定されている場合について、現状の技術を用いた検証を行った。

9.2. 項目の自動抽出の検証

9.2.1. EFSA chemical contaminants in food and feed のスクレイピングのからの抽出

9.2.1.1. 実施事項

化学物質のリスク評価のドキュメントにおいて、構造化の抽出項目として

1. 対象物質
2. 対象動物
3. ハザードの概要
4. 毒性に関する情報
5. ばく露に関する情報
6. 他機関の評価情報
7. 評価結果
8. 評価の限界や今後の課題

を定めた。この項目について EFSA の chemical contaminants in food and feed (図 9-1)からスクレイピングで取得した記事から情報抽出を行った。ここで EFSA の記事において「Events」はリスク評価のドキュメントでないため「News」と「Scientific output」の記事を対象とし、期間を2024年2月から2025年2月の12記事を対象とした。対象記事の一覧を表9-1に示した。スクレイピングで取得したhtmlファイル(図9-2参照)をテキストに変換し、この文字列を加えて表9-2に示した質問文(上段)を英訳(下段)して大規模言語モデル(LLM)に問い合わせることにより情報抽出を行った(英語の回答を日本語に翻訳)。LLMはClaude 3.5 Sonnet v1、GPT-4o、Gemini 1.5 Proの3つを試した。この結果を

「250305 構造化項目抽出結果報告」フォルダーの「ChemicalContaminantsScraping.xlsx」に示した。

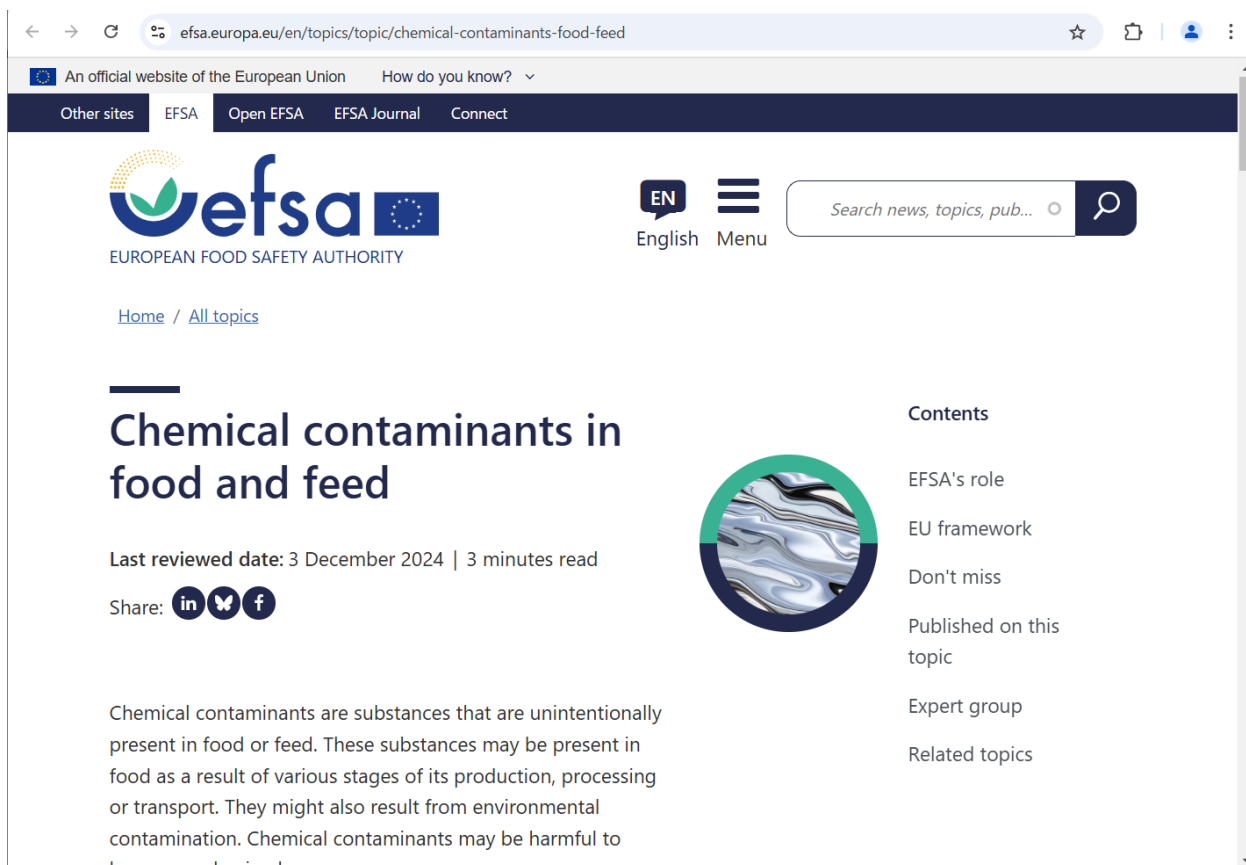


図 9-1 EFSA の chemical contaminants in food and feed

表 9-1 2024 年 2 月から 2025 年 2 月までの EFSA chemical contaminants in food and feed の News と Scientific output の記事一覧

date	title
20240306	Report for 2022 on the results from the monitoring of veterinary medicinal product residues and other substances in live animals and animal products
20240306	Veterinary drug residues in animals and foods remain low
20240312	Risks for animal and human health related to the presence of polychlorinated naphthalenes (PCNs) in feed and food
20240702	Risk assessment of small organoarsenic species in food
20240711	Animal dietary exposure in the risk assessment of contaminants in feed
20240715	Update of the scientific opinion on tetrabromobisphenol A (TBBPA) and its derivatives in food
20240823	EFSA Project on the use of NAMs to explore the immunotoxicity of PFAS
20241009	Genotoxicity of beauvericin
20241023	Update of the risk assessment of brominated phenols and their derivatives in food
20241209	Risk assessment of complex organoarsenic species in food
20241211	Have your say - draft risk assessment of fluoride in food and drinking water

The screenshot shows a web browser displaying the article page for 'Animal dietary exposure in the risk assessment of contaminants in feed' on the EFSA Journal website. The page includes the title, publication and adoption dates, social media sharing options, and a table of contents with 'Abstract' selected. Below the title, there is a 'Meta data' section with fields for DOI, KEYWORDS, PANELS, and CONTACT. The 'Abstract' section contains the main text of the article, and the footer indicates it is from the European Food Safety Authority.

