地名に関しては、MT ソフトウエアは、Nonthaburi をカタカナに翻字化することで対処していると思われる。一部の有名な外国の地名については、日本語の形態素解析器の辞書に収載されているため、GHIS がヘルスイベント発生場所を自動的に特定する上でおおいに役立つ。あまり知られていない地名でも自動的に翻字されており、エンドユーザが、イベントの発生場所、さらに、その場所における歴史的および地理的な背景について理解するために、日本語で記述された Wiki などのウェブページを検索する際に役に立つ（この例では、日本語 Wiki ページにノンタブリに関する記述がある）。

<WHO 例 3>

Field investigations have not found any indications of respiratory illness in close contacts of the patient.

<Atlas の結果>

現地調査は患者の近接の呼吸器疾患のどんなしるしにも当たっていません。

<人間による翻訳>

フィールド調査によると、患者と密接接触を持った者の間で呼吸器疾患を発症した者は今のところいない。

この例では、付加詞「in close contacts of」の日本語訳は適切ではない。また、動詞句「have not found」も正しく翻訳されていない。この場合、付加詞および動詞句の翻訳の質が低いことが、自動イベント抽出および人間の理解に大きな影響を及ぼす。専門用語に関しては、一部のあまり一般的でない用語が誤って翻訳されているが（例：「constitutional syndrome（全身性症候群）」が「本質的なシンドローム」と訳されている）、大部分の用語は正しく翻訳されている。

以下は、上記の 3 つの例を SYSTRAN と、無料で利用可能な Google 翻訳（β版）の 2 つの MT システムを使用して機械翻訳したものである。例 1 の結果は以下の通りである。

<SYSTRAN の結果>

疫学的な調査は伝染の本当らしいもととして感染させた家禽への露出のいずれの場合も覆いを取った。今まで、インドネシアは H5N1 鳥インフルエンザの 7 つの人間の場合を報告した。これらの場合の 4 つは致命的だった。

<Google 翻訳の結果>

疫学的調査の両方のケースで感染の可能性、感染した家禽への曝露源として明らかになった。現在までに、インドネシア、H5N1 亜型鳥インフルエンザのヒト症例を報告した 7。これら 4 つの例死亡している。

「avian flu」および「Epidemiological investigation」などの専門用語は正しく翻訳されている。この点に関しては Atlas と同じであったが、領域に依存する表現は上手く翻訳されていない。領域に依存する表現である「human case」における「case」および「source of infection」における「source」を、SYSTRAN は、誤ってそれぞれ「場合」および「もととして」と翻訳している。Google 翻訳は、「source of infection」を誤って「exposure source」と認識している（結果、「曝露源」として翻訳している）。さらに、Google 翻訳は、「seven human cases」の区切りを誤っており、Google 翻訳の結果から「seven cases（7 例）」を理解することができない。例 2 の翻訳結果は以下の通りである。

<SYSTRAN の結果>

患者は Nonthaburi の地域、裏庭の鶏が数日先に死に始めたバンコクの北の彼女の夫を訪問した。

<Google 翻訳の結果>

ノンタブリ県では患者さんの夫、バンコク北部の裏庭で鶏を数日前に死亡し始めて訪問した。

この例では、SYSTRAN は地名を翻訳していないが、Google は正しくカタカナに翻訳している。例 3 の翻訳は以下の通りである。

<SYSTRAN の結果>

実地調査は患者の近い接触の呼吸の病気の徴候を見つけなかった。

<Google 翻訳の結果>

フィールド調査は、患者の密接接触者の呼吸器疾患の兆候を発見していない。

付加詞「in close contacts of」は、Google 翻訳では正しく翻訳されているが、SYSTRAN は正しく翻訳していない。したがって、MT による付加詞の翻訳は安定しているとは言い難い。MT システムの結果間には軽微な差が存在するが、傾向として、保健関連の用語は翻訳されるが、動詞句あるいは付加詞の翻訳は得意ではないようである。

ここに示した例の検討から、MT システムがニュースのトピックの自動分類において大変役立つと考えられる。対照的に、テキストマイニングシステムへの入力として MT を使用するイベント抽出では、良好な結果を期待することはできない。人間による最終的な判断の支援としては、ニュース報道の一部の簡単な表現は要約に役立つものと考えられたが、上記の例では、読者にニュース報道の元の言語に関する知識がまったくない場合、翻訳を読んでニュース報道の詳細について理解することは困難であると思われた。これは、一般的な保健関連のニュースでは動詞句および付加詞の処理があまり良くないという翻訳精度のためである。

参考文献・資料

[1] Wilks, Y. (2009), "Machine Translation - Its Scope and Limits", SpringerScience.

[2] 永田昌明 渡辺太郎 塚田元, 機械翻訳最新事情：(上) 統計的機械翻訳入門, 情報処理, Vol.49, No.1, pp. 89-95, 2008.

[3] 塚田元 永田昌明 隅田英一郎 黒橋禎夫, 機械翻訳最新事情：(下) 評価型ワークショップの動向と日本からの貢献, 情報処理学会会誌, vol. 49, No. 2. pp. 70-80,2008.

[4] Asia-Pacific Association for Machine Translation (2009), "The list of machine translation software", http://www.aamt.info/japanese/mtsys-j.htm (last accessed 15th January 2009).

[5] John, W. and Somers, H. L. (1992), "An Introduction to Machine Translation", Academic Press.

[6] Miller, K.,, Gales, D., Underwood, N. and Magdalen, J. (2001), "Evaluation of Machine Translation Output for an Unknown Source Language: Report of an ISLE-based Investigation"

[7] Hovy, E., (1999), "Toward finely differentiated evaluation metrics for machine translation", in Proc. Eagles Workshop on Standards and Evaluation, Pisa, Italy.

[8] Reeder, F. (2001), "Additional MT-eval references", Tech. Report of the International Standards for Language Engineering, Evaluation Working Group. Available at http://issco-www.unige.ch/projects/isle/fermti/. Last accessed December 2nd 2008.

[9] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002), "BLEU: a Method for Automatic Evaluation of Machine Translation", in Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July, pp. 311-318.

1.3 オントロジー

オントロジーの設計、構築、および維持方法について簡単に説明することは不可能である。むしろ、ここでは、ユーザの観点から簡潔に主要な考え方を特定し、本報告書の後半で取り上げる公衆衛生の応用と関連付ける。

まず、未来型のコンピュータシステムによる検索について考える。分析者は、「9 月 6 日から 9 日の間にベトナム北東部で報告された呼吸症候群のすべてのヒト症例」を探すためのリクエストを入力する。コンピュータが、分析者が期待する文書に高い割合で一致する文書を検出するためには、どのような知識が必要か？コンピュータは、少なくとも以下について知っている必要がある。

- 「A（H5N1）型インフルエンザ」には「呼吸症候群」と関連した症状がある
- 「A（H5N1）型インフルエンザ」と同義の言葉は 12 個ある
- 「ベトナム北東部」には、Cao Bang、Lao Cai、Yen Bai、Ha Giang、Tuyen Quang などが含まれる
- 「A（H5N1）型インフルエンザ」は「A 型インフルエンザウイルスの亜型である H5N1 型」によって引き起こされる
- 他の言語にも同義語が存在する

これは、オントロジーに含まれる一般的な種類の情報である。インターネット上の通常のキーワードによる検索エンジンでは、この種の詳細な情報が欠けており、この例のような人間の分析者の作業を支援することはできない。

オントロジーの考え方は、最も広い意味において、具体的または抽象的、現実または想像上の、などのオブジェクトの異なるカテゴリ間の区別およびそれらの間の関係を形式化する。これは、コンピュータ科学および自然科学などデータを中心とした多くの学問分野の中心にあり、また、オントロジーに関する公式・非公式の研究が近年になって研究トピックとして復活した理由でもある。現実における基本的な性質を説明することは新しいことではなく、パルメニデス、アリストテレス、およびプラトンの時代にまで遡ることができる。オントロジーは、哲学者だけが使用する抽象的な理論的概念ではなく、現在では、領域における用語を形式化的に構築するための手段として研究の主流となっている[1]。それらの最も重要な機能は、コミュニティにおける共通言語となること、言語と人間およびコンピュータの理解との間のギャップを埋めることである。つまり、オントロジーは専門知識の領域のモデルとなる。より理論的な形では、オントロジーはコンピュータによる推論の基盤となり、生物医学的情報科学のような分野における実世界についての知的な意思決定を支援する。

入門文献[2]において、Noy および McGuiness は、オントロジーを作成する他の理由についても言及している。

(a) 分野依存知識の再使用を可能にする。時間や位置などの共通概念は、異なる領域においても使用されている。新しい領域で定義を再使用することで労力

を抑えることができる。この考え方によりオントロジーが統合されることが望まれる。

(b) 領域の条件を明確にする。これにより、将来、概念あるいは関係についての考え方を変更する必要がある場合、よりよい変更管理ができる。

(c) 運用知識から分野依存知識を分離する。知識がそれを処理するソフトウエアから明示的に分離されるため、多くの異なるコンピュータシステムにおいて知識を独立して使用および維持することができる。

(d) 分野依存知識を分析する。明示的な形式で用語の語彙を持つことで、研究者がそれらを拡張するために自動分析を適用することができる。

オントロジー作成の方法論に関して、研究者の間で広く受け入れられているコンセンサスが存在しないため、標準的なオントロジーは存在せず、そのようなものが存在するか、あるいは、必要かについては考え方による。哲学的な理論については進展中であるにもかかわらず、既に数多くのオントロジーが実用的な使用のため開発されている。これらは様々な異なる目的のために構築されていることが多いが、多くが互いに重複している。現実には、重複したオントロジーでもモデル化する領域について異なる「見解」を取ることがあり、相反していることが多くある。しかし近年、Open Biomedical Ontologies（OBO）[3]などのコミュニティ努力によって、オントロジー構築のための標準化にむけて大きく前進し、オントロジー構造の矛盾が明らかにされ、相互操作性が促進されている。

通常オントロジーには、少なくとも用語と「type of」関係によって構築される人間による定義が含まれる。例えば、Medical Subject Headings（MeSH）オントロジー[4]では、「influenza（インフルエンザ）、human（ヒト）」は「respiratory tract infection（気道感染症）」の一種（type of）と定義されている。人間が理解しているその他の広く知られているオントロジー例としては、SNOMED Clinical Terms[5]、Unified Medical Language System（UMLS）[6,7]およびAGROVOC[8]がある。「type of」関係以外では、多くのオントロジーが以下をエンコードしている。