Animal dietary exposure in the risk assessment of contaminants in feed

Published: 11 July 2024 | Adopted: 6 June 2024

Share: [in](#) [Twitter](#) [Facebook](#)

efsa JOURNAL
Wiley Online Library

Full article:
[Read online at EFSA Journal](#) | [Full article \(online viewer\)](#)

Meta data

DOI: <https://doi.org/10.2903/j.efsa.2024.8858>

KEYWORDS: animal dietary exposure, complete feed, contaminants, diets, feed materials

PANELS: [Additives and Products or Substances used in Animal Feed](#)

CONTACT: feedap@efsa.europa.eu

Abstract

EFSA performs dietary exposure assessments for food - producing and non - food - producing animals to deliver risk assessment for mandates on the presence of contaminants in feed. The CONTAM and FEEDAP Panels identified the need to update the animal dietary exposure assessment model used in those assessments in CONTAM Scientific Opinions since 2011 in cases where insufficient occurrence data are available on species specific compound feeds. The Panels proposed in this statement a series of model diets based on groups of feed materials with the possibility to use different feed materials in their formulation. The Panels considered that the currently proposed model diets cover the need of the CONTAM Panel to assess the dietary exposure of animals to contaminants in feed.

© European Food Safety Authority

図 9-2 スクレイピングで取得している記事の例

表 9-2 項目を抽出するための質問文。上段の日本語の質問を英訳して LLM に問い合わせた。

項目	質問文
対象物質	下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから 対象物質の名称 のみを教えてください。余計な説明は不要で、対象物質名が見つからない場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ教えてください。

	The following document is a report on the risk assessment of chemical substances. Please provide only the name of the target substance from this document. No additional explanations are necessary. If the name of the target substance cannot be found or if it is not a report on the risk assessment of chemical substances, respond only with "NA".
対象動物	下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから ヒトやウシなどの対象動物の名称 のみを教えてください。余計な説明は不要で、対象動物が見つからない場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ教えてください。
	The following document is a report on the risk assessment of chemical substances. Please provide only the name of the target animals, such as humans or cattle , from this document. No additional explanations are necessary. If the name of the target animals cannot be found or if it is not a report on the risk assessment of chemical substances, respond only with "NA".
ハザードの概要	下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから ハザードの概要 を抽出してください。余計な説明は不要で、ハザードの情報が無い場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ教えてください。
	The following document is a report on the risk assessment of chemical substances. Please extract only the summary of the hazard from this document. No additional explanations are necessary. If there is no hazard information or if the document is not a report on the risk assessment of chemical substances, respond only with "NA".
毒性に関する情報	下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから 毒性に関する情報 を抽出してください。余計な説明は不要で、毒性に関する情報が無い場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ教えてください。
	The following document is a report on the risk assessment of chemical substances. Please extract only the information related to toxicity from this document. No additional explanations are necessary. If there is no information on toxicity or if the document is not a report on the risk assessment of chemical substances, respond only with "NA".
ばく露に関する情報	下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから ばく露に関する情報 を抽出してください。余計な説明は不要で、ばく露に関する情報が無い場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ教えてください。
	The following document is a report on the risk assessment of chemical substances. Please extract only the information related to exposure from this document. No additional explanations are necessary. If there is no information on exposure or if the document is not a report on the risk assessment of chemical substances, respond only with "NA".
他機関の評価情報	下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから 他機関の評価情報 を抽出してください。余計な説明は不要で、他機関の評価情報が無い場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ教えてください。
	The following document is a report on the risk assessment of chemical substances. Please extract only the evaluation information from other organizations in this document. No additional explanations are necessary. If there is no evaluation information from other organizations or if the document is not a report on the risk assessment of chemical substances, respond only with "NA".
評価結果	下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから 評価結果 を抽出してください。余計な説明は不要で、評価結果の情報が無い場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ教えてください。
	The following document is a report on the risk assessment of chemical substances. Please extract only the evaluation results from this document. No additional explanations are necessary. If there is no information on evaluation results or if the document is not a report on the risk assessment of chemical substances, respond only with "NA".

<p>評価の限界や今後の課題</p>	<p>下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから評価の限界や今後の課題を抽出してください。余計な説明は不要で、評価の限界や今後の課題の情報が無い場合や下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は NA とのみ答えてください。</p> <p>The following document is a report on the risk assessment of chemical substances. Please extract only the limitations of the assessment or future challenges from this document. No additional explanations are necessary. If there is no information on the limitations or future challenges, or if the document is not a report on the risk assessment of chemical substances, respond only with "NA".</p>
<p>全項目</p>	<p>下記のドキュメントは化学物質のリスク評価に関する報告です。このドキュメントから対象物質、対象動物、ハザードの概要、毒性に関する情報、ばく露に関する情報、他機関の評価情報、評価結果、評価の限界や今後の課題の情報を抽出して項目別にまとめてください。各項目の情報がなければ無理に答えず、項目の情報としては NA とのみ答えてください。下記のドキュメントが化学物質のリスク評価に関する報告で無い場合は mismatched と答えてください。</p> <p>The following document is a report on the risk assessment of chemical substances. Please extract and organize the information into the following categories: target substance, target animal, hazard summary, information on toxicity, information on exposure, evaluation information from other organizations, evaluation results, and limitations of the assessment or future challenges. If there is no information for a particular category, do not force an answer and respond with "NA" for that category. If the document is not related to the risk assessment of chemical substances, respond with "mismatched".</p>