- 「black death（黒死病）」が「bubonic plague（腺ペスト）」の別の表現であるなどの類義性（同義語）。
- 他のオントロジーの用語への外部参照。これにより、ユーザは異なる命名則間のギャップを埋めることができる。
- オントロジーにおける他の用語と名前の付けられた関係。例えば、特定のウイルスは、ヒトに特定の疾患を引き起こす可能性がある。

より具体的な例について検証する。コンピュータシステムが感染症について記述された文書を部分的に理解するためには、その領域に関する形式化された知識がある程度必要なのは明らかである。それがオ

ントロジーの形の場合、必要な用語カテゴリには、DISEASE（疾病）、SYMPTOM（症状）、および SPECIES（種）が含まれる。また、治療の知識が重要な場合には、DRUG（薬剤）を含める可能性がある。禾穀類の疾病を含めるようにオントロジーを拡張する場合には、オントロジーに PHENOTYPE GENE（表現型遺伝子）が含まれることが考えられる。疾病とともに地理空間的情報を収集するためシステムを拡張する場合には、COUNTRY（国名）、PROVINCE（州、省、県）、LATITUDE（緯度）、LONGITUDE（経度）、および POPULATION（人口）を定義することが考えられる。これは、システムからの情報を Google マップあるいは NASA の World Wind などの地理空間ブラウザと併用する際に役立つ。

オントロジーの重要な目的は、用語および概念を明確に定義することである。例えば、「VIRUS（ウイルス）」と言う概念は、容易にコンピュータに感染する有害ソフトウエアまたは微生物と取り違えられる可能性がある。これは、人間では自然言語による定義を、コンピュータでは論理的規則を使用することによってなされる。

公衆衛生領域では、多くのオントロジーが個々のコンピュータシステムで使用するために手で作成されている。これらに含まれる知識は長年にわたって行われたカスタマイズのために極めて有用であるが、一般公開されているのはこれらのうちの少数である。BioCaster オントロジー（BCO）[9]は、一般公開されている人間およびコンピュータの相互理解におけるリソースの1例である。規模は SNOMED あるいは UMLS と比較してはるかに小さいものの、感染症を記述するための基礎語彙を多数の言語で提供している。下記の図2はその構造を示している。DISEASE（疾病）および SYMPTOM（症状）の概念と「ルート用語」、また、それらの様々な言語による表現が示されている。これにより、言語を超えて、記述するトピックについて文書を検索および比較することが可能になる。

オントロジーが含む情報、およびコンピュータによる文書の知的処理における役割について極めて簡単に説明した。これは、必要な分野依存知識および背景情報を提供することによって行われる。オントロジーは分野依存知識をエンコードする強力な方法であるが、単独では分析システムとして機能するものではない。コンピュータシステムにおけるそれらの位置、つまり、コンピュータによる解釈を支援するが、コンピュータが理解できる形式へのフリーテキストの変換がテキストマイニングシステムなどの言語処理技術を使用して他の場所で行われることは重要である。参照した地理空間アプリケーションについては[10]を、また、サーベイランスにおける知識に関する問題については[11]を参照とする。

参考文献・資料

[1] Gruber, T.R. (1993), "A Translation Approach to Portable Ontology Specification", Knowledge Acquisition 5: 199-220.

[2] Noy, N. and McGuinness, D. (2009), "Ontology development 101: A guide to creating your first ontology", available from http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html (last accessed 1st March 2009).

[3] The Open Biomedical Ontologies (2009), available from http://www.obofoundry.org/ (last accessed 1st March 2009).

[4] Lowe, H. & Barnett, G. (1994), "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches", J. Am. Med. Assoc. 271, 1103–1108.

[5] Price, C. and Spackman, K. (2000),

"SNOMED clinical terms", British Journal of Healthcare Computing & Information Management 17(3): 27-31.

[6] Humphreys, B. and Lindberg, D.. (1993), "The UMLS project: making the conceptual connection between users and the information they need", Bulletin of the Medical Library Association 81(2): 170.

[7] Bodenreider, O. (2004), "The Unified Medical Language System (UMLS): integrating biomedical terminology", Nucleic Acids Res. 32 (database issue): D267–D270.

[8] Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J. and Katz, S., (2004), "Reengineering thesauri for new applications: the AGROVOC example", J. Digital Inform, 4.

[9] Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D. Barrero, R., Takeuchi , K.and Kawtrakul, A. (2007), "A multilingual ontology for infectious disease surveillance: rationale, design and challenges", Language Resources and Evaluation, Elsevier, DOI: 10.1007/s10579-007-9019-7.

[10] Scharl, A. and Tochtermann, K. (eds) (2007), The Geospatial Web – How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society, Springer-Verlag, London.

[11] Mehrotra, S., Zeng, D., Chen, H., Thuraisingham, B. and Wang, F. (eds) (2006), Intelligence and Security Informatics, Springer-Verlag, Berlin.