9.2.1.2. 取得結果例

例として 2024 年 12 月 9 日の「Risk assessment of complex organoarsenic species in food」(食品中の複雑な有機ヒ素化合物のリスク評価)を下記に示した。想定していた情報の取得状況、LLM による取得情報の揺らぎ等を確認し、評価方法の検討を行う必要がある。

9.2.1.2.1. 元記事

元記事情報からリンク等を除いた主要部分を下記に示した。

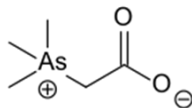
<p>Keywords Arsenobetaine, arsenolipids, arsenosugars, complex organoarsenic species, margin of exposure (MOE), risk assessment</p> <p>Abstract The European Commission asked EFSA for a risk assessment on complex organoarsenic species in food. They are typically found in marine foods and comprise mainly arsenobetaine (AsB), arsenosugars and arsenolipids. For AsB, no reference point (RP) could be derived because of insufficient toxicity data. AsB did not show adverse effects in the two available repeat dose toxicity tests in rodents. It has not shown genotoxicity in in vitro assays. There is no indication of an association with adverse outcomes in human studies. The highest 95th percentile exposure for AsB was observed in 'Toddlers' with an estimate of 12.5 µg As/kg bw per day (AsB expressed as elemental arsenic). There is sufficient evidence to conclude that AsB at current dietary exposure levels does not raise a health concern. For glycerol arsenosugar (AsSugOH) a RP of 0.85 mg As/kg bw per day was derived based on the BMDL10 values for cognitive and motor function in mice. A margin of exposure (MOE) of 1000 would not raise a health concern. The highest 95th percentile estimate of exposure for AsSugOH (for adult consumers of red seaweed Nori/Laver) was 0.71 µg As/kg bw per day (AsSugOH expressed as elemental arsenic), which results in an MOE>1000, not raising a health concern. Based on qualitative consideration of all identified uncertainties, it is regarded likely that the dietary exposures to AsB and AsSugOH do not raise a health concern. No conclusions could be drawn regarding other arsenosugars. No risk characterisation could be conducted for arsenolipids, due to the lack of data.</p> <p>キーワード :</p>

アルセノベタイン、アルセノリピド、アルセノ糖、複雑な有機ヒ素化合物、暴露マージン (MOE)、リスク評価

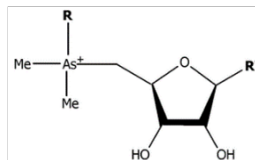
要旨

欧州委員会は EFSA に食品中の複雑な有機ヒ素化合物のリスク評価を依頼した。これらは主に海産物に含まれており、主にアルセノベタイン (AsB)、アルセノ糖、アルセノリピドで構成される。AsB については、毒性データが不十分なため参照点 (RP) を導出できなかった。AsB は、げっ歯類を用いた 2 つの利用可能な反復投与毒性試験において有害な影響を示さなかった。in vitro 試験では遺伝毒性を示していない。ヒトを対象とした研究では、有害な結果との関連性は示されていない。AsB の最も高い 95 パーセントイル暴露は「幼児」で観察され、12.5 $\mu\text{g As/kg}$ 体重/日 (元素ヒ素として表現された AsB) と推定された。現在の食事暴露レベルの AsB が健康上の懸念を引き起こさないと結論付けるには十分な証拠がある。グリセロールアルセノ糖 (AsSugOH) については、マウスの認知機能と運動機能に関する BMDL10 値に基づいて、0.85 mg As/kg 体重/日の RP が導出された。暴露マージン (MOE) が 1000 以上であれば健康上の懸念は生じない。AsSugOH の最も高い 95 パーセントイル暴露推定値 (赤海藻ノリ/ラバーの成人消費者) は 0.71 $\mu\text{g As/kg}$ 体重/日 (元素ヒ素として表現された AsSugOH) であり、これは MOE > 1000 となり、健康上の懸念を引き起こさない。特定されたすべての不確実性を定性的に考慮すると、AsB と AsSugOH の食事暴露が健康上の懸念を引き起こす可能性は低いと考えられる。他のアルセノ糖については結論を出すことができなかった。アルセノリピドについては、データ不足のためリスク特性評価を行うことができなかった。

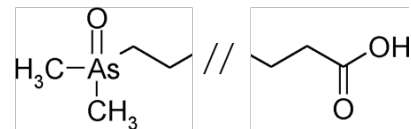
arsenobetaine (AsB)



arsenosugars



arsenolipids



Claude 3.5 Sonnet v1 による現システムでの要約は下記である。

欧州食品安全機関 (EFSA) は、主に海産物に含まれる複雑な有機ヒ素化合物 (アルセノベタイン (AsB)、ヒ素糖、ヒ素脂質) に関する食品中のリスク評価を実施した。主な調査結果は以下の通りである。

1. アルセノベタイン (AsB) :

- 毒性データが不十分なため、参照値を設定できなかった。
- げっ歯類の研究や in vitro 遺伝毒性試験で有害影響は観察されなかった。
- 最高曝露量 (95 パーセントイル) は幼児で 12.5 $\mu\text{g As/kg}$ 体重/日であった。
- 現在の食事による曝露レベルは健康上の懸念を引き起こさない。

2. グリセロールヒ素糖 (AsSugOH) :

- マウスの認知機能と運動機能に基づき、参照値を 0.85 mg As/kg 体重/日と設定した。
- 曝露マージン (MOE) 1000 が安全とみなされる。
- 最高曝露量 (95 パーセントイル) は赤藻類を摂取する成人で 0.71 $\mu\text{g As/kg}$ 体重/日であった。
- MOE > 1000 であり、健康上の懸念はないことを示している。