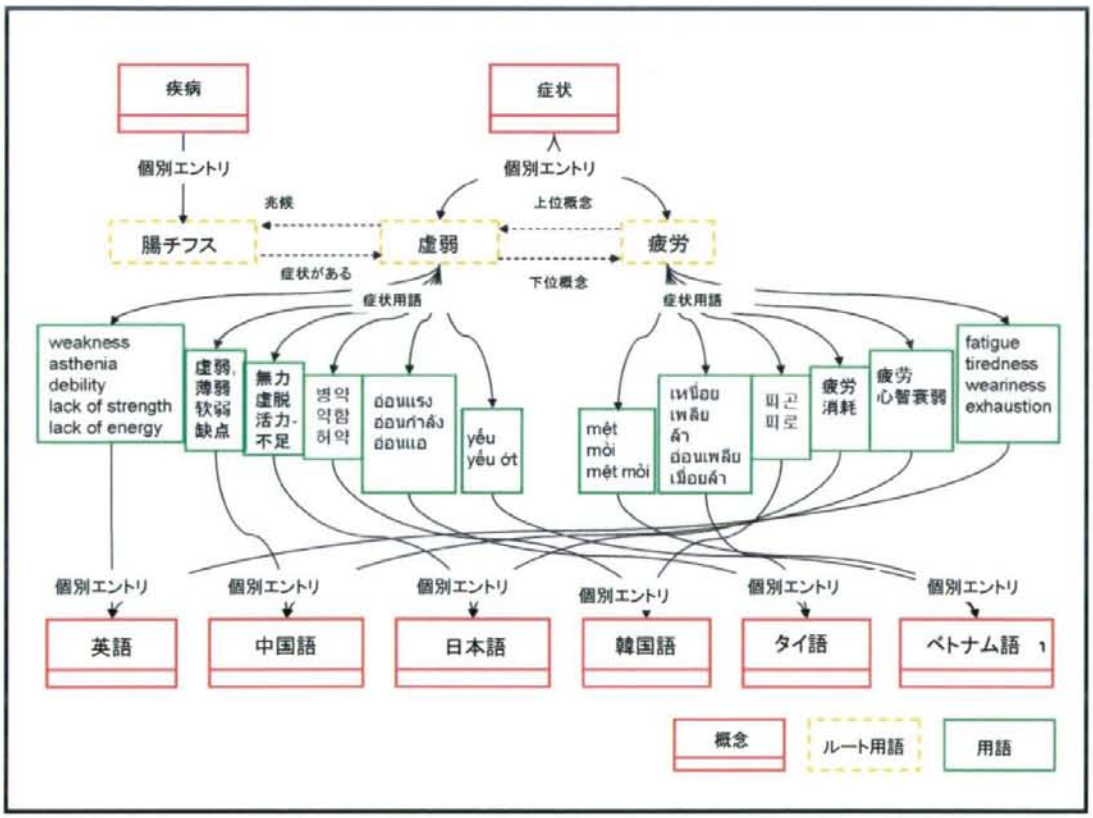

図 2：BioCaster オントロジーにおける DISEASE（疾病）および SYMPTOM（症状）の概念、「Typhoid fever（腸チフス）」、「weakness（虚弱）」および「fatigue（疲労）」のルート用語および同義語を 6 つの言語で示している。

結果および考察

2．ケーススタディ

2. ケーススタディ

ここで報告される構造化されたケーススタディは、2008 年 11 月後半から 2009 年 3 月前半にかけて面談および文献調査を併用して実施された。実施した調査対象のすべてのシステムについて、現場訪問あるいはメールによってシステムの所有者と連絡を取り、詳細を確認している。調査では、各システムの背景、対象範囲対象範囲、方法、および将来的な方向性について注目した。

国家／多国間システム

2.1GPHIN（Public Health Canada）

2.1.1 背景

システムの歴史

Global Public Health Intelligence Network（GPHIN）[1-3]は、1 日 24 時間、週 7 日間、市民緊急事態の検出、リスク評価、および対応を行っている安定した公衆衛生サーベイランスシステムである。本システムは、カナダ保健省（PHAC）傘下の Centre for Emergency Preparedness and Response（CEPR）によって運営されている。なお、PHAC の責任者は Canadian Minister of Health 直属の Chief Public Health officer である。11 年間におよぶ運用歴がある GPHIN は、緊急対応の技術面および管理面双方に関する専門的な知識を有しており、州および中央政府に対し、初動対応および公衆衛生セキュリティの分野において実践的な支援を提供している。

GPHIN の概念は、1990 年代中頃にまで遡ることができる[4]。1994 年のインドのスーラト（Surat）における肺ペストの流行時に、アウトブレイクの初期段階における報告は、他の公式の公衆衛生ソースよりも CNN International の方が適時的であったことにカナダ保健省の医師 2 名が気付いた。1997 年には、National Health Surveillance Infrastructure[2]のもと、多くのパイロットプロジェクトの予算が組まれた。これらのプロジェクトの中には、GPHIN の最初のプロトタイプシステムが含まれた。この予算は、ウェブを利用して世界的な疾病アウトブレイクに関する情報が適時的に収集できるかどうかを検討するためのものであった。GPHIN は最初のバージョンでは、主にウェブベースの世界的なメディア（全国紙および地方新聞）を英語とフランス語の二言語で精査した。また、確認されていない風説的なニュースについて対処する必要があったため、GPHIN は早期段階において世界保健機関（WHO）とのパートナーシップを確立している。

1998～2001 年の WHO による初期評価では、WHO が確認したアウトブレイクの 56%をこのシステムが検出していたことが示された。この数値からは適時性や特異性については分からないが、たった二言語だけでも、ウェブニュースソースを使用することでイベントを検出可能なレベルにまで技術が成熟していることを示している。これにより、プロトタイプの概念（ウェブソース、プロセス、および基盤）がカナダ国民に対する公衆保健上の脅威の早期警戒として有効であることが確認されることとなった。

2004 年以降、GPHIN は、対応する言語数を増やし、警報に関する規則を調整し、また、人間による分析能力を増強している。ニュース報道のタクソノミーによる分類は、技術の成熟に伴って完全に自動化されており、処理可能なニュース記事の数も急速に増加している。