3. その他のヒ素糖とヒ素脂質 :

- 他のヒ素糖については結論を出すことができなかった。
- データ不足のため、ヒ素脂質のリスク特性評価は不可能であった。

総じて、現在の証拠と不確実性を考慮すると、AsB と AsSugOH の食事による曝露は健康上の懸念はないと考えられる。

9.2.1.2.2. 対象物質

Claude 3.5 Sonnet v1 は日本語訳の時に説明文が付与されていたため英語の回答を示した。

Claude 3.5 Sonnet	complex organoarsenic species
-------------------	-------------------------------

GPT-4o	複雑な有機ヒ素化合物
Gemini 1.5 Pro	複雑な有機ヒ素種

9.2.1.2.3. 対象動物

Claude 3.5 Sonnet v1 は日本語訳の時に「なし」に変換されてしまったため英語の回答を示した。

Claude 3.5 Sonnet	humans
GPT-4o	人間
Gemini 1.5 Pro	ヒト、マウス、げっ歯類

9.2.1.2.4. ハザードの概要

Claude 3.5 Sonnet	なし
GPT-4o	ドキュメントから抽出された危険性の概要：,アルセノベタイン (AsB)：毒性試験で有害な影響は観察されず、遺伝毒性もなく、人間の研究において有害な結果との関連性も認められませんでした。現在の食事による摂取量は健康上の懸念を引き起こしません。グリセロールアルセノシュガー (AsSugOH)：基準値 (RP) として 0.85 mg As/kg 体重/日が導き出されました。暴露マージン (MOE) が 1000 以上の場合、健康上の懸念はありません。現在の食事による摂取量は健康上の懸念を引き起こしません。、その他のアルセノシュガー：結論を導き出すことはできませんでした。、アルセノリピッド：データ不足のためリスクの特性評価は行われていません。
Gemini 1.5 Pro	アルセノベタイン (AsB) は、げっ歯類の研究では有害な影響を示さず、in vitro での遺伝毒性も示さなかった。現在の曝露レベルでは、ヒトの研究において有害な結果は示されていない。グリセロールアルセノシュガー (AsSugOH) については、暴露マージン (MOE) > 1000 が算出され、現在の曝露レベルでは健康への懸念はないことが示された。他のアルセノシュガーについては結論を導き出すことができず、データ不足のため、アルセノリピッドのリスク特性評価は不可能であった。

9.2.1.2.5. 毒性に関する情報

Claude 3.5 Sonnet	この文書には、毒性に関する以下の情報が含まれています：,アルセノベタイン (AsB) については、毒性データが不十分なため、参照ポイントを導き出すことができませんでした。AsB は、げっ歯類での 2 つの利用可能な反復投与毒性試験において有害な影響を示しませんでした。また、in vitro アッセイで遺伝毒性を示していません。、グリセロールヒ素糖 (AsSugOH) については、マウスの認知機能と運動機能に関する BMDL10 値に基づき、0.85 mg As/kg 体重/日の参照ポイントが導き出されました。、AsSugOH の曝露マージン (MOE) が 1000 以上であれば、健康上の懸念は生じないと考えられます。、他のヒ素糖の毒性に関しては、結論を出すことができませんでした。、データ不足のため、ヒ素脂質についてはリスク特性評価を行うことができませんでした。
GPT-4o	欧州委員会は、食品中の複雑な有機ヒ素化合物に関するリスク評価を EFSA (欧州食品安全機関) に依頼しました。これらは主に海産食品に含まれ、主にアルセノベタイン (AsB)、アルセノシュガー、アルセノリピッドで構成されています。AsB については、毒性データが不十分であるため、基準点 (RP) を導出することができませんでした。AsB は、げっ歯類を用いた 2 つの反復投与毒性試験において有害な影響を示さず、また、in vitro 試験において遺伝毒性も示しませんでした。人間の研究においても有害な結果との関連性は示されていません。AsB の 95 パーセンタイルの最高曝露量は「幼児」で観察され、12.5 µg As/kg 体重/日 (AsB を元素ヒ素として表現) と推定されました。現在の食事性曝露レベルにおいて、AsB が健康上の懸念を引き起こすことはない結論付けるのに十分な証拠があります。、グリセロールアルセノシュガー (AsSugOH) については、マウスの認知機能および運動機能に基づく BMDL10 値に基づき、0.85 mg As/kg

	体重/日という RP が導出されました。暴露マージン (MOE) が 1000 以上であれば健康上の懸念を引き起こさないとされます。AsSugOH の 95 パーセントイルの最高暴露量 (赤い海藻ノリ/ラバーを消費する成人) は $0.71 \mu\text{g As/kg}$ 体重/日 (AsSugOH を元素ヒ素として表現) と推定され、MOE が 1000 を超えるため、健康上の懸念を引き起こすことはありません。特定されたすべての不確実性を定性的に考慮した結果、AsB および AsSugOH の食事性暴露が健康上の懸念を引き起こさない可能性が高いと考えられます。他のアルセノシュガーについては結論を出すことができませんでした。アルセノリピッドについてはデータが不足しているため、リスク特性評価を行うことができませんでした。
Gemini 1.5 Pro	AsB については、毒性データが不十分なため、参照点 (RP) を導出できませんでした。AsB は、入手可能な 2 つのげっ歯類を用いた反復投与毒性試験において有害作用を示しませんでした。また、 <i>in vitro</i> 試験では遺伝毒性を示していません。グリセロールアルセノシュガー (AsSugOH) については、マウスにおける認知機能および運動機能に対する BMDL10 値に基づき、 0.85 mg As/kg 体重/日の RP が導出されました。