技術的レベルに目を向けると、システムは、カナダのモントリオールにある Web コンテンツ管理およびテキストマイニング技術を専門とする NStein technologies との協力で開

発されている。

2009 年以降は、NStein の役割はハイエンド技術サポートになり、日々の技術的保守および修理は専門の情報技術担当職員が行う予定である。

図 3：GPHIN の安全なポータルログイン

財源

システムの財源に関しては、これまでに 3 つの大きな変化があった。1997 年から 2003 年ごろまでのプロトタイプシステムは政府の助成金によって運用され、2004 年から 2008 年の間は、Ted Turner 基金、カナダ政府からの助成金、および利用各国からの寄付によって運用された。カナダ政府は、2009 年以降の GPHIN の開発および改良費の新しい財源を模索している。

現在のユーザ

GPHIN は、外務省、カナダ保健省、カナダ農務・農産食品省、カナダ食品検査局、および王室カナダ騎馬警察を含めたカナダ全土の政府機関に情報を提供している。国際パートナーおよび GPHIN 報告の確認者としての WHO の役割は、GPHIN の重要な特徴であり、無視することはできない[4]。また、GPHIN は WHO の Global Outbreak Alert Response Network（GOARN）の一部となっている。報告書は、国連食糧農業機関、国際獣疫事務局、英国健康保護局 HPA（英国）、ECDC（EU）、MedISys（EU）および様々な非政府組織にも提供されている。システムへのユーザアクセスは、安全なポータルサイトを介して提供されている。

なお、GPHIN は、ニュースソースのテキストの著作権に関する制約のため、元のニュース報道を通信社などによる再放送用には公開していない。

GPHIN は、ヒンディー語などの他言語のニーズについて継続して検討している。例えばインドの場合、現在のところ、英語の媒体報道は適時性および対象範囲が良好であることを示している。

2.1.2 対象範囲

言語

GPHIN は、6 つの国連公用語（アラビア語、簡体字中国語、英語、フランス語、ロシア語、およびスペイン語）および繁体字中国語、ポルトガル語（WHO のいくつかの地域委員会の公用語）の分析を行っている。国連公用語は、いくつかの理由から選択されている。(a) 英語ではあまり対象とされていない地域を含めることで世界的なメディアの対象範囲を最大化すること、(b) 英語圏外の地方で発生したイベントが英語報道されるまでの遅延を回避すること、および (c) WHO がシステムの主要ユーザであること。WHO がこれら 6 つの言語が通商および外交の言語とみなされ、多くの国で高レベルの意思決定者および行政官によって使用されている場所[5]におけるその公用語の選択について検討している。これに対し、ベンガル語、ヒンディー語、および日本語は、何百万も人が使用しているが、国際政治ではあまり使われていない。

GPHIN は、ヒンディー語などの他言語のニーズについて継続して検討している。例えばインドの場合、現在のところ、英語の媒体報道は適時性および対象範囲が良好であることを示している。

疾病

ヒトおよび動物のほとんどの感染症（現在オントロジーには 700 のエントリ）、さらに、一部の植物病害を扱っている。自然発生した疾病、および病原が故意にリリースされた疑いのある疾病の両方をモニタリングしている。

その他の公衆衛生上の脅威

GPHIN は、公衆衛生サーベイランスに対してオールリスクアプローチを取っている。これには、化学物質、放射線および核への曝露、バイオテロ、危険な消費製品、および自然災害のモニタリングが含まれている。

地理

GPHIN は、基本的に全世界を対象としているが、南米やカリブ海諸国などの地元のニュース提供者が情報サービス企業と契約していない地域、あるいは GPHIN が現在使用する 9 言語以外の地域については対象力が弱い。

2.1.3 方法

データソース

ニュース記事のフルテキストについては、Al Bawaba（アラビア語）と Factiva（ペルシア語以外のすべての言語）の 2 つの主要な情報サービス会社と契約している。ペルシア語の場合、情報サービス会社から自動的に取得する計画があるが、現在は、人間の分析者がウェブから手動で情報を収集、入力している。現時点では、Al Bawaba および Factiva が、GPHIN の検索構文を 23,000 以上のニュースソース（GPHIN の言語の全国紙および地方新聞の両方）に適用していると推定される。

分析者は、GPHIN が処理する報道に加え、MedISys および欧州連合（EU）の Europe Media Monitor（EMM）からの出力についても検証している。

また、GPHIN では 3 番目の情報ソースとして、疾病の確認を行っている保険機関を介した情報が使用されることがある。この情報は、GPHIN の検証プロセスの一部として使用される。

知識源

GPHIN の自動コンポーネントの中心は、多言語に対応した公衆衛生オントロジーの開発である。オントロジーは、ヒト疾病、動物病、植物病害、化学事故、放射線曝露、危険な製品、および自然災害の分野を対象としている。オントロジーには、用語とともに重みがエンコードされている。システムは、この重みを使用して、報道内容の警戒レベルを検出および評価し、人間の分析者による確認および詳細な調査が必要かどうか判断する。

人間による分析

システムの主要な「コンポーネント」の 1 つは、人間による分析である。現在、生物学、化学、ジャーナリズム、サーベイランス、経済学、および環境科学などの様々な専門知識を有する 14 名が担当している。分析者は多くの役割を担うが、主要な機能は、専門領域の知識、担当地域の理解、そして現在の状況に関する直感に基づいてニュース項目を順位付けすることである。この理解には、アウトブレイクが発生し、対応が行われる地域の社会経済的背景に関する詳細な知識が含まれる。分析者の副次的なタスクには、ニュースソースのマニュアル調査、翻訳システムの出力の確認および編集、特定のクエリに関するユーザからの要求への対応、さらに、検索クエリの開発が含まれる。

2004年以降、GPHINはほぼリアルタイムベースの年中無休対応の提供を目指している。それによってスタッフのスキルの獲得、向上、および維持を明確に強調し、彼らの仕事量を、ニュース提供のピーク時と一致するようなスケジューリングに最大の配慮をもたらすことが出来る。

分析者は通常、ニーズに応じて変更できる柔軟なスケジュールに従い業務を行っている。基本方針は、分析者の生活リズムを「太陽に合わせる」こと、そして、対象言語の地域において最初のニュースが報道された時に分析者が待機していることである。例えば、アラビア語分析者のシフトには、中東および北アフリカのタイムゾーンの違いにより、カナダ時間の夜間および早朝のシフトが含まれる。ピーク時間は、言語や国によって異なる。

図 4：作業中の GPHIN 分析者

ハードウエア

物理的なコンピュータサーバ（プロセッサおよび記憶装置）は、第三者の専門家に外部委託されているため、プロジェクト職員は保全業務から解放されている。各分析者は通常、サーバへの高速ネットワーク通信が可能な仕事場とスクリーンを有し、安全で信頼できる電子メールシステムが使用できる。

アルゴリズム

図4に示されるように、GPHINはモジュール性の高いプロセスによって構成されている。主要な分析は、基本的に以下の2段階に分けることができる。

第1段階：自動分析。情報サービス会社から大量のテキストをダウンロードし、重複ニュース項目を特定および削除する。英語以外のテキストは、機械翻訳（MT）ソフトウエアを使用して英語に翻訳される。翻訳元の言語から英語の翻訳には、様々な最高品質の商用ソフトウエアパッケージが使用されている。その後、テキストはGPHINオントロジーによる重要なキーワードについて、スキャニングにより自動的に分類され、公衆衛生イベントの重要度に基づき、文書にスコアが付与される。分類スコアに基づいて、文書は、掲載（深刻な公衆衛生イベントとしての自動警報ステータス）、チェック（人間の分析者による評価が必要）または破棄（関連性なし）に割り振られる。破棄カテゴリの文書も、体系的な評価および分析者による将来的な参照のために保管されるため、実際にファイルが削除されることはない。関連する記事は、さらに、同様の重み付きキーワードによって自動的に、+動物、+ヒト、+薬品などの非排他的な保健上の脅威カテゴリに分類される。なお、キーワードの重みは、現在の保健上の脅威における優先順位を反映するように随時検証されている。これにより、文書の重要度確認システムの調整機構を簡単かつ柔軟なものにする。

時々、情報サービス会社からの追加的なソースによる補完が必要な場合などの例外が発生することがある。例えば、通常のニュースサービスが制限されている場合。これらの場合、人間の分析者がウェブから新しいソースを入力する。

第2段階：人間による分析。文書の重要度

については、訓練されたバイリンガルの分析者が最終的に判断する。一般に、人間の分析者は2ヵ国語(WHO公用語+英語、または、ペルシア語+英語)を使うことができ、専門領域の教育を受けている。すべての判断は、改訂国際保健規則(2005)を支持してWHO annex 2[6](詳細については[2]を参照)と同様の判断手段に従って行われる。これは、高いレベルの訓練、経験、そして共有された判断によって補完される。

エンドユーザへの出力は、警報、生物地理学マップ、およびランク付けされた検索可能な記事リストなどの様々な形式で、安全なポータルサイトを介して提供される。さらに、ユーザは独自の検索クエリを指定することができる。ユーザは、元のニュース記事および商用MTソフトウエアを使用して英訳された記事の要約にアクセスすることができる。MT出力の確認は、別の自動ソフトウエアモジュールによって行われる。なお、分かりにくいと判断された場合には、記事にフラグが立てられる。これらは、分析者が対処する。