9.2.1.2.6. ばく露に関する情報

Claude 3.5 Sonnet	ドキュメントには、ばく露に関する以下の情報が含まれています：,アルセノベタイン (AsB) の 95 パーセントイルばく露の最高値は、「幼児」で観察され、1 日あたり $12.5 \mu\text{g As/kg}$ 体重 (元素状ヒ素として表現された AsB) と推定されました。、グリセロールアルセノシュガー (AsSugOH) の 95 パーセントイルばく露推定値の最高値は (赤海藻ノリ/ラバーを摂取する成人について)、1 日あたり $0.71 \mu\text{g As/kg}$ 体重 (元素状ヒ素として表現された AsSugOH) でした。
GPT-4o	AsB の 95 パーセントイルにおける最高の暴露量は「幼児」で観察され、その推定値は $12.5 \mu\text{g As/kg}$ 体重/日 (AsB は元素ヒ素として表現) でした。、AsSugOH の 95 パーセントイルにおける最高の暴露量の推定値は、(赤い海藻であるノリ/ラバーを消費する成人の場合) $0.71 \mu\text{g As/kg}$ 体重/日 (AsSugOH は元素ヒ素として表現) でした。
Gemini 1.5 Pro	ヒ素 (AsB) の 95 パーセントイル最高曝露量は、「幼児」で観察され、推定値は 1 日あたり $12.5 \mu\text{g As/kg}$ 体重 (AsB は元素ヒ素として表される) であった。、紅藻ノリ/アマノリを摂取する成人消費者におけるヒ素糖 (AsSugOH) の 95 パーセントイル曝露量の最高推定値は、1 日あたり $0.71 \mu\text{g As/kg}$ 体重 (AsSugOH は元素ヒ素として表される) であった。

9.2.1.2.7. 他機関の評価情報

Claude 3.5 Sonnet	なし
GPT-4o	なし
Gemini 1.5 Pro	なし

9.2.1.2.8. 評価結果

Claude 3.5 Sonnet	なし
GPT-4o	アルセノベタイン (AsB): 現在の食事摂取レベルでは、AsB が健康上の懸念を引き起こさないことを結論付けるのに十分な証拠がある。、グリセロールアルセノシュガー (AsSugOH): 参照値 (RP) として 0.85 mg As/kg 体重/日が導き出された。暴露マージン (MOE) が 1000 以上の場合、健康上の懸念を引き起こさない。AsSugOH の食事摂取は健康上の懸念を引き起こさない。、その他のアルセノシュガー: 結論を導き出すことはできなかった。、アルセノリピッド: データ不足のためリスク特性評価を行うことができなかった。
Gemini 1.5 Pro	AsB については、参照点 (RP) を導出できませんでした。げっ歯類を用いた 2 つの反復投与毒性試験では有害作用は認められませんでした。* <i>in vitro</i> *試験では遺伝毒性は認められませんでした。ヒト試験では有害事象との関連性は示唆されませんでした。幼児における 95 パーセントイル曝露量の上限值は 1 日あた

	り 12.5 $\mu\text{g As/kg}$ 体重でした。現在の食事曝露レベルでの AsB は健康上の懸念を引き起こしません。、グリセロールアルセノシュガー (AsSugOH) については、RP は 1 日あたり 0.85 mg As/kg 体重です。MOE が 1000 以上であれば健康上の懸念を引き起こしません。紅藻類を摂取する成人における 95 パーセントイル曝露量の上限值は 1 日あたり 0.71 $\mu\text{g As/kg}$ 体重であり、MOE は 1000 を超えるため、健康上の懸念を引き起こしません。、AsB および AsSugOH への食事曝露は健康上の懸念を引き起こしません。他のアルセノシュガーに関する結論はありません。データ不足のため、アルセノリピドのリスク特性評価は行われていません。
--	--

9.2.1.2.9. 評価の限界や今後の課題

Claude 3.5 Sonnet	文書から抽出された制限事項と今後の課題は以下の通りです:、1. アルセノベタイン(AsB)の毒性データが不十分で、参照点を導き出すことができない。、2. グリセロールアルセノ糖(AsSugOH)以外のアルセノ糖に関する結論が出せない。、3. データ不足のため、アルセノリピドのリスク特性評価を実施できない。
GPT-4o	・ アルセノベタイン (AsB) については、基準点 (RP) を導出するための毒性データが不十分です。、他のアルセノシュガーに関しては結論を出すことができませんでした。、データ不足のため、アルセノリピドについてのリスク評価は実施できませんでした。
Gemini 1.5 Pro	他の砂糖類については結論を導き出すことができませんでした。、データがないため、砒脂質のリスク特性評価を行うことができませんでした。

9.2.1.2.10. 全項目まとめて抽出

- ・ 項目別抽出では抽出していない内容を赤字で示した。
- ・ 英語で質問し、回答を日本語訳しているために回答の項目名は LLM 間で一致していない。
- ・ GPT-4o 及び Gemini 1.5 Pro ではマークダウン形式で出力されていた。