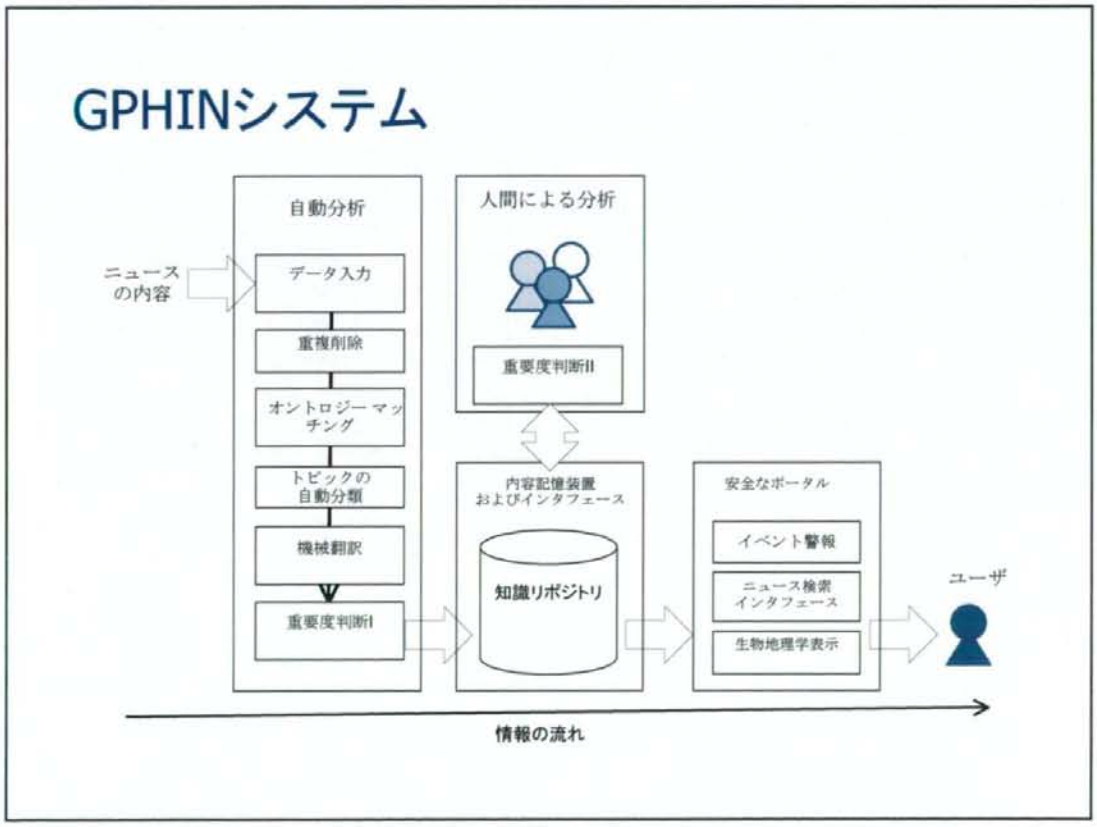

図5：GPHINシステムにおける情報の流れ

2.1.4 考察

成功事例

GPHIN の最も大きな成功と見なされている事例は、2002 年 11 月の中国南部の広東省で発生した異常なイベントの検出である[4]。この中国語新聞の最初の報道が、当時、既に拡がりつつあった SARS 流行の第一報であった。対照的に、英語で最初に報道されたのは 2003 年 1 月であり、迅速なサーベイランスにおける多言語能力の必要性を裏付けている。その後の流行中も、GPHIN は、ほぼリアルタイムで SARS 疑惑例（未確認）に関する情報を提供し続けた。これらの情報は、WHO の症例報告よりも 2～3 日早かった。Mawudeku および Blench[1]が、風説的なニュースソースから症例数を算出したところ、症例数は、実際の数よりも約 50%上回っていたが、その動向は WHO の公式症例数の動向とほぼ一致していた。

より大きな観点からは、GPHIN は、WHO と GOARN における従来のコミュニケーションを変更し、国際的な公衆衛生イベントの報告における遅延を減らしたグローバル・ヘルスサーベイランスの情報収集および共有における分岐点であると言える。この影響は、WHO／GOARN システムに流れ込んだ報告の量に見ることができる。2004 年に、Mykhalovskiy および Weir[4]は、「GPHIN は、WHO の早期警戒アウトブレイク情報の約 40%を提供している」と報告している。

将来的な方向性

2009 年には、GPHIN の開発および強化の費用を支えるための新しい財源モデルが明らかにされる予定である。これにより、システムを維持するための新しい技術を導入することが可能になる。

2007 年に施行された 2005 年の国際保健規則は、GPHIN の新しい方向性を示している。GPHIN は、国の新しい要件の支援を助け、カナダにおける包括的な保健サーベイランス基盤と、それが経済的に実現不可能である地域へと拡大する予定である。

参考文献・資料

[1] Mawudeku, A. and Blench, M. (2006), "Global Public Health Intelligence Network (GPHIN)", Proc. 7th Conference of the Association for Machine Translation in the Americas, 8-12 August, Cambridge, MA, USA. Available from http://www.mt-archive.info/MTS-2005-Mawudeku.pdf (last accessed 15th December 2008).

[2] Mawudeku, A., Lemay, R., Werker, D., Andraghetti, R. and St. John, R. (2007), "The Global Public Health Intelligence Network", in Infectious Disease Surveillance, M'ikantha, N., Lynfield, R., Van Beneden, C. and de Valk, H. (eds), Blackwell Publishing, pp. 304-317.

[3] Eysenbach, G. (2003), "SARS and population health technology", J. Medical Internet Research, 5. Available from http://www.jmir.org/2003/2/e14/ (last accessed 15th December 2008).

[4] Mykhalovskiy, E. and Weir, L. (2006), "The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health", Canadian J. of Public Health, 97(1), pp. 42-44.

[5] World Health Organization, (1999), "Reaching out to the largest audience: languages for communication in WHO", Technical report EV105/20, Executive Board, 105th Session, Provisional agenda item 7.2, 24th November. Available from http://ftp.who.int/gb/archive/pdf_files/EB105/ee20.pdf (last accessed 15th December 2008).

[6] World Health Organization, (2005), "Decision instrument for the assessment and notification of events that may constitute a public health emergency of international concern", Intergovernmental working group on revision of the International Health Regulations,

Technical report A/IHR/IGWG/2/INF.DOC./4, 22nd February. Available from http://www.who.int/gb/ghs/pdf/IHR_IGWG2_ID4-en.pdf (last accessed 17th December 2008).