Claude 3.5 Sonnet	<p>標的物質: 複雑な有機ヒ素化合物、特にアルセノベタイン(AsB)、アルセノ糖 (特にグリセロールアルセノ糖 AsSugOH)、およびアルセノリピド、</p> <p>標的動物: げっ歯類 (毒性研究用)、ヒト (曝露評価用) 、</p> <p>ハザードの要約:、AsB: げっ歯類研究で有害影響は観察されず、in vitro で遺伝毒性なし、AsSugOH: マウスで認知機能と運動機能への潜在的影響、その他のアルセノ糖とアルセノリピド: データ不足、</p> <p>毒性に関する情報:、AsB: 参照点を導出するには不十分なデータ、AsSugOH: マウスでの認知および運動効果に基づく BMDL10 から、0.85 mg As/kg 体重/日の参照点、</p> <p>曝露に関する情報:、AsB: 幼児で最高 95 パーセントイル曝露が 12.5 $\mu\text{g As/kg}$ 体重/日、AsSugOH: 赤海藻を消費する成人で最高 95 パーセントイル曝露が 0.71 $\mu\text{g As/kg}$ 体重/日、</p> <p>他の組織からの評価情報: なし、</p> <p>評価結果:、AsB: 現在の食事曝露レベルでは健康上の懸念なし、AsSugOH: 健康上の懸念なし (MOE > 1000) 、その他のアルセノ糖とアルセノリピド: データ不足のため結論なし、</p> <p>評価の限界または将来の課題:、AsB の参照点を導出するには毒性データが不十分、アルセノリピドのリスク特性評価のためのデータ不足、AsSugOH 以外のアルセノ糖については結論不可能、不確実性の定性的考慮により、AsB と AsSugOH の食事曝露は健康上の懸念を引き起こす可能性は低い</p>
GPT-4o	<p>### 抽出および整理された情報:</p> <p>**対象物質:**、アルセノベタイン (AsB)、グリセロールアルセノシュガー (AsSugOH)、その他のアルセノシュガー、アルセノリピド、</p> <p>**対象動物:**、齧歯類 (AsB の反復投与毒性試験に使用)、マウス (AsSugOH の認知および運動機能に関する研究から参照ポイントを導出するために使用)、</p> <p>**ハザード概要:**、**アルセノベタイン (AsB):** 齧歯類での反復投与毒性試験で有害影響は観察されず。*in vitro* 試験で遺伝毒性なし。人間の研究でも有害な結果との関連性は認められない。、**グリセロールアルセノシュガー (AsSugOH):** マウスで特定の用量で認知および運動機能への影響が</p>

観察された。,**その他のアルセノシュガー:** データ不足のため結論は得られなかった。,**アルセノリピッド:** データ不足のためリスク評価は実施できなかった。 ,
****毒性に関する情報:****,**アルセノベタイン (AsB):** 参照ポイント (RP) を導出するには毒性データが不十分。齧歯類での反復投与毒性試験や*in vitro* 試験で有害影響や遺伝毒性は観察されなかった。,**グリセロールアルセノシュガー (AsSugOH):** マウスの認知および運動機能に関するベンチマーク用量下限信頼限界 (BMDL10) 値に基づき、0.85 mg As/kg 体重/日という参照ポイント (RP) が導出された。 ,
****曝露に関する情報:****,**アルセノベタイン (AsB):** 幼児における 95 パーセントイルの最高曝露量は 12.5 μ g As/kg 体重/日 (元素ヒ素として表記)。,**グリセロールアルセノシュガー (AsSugOH):** 赤海藻 (ノリ/ラバー) を消費する成人における 95 パーセントイルの最高曝露量は 0.71 μ g As/kg 体重/日 (元素ヒ素として表記)。 ,
****他の組織からの評価情報:****なし,
****評価結果:****,**アルセノベタイン (AsB):** 現在の食事性曝露レベルでは健康上の懸念はない。,**グリセロールアルセノシュガー (AsSugOH):** 曝露マージン (MOE) が 1000 以上の場合、健康上の懸念はない。,**その他のアルセノシュガー:** 結論は得られなかった。,**アルセノリピッド:** データ不足のためリスク評価は実施できなかった。 ,
****評価の限界または将来の課題:****アルセノベタイン (AsB) の参照ポイントを導出するには毒性データが不十分。 ,アルセノリピッドのリスク評価にはデータが不足している。 ,その他のアルセノシュガーについてはデータ不足のため結論は得られなかった。

Gemini 1.5 Pro
****対象物質:**** 食品中の複合有機ヒ素種。主にアルセノベタイン (AsB)、アルセノ糖 (特にグリセロールアルセノ糖 - AsSugOH)、およびアルセノ脂質を含む。 ,
****対象動物:**** げっ歯類 (AsSugOH 毒性試験では特にマウス)、ヒト (曝露および疫学研究)。 ,
****有害性概要:**** AsB はげっ歯類試験で有害作用を示さず、in vitro で遺伝毒性も示さない。AsSugOH は、マウスの認知および運動機能への影響に基づいて導出された参照点を持つ。アルセノ脂質は有害性評価のためのデータがない。 ,
****毒性情報:**** AsB : げっ歯類での 2 回の反復投与毒性試験で有害作用なし、in vitro で遺伝毒性なし。AsSugOH : マウスの認知および運動機能について導出された BMDL10 値を用いて、参照点 (RP) 0.85 mg As/kg 体重/日を確立。アルセノ脂質 : データなし。 ,
****曝露情報:**** AsB : 幼児における曝露量の 95 パーセントイル値は 12.5 μ g As/kg 体重/日と推定。AsSugOH : 紅藻類 (海苔) を摂取する成人における曝露量の 95 パーセントイル値は 0.71 μ g As/kg 体重/日と推定。 ,
****他機関による評価情報:**** なし,
****評価結果:**** AsB : 食事による曝露は健康上の懸念を引き起こさない。AsSugOH : MOE > 1000 であり、現在の曝露レベルでは健康上の懸念はない。アルセノ脂質 : データ不足のためリスク評価は不可能。全体として、不確実性を考慮すると、AsB および AsSugOH への食事曝露は健康上の懸念ではない可能性が高い。他のアルセノ糖については結論を導き出すことができなかった。 ,
****評価の限界または今後の課題:**** アルセノ脂質のデータ不足によりリスク評価ができない。AsSugOH 以外の他のアルセノ糖に関するデータが必要。

項目別抽出との主な違いを下記にまとめた。

- Claude 3.5 Sonnet
 - ハザードの概要が抽出されている。
 - 評価結果が抽出されている。
- GPT-4o
 - 対象動物でヒトが抽出されていない。
 - 毒性に関する情報が、項目別ではほぼ全訳であったが要約されている。
- Gemini 1.5 Pro
 - 評価結果にあった参照点(RP)の情報が消えている。

9.2.2. EFSA chemical contaminants in food and feed の full article からの抽出

9.2.2.1. 実施事項

スクレイピングで取得する記事は「News」及び「Scientific Opinion」の「Abstract」であるため情報量が少なく、「食品安全関係情報データベース」に登録する記事としては全文の翻訳になることが予想される。そこで「Scientific Opinion」の9つについてEFSA Journalの「Full article」のPDFファイルをダウンロードし、これをマークダウン形式にテキスト化した記事に対して9.2.1.1節の項目の抽出を行った。この結果を「250305 構造化項目抽出結果報告」フォルダーの「ChemicalContaminantsFullArticle.xlsx」にまとめた。Claude 3.5 Sonnet v1では1つのドキュメント、GPT-4oは3つのドキュメントがトークン数を超えたため回答を得ることが出来なかった。PDFファイルは23ページから146ページであり、トークン数を考慮した抽出項目の評価方法の検討が必要であろう。

9.2.3. LLMによるデータ抽出のまとめ

今回のLLMによる項目抽出の結果からは、項目別抽出とまとめて抽出では抽出される情報が異なることがあった。この結果から、質問文によって抽出項目を調整することが出来ると思われる。よって項目毎に欲しい情報(正解データ)を作成しておくことにより、質問文を調整して精度を高められる可能性がある。

今後もLLMはバージョンアップ及び新規ベンダーによる提供が予想されることから、LLMの評価方法を確立し、結果を自動的に評価する仕組みを構築しておくことが望ましいと思われる。これらの観点からも、統計情報として扱える量の正解データ(抽出したい情報)を構築しておくことが有益であると考えられる。

9.2.4. ハルシネーションの判定

LLMを用いて指定した項目の情報抽出を行った場合、記事に情報が無いにも関わらず別の回答をするハルシネーションが課題となっている。近年推論機能を持つLLMが提供されており、ハルシネーションを判定できる可能性がある。そこでLLMを使って抽出した情報がハルシネーションかどうかをLLMに判定させることを試みた。

Claude 3.5 Sonnet v1 (anthropic.claude-3-5-sonnet-20240620-v1:0)で農薬評価書の文書断片から発がん性の有無を回答させる実験から得られたハルシネーション例(指定した農薬の発がん性の情報が無いにも関わらず、発がん性「無し」と答えた場合)を表9-3に示した。最後の「フロルピラウキシフェンベンジル」の場合は本文中に「本剤に発がん性は認められない」とあるが、この情報は「フェノチオカルブ」であるためNAが正解となる例である。これらの回答について図9-3に示した質問文をLLMに投げてハルシネーションが判定できるのかを試した。

表 9-3 Claude 3.5 Sonnet v1 で与えた文書断片に情報がなく「NA」と回答するべき場合に「無」と回答した時の与えた文書断片

ID	農薬名	専門家抽出	LLM 抽出
PN_170	チアジニル	有	無
評価に用いた試験成績は、動物体内運命（ラット、ヤギ及びニワトリ）、植物体内運命（水稻）、作物等残留、急性毒性（ラット、ウサギ）、亜急性毒性（ラット、イヌ）、慢性毒性（イヌ）、慢性毒性/発がん性併合（ラット）、 発がん性（マウス） 、2世代繁殖（ラット）、発生毒性（ラット、ウサギ）、遺伝毒性等である。各種毒性試験結果			

から、チアジニル投与による影響は、主に 肝臓（重量増加、肝細胞肥大等） 及び 腎臓（尿管上皮空胞化等） に認められた。			
PN_203	トリフロキシストロビン	無	無
評価に用いた試験成績は、動物体内運命（ラット、ヤギ及びニワトリ）、植物体内運命（りんご、小麦等）、作物等残留、急性神経毒性（ラット）、亜急性毒性（ラット及びイヌ）、慢性毒性（イヌ）、慢性毒性/ 発がん性併合（ラット） 、 発がん性（マウス） 、2世代繁殖（ラット）、発生毒性（ラット及びウサギ）、遺伝毒性、免疫毒性等である。各種毒性試験結果から、トリフロキシストロビン投与による影響は、主に 肝臓（肝細胞肥大等） に認められた。			
PN_217	ピコキシストロビン	有	無
各種毒性試験結果から、ピコキシストロビン投与による影響は、主に体重（増加抑制）、肝臓（肝細胞肥大：マウス）及び十二指腸（粘膜過形成及び粘液腺拡張：マウス）に認められた。神経毒性、繁殖能に対する影響、催奇形性、生体にとって問題となる遺伝毒性及び免疫毒性は認められなかった。			
PN_263	フェンブコナゾール	有	無
評価に用いた試験成績は、動物体内運命（ラット、ヤギ及びニワトリ）、植物体内運命（小麦、らっかせい等）、作物等残留、亜急性毒性（ラット、マウス及びイヌ）、慢性毒性（イヌ）、慢性毒性/ 発がん性併合（ラット） 、 発がん性（マウス） 、2世代繁殖（ラット）、発生毒性（ラット及びウサギ）、遺伝毒性等である。各種毒性試験結果から、フェンブコナゾール投与による影響は、主に 肝臓（肝細胞肥大及び空胞化等） に認められた。 催奇形性及び遺伝毒性は認められなかった。			
PN_278	フルオピラム	有	無
各種毒性試験結果から、フルオピラム投与による影響は、主に眼（ラット：角膜混濁、網膜退色等）、肝臓（重量増加、小葉中心性肝細胞肥大等）、腎臓（重量増加、慢性腎症等）及び甲状腺（ろ胞上皮細胞肥大等）に認められた。神経毒性、繁殖能に対する影響、免疫毒性及び遺伝毒性は認められなかった。（参照1、45、68）13. 遺伝毒性試験フルオピラム原体の細菌を用いた復帰突然変異試験、チャイニーズハムスター肺由来細胞（V79）を用いた染色体異常試験及びHprt 遺伝子座突然変異試験、マウスを用いた in vivo 小核試験が実施された。結果は表43に示されているとおり全て陰性であったことから、フルオピラムに遺伝毒性はないものと考えられた。（参照1、46～50、68）51 表43 遺伝毒性試験概要（原体）試験対象処理濃度・投与量結果 invitro 復帰突然			
PN_321	フロルピラウキシフェンベンジル	無	無
しかしながら、マウス骨髄での小核の誘発は低体温に起因する可能性もあり、いずれにせよ、 本剤に発がん性は認められない ことから、これらの遺伝毒性陽性反応は発がん性と無関係であると考えられ、ADI及びARfDの設定は可能と考えられた。各種試験結果から、農産物中の暴露評価対象物質を フェノチオカルブ （親化合物のみ）と設定した。			

下記の問い合わせを行った際の回答が"無"でした。この回答が正しいかどうか判断してください。

[参考]部分の情報は PDF から markdown に変換したドキュメントです。PDF からの文章の抽出は前後が入れ替わることや誤字脱字、途中で切れていることがあるため、内容を予測しながら質問に回答してください。

[質問]

{chemical}の発がん性の有無を"有"もしくは"無"で教えてください。プロンプト中に答えがなければ NA とのみ教えてください。補足情報等は回答に加えなくてください。

[参考]

{rag_context}

図 9-3 ハルシネーション判定に用いたプロンプト。{chemical}には農薬名、{rag_context}には農薬評価書から得られた文書断片に置き換えて LLM に問い合わせる。

図 9-3 の質問に対して、間違いと指摘した場合、及び「NA」と回答した場合を正解としたときの各

LLM の正解の回数を表 9-4 に示した。LLM は問い合わせる度に回答が異なるため 3 回同じ質問を繰り返している。全ての回答結果を「250312 第 8 回業務進捗報告会」フォルダーの「Hallucination 判定回答.xlsx」に示した。一部に回答が間違っていると指摘していながら「無」と回答している場合や回答が正しいと述べながら「NA」と答えている場合があった(いずれも不正解にカウントした)。推論機能のある Gemini 2.0 Flash Thinking や GPT-o1 ではハルシネーションと判定出来ている場合もあり、特に Gemini 2.0 Flash Thinking はかなりの頻度でハルシネーションを判定出来ていた。

表 9-4 LLM 毎にハルシネーションと判定した回数。同じ質問を 3 回行っている。

	PN_170	PN_203	PN_217	PN_263	PN_278	PN_321	合計
Gemini 2.0 Flash	0	0	0	0	0	0	0
Gemini 2.0 Flash Thinking	3	3	3	2	3	1	15
Gemini 2.0 Pro	0	3	0	0	0	0	3
GPT-4o	0	0	0	0	1	0	1
GPT-o1	1	0	3	3	3	0	10
GPT-o3-mini	0	0	1	0	3	0	4

LLM の推論機能の進展により、ハルシネーションを判定できるようになってきた。複数の LLM の組み合わせや、同じ質問を何度か行った際の多数決を取るなどにより精度を高められる可能性がある。項目の抽出についても推論機能のある LLM で行うことにより、プロンプトを最適化することで高い精度で実現できる可能性があると思われる。

9.3. 構造化抄録データベース構築における課題のまとめ

9.3.1. カテゴリー化と抽出項目の検討

食品安全領域の論文は多様な領域を含んでおり、構造化抄録を構成する項目は領域ごとに異なる。そこで食品安全情報を提供する記事のカテゴリー化を行い、カテゴリー毎に抽出項目を決定する必要がある。これは日常業務において食品安全情報を作成、整理している食安委職員による作業となるだろう。ただしこの作業ではカテゴリー毎の抽出項目が一致するような分類を行う必要がある。食品安全領域の分類と抽出項目の決定は独立に行うのではなく、抽出項目を検討しながら分類を決定する必要がある。

9.3.2. 専門用語辞書の作成

元記事に表現の揺らぎがあるため、抽出項目の一致をみる解析では表現の揺らぎ対策が必要である。既存の同義語辞書に加えて専門用語同義語辞書も必要になる。

9.3.3. 運用による修正

食品安全領域の分類と抽出項目を決定後、ある程度の期間運用し、これを修正する作業が発生すると考えられる。運用には LLM を用いた項目の抽出が必要となるが、この精度と LLM の回答の揺らぎが課題となる。

9.3.4. 自動抽出の最適化

LLM の開発速度が速いため、評価を迅速に行える環境を構築する必要がある。LLM の精度評価を機械的に行うためには、エキスパートによる項目の抽出結果が必要となる。統計情報として算出できる規模の正解データ(例えば 200 件)をカテゴリー領域毎に作成し、且つ評価方法を検討する必要がある。項目の抽出では、LLM の選定とプロンプトの最適化を行ってから実行することが望ましい。この目的のためにエキスパート抽出データを学習用(training)と検証用(test もしくは hold-out)に分けて、学習用データで LLM とプロンプトの組合せを最適化し、検証用データで精度情報を算出するなどの方法が考えられる。

一方で推論機能付き LLM を用いるとハルシネーションがある程度判定出来る様になったなど、技術革新によりこれまでの課題が解決されることもある。最新情報を収集し続け、新たな方法論を考察し続けることも必要である。