2.2 ARGUS (Georgetown University, USA)

2.2.1 背景

システムの歴史

Georgetown University Medical Center の Project Argus[1,2]は、2004 年以降、公開されているウェブメディアから疾病アウトブレイクの迅速な検出を行っている。システムは、主に、米国外の生物学的イベントを検出および追跡している。米国内のサーベイランスは、疾病対策センターやその他の米国連邦政府機関が担当している。Project Argus は、自動分析と人間による分析の混合システムを運用しており、自動的に収集した情報と、専門分析者チームを利用するエンドユーザとの間を仲介する、現在存在する 2 つのうち 1 つのシステムである。

歴史的な観点からは、Argus の最初の自動分析技術は、2001～2004 年に構築された MITRE Corporation[3-4]（下記参照）の MiTAP（MITRE Text and Audio Processing）システムの採用にまで遡る。現在、そのオリジナルの MiTAP システムの中核的な部分は、World Wide Web から様々な言語のニュース報道を収集するために使用されている。収集された情報は、人間の分析者の各ニーズに合わせてフィルタリングされる。

Argus は、2004 年に早期警戒インジケータのプロトタイプとして確立されており、その後、プロトタイプが成功したとの判断をもって、2006 年にフル稼働モードに移行した。この年に、Project Argus が H5N1 型鳥インフルエンザの検出および追跡のために樹立された。Argus の能力は、約 150 のヒトおよび動植物の感染症にまで拡張された。

Argus の人間とコンピュータの両方を使用する分析方法にはいくつか特徴があり、これらは注目に値する。(a) 自然災害および気象学コミュニティで使用される方法に基づく段階モデル[5]を使用した生物学的イベントの記述、および (b) 直接的、間接的な環境、および気象インジケータの統合。ここでは、直接的インジケータは、臨床診断、地理、罹患／死亡例数、および人口インジケータなどの標準的な疫学的インジケータをさす。対照的に、間接的インジケータには、地震、洪水、および他の自然災害、さらに、社会的混乱（例：病院の閉鎖、一般用医薬品の不足）、公式発表、業務習慣の変更などの前駆的なイベントが含まれる。なお、(b) が開発された背景には 9/11 委員会報告書の提言があり、とくに突発的な感染症の検出に注力している。

Project Argus は、自動化されたコンポーネントに加え、人間による分析も重視している。40 名以上の分析者がいるため、言語的および文化的知識は広範囲におよび、約 40 の言語に対応することができる。分析者に加え、チームには、コンピュータシステムの保守および開発を担当する 8ー10 名の技術専門家、さらに、中心的な管理チームである公衆衛生に関する専門家が含まれる。

MiTAP に関する注記：MiTAP システム[3,4]は、2001ー2004 年の間に、Defense Advanced Research Projects Agency（DARPA）の助成金の下、MITRE Corporation によって Translingual Information Detection Extraction and Summarization（TIDES）[多言語情報の検出、抽出、および要約]プログラムとして開発された。この時期における MiTAP の目的は、「分析者、医療専門家、医療サービス、および人道支援および救済事業に関与する人員に、多言語情報への適時的なアクセスを提供すること」であった[3]。計画は、ウェブからのニュースおよび他の情報（例：株価指標、交通データ）などを捕捉、フィルタリング、翻訳、分類、および要約するための技術を融合するシステムの開発することであ

った。これには、現在テキストマイニングとして知られる、構造化されていない疾病イベントに関する情報の構造化フォーマットへの変換技術が必要であった。計画では、専門家が設定したインジケータおよび警戒閾値に基づくニュースイベントの自動分析の提示が含まれていた。イベントが検出されると、専門家によって書かれた重み付けルールに基づく自動コンポーネントが警報の発令について判断する。2002 年に起動された最初のプロトタイプには、現在ほどではないが、データ収集コンポーネント、機械翻訳エンジン、ルールベースのエキスパートシステム、人間がデータ空間を視覚的な要約を介して確認することができるグラフィカルユーザインタフェース（GUI）などの強力な機能が備わっていた。

現在のユーザ

現在、Argus へのアクセスは、特定の米国連邦政府機関に制限されており、毎日の各国の報告書と安全なポータルサイトを介したシステムへのログインが前提とされている。Argus の興味深い側面の 1 つは、警報ガイドラインの策定およびユーザの改訂への関与である。

2.2.2 対象範囲

言語

Argus の人間による分析は約 40 言語、インデックス化された記事検索システムは 29 言語、そして、機械翻訳は 13 言語に対応している。これらは、ソーステキストおよび地域担当分析者の可用性に基づいて選択された。

疾病

オントロジーには、現在、ヒト、動植物の病原を含む約 150 の急速に拡散する疾病が含まれている。選択は、現在のプロジェクトの主要な使命に基づいているが、さらに、生物学的脅威の範囲を自主的に拡張して社会経済リスクの高い病原を含めている。

その他の公衆衛生上の脅威

Project Argus は、主に健康に対する生物学的脅威の分析を行っている。

地理

本システムは、米国を除く世界のほぼ全て（200 ヵ国以上）を対象としている。

2.2.3 方法

データソース

システムは、独自に開発したウェブ巡回エンジンを使用している。このエンジンは、1 日あたり約 250,000 の記事から 130 の報告を作成している。保全タスクには、ウェブのメディアソースからの直接的な情報の収集および不要データの削除など多くの作業が含まれる。他のシステムとは異なり、Project Argus は現在、外部の情報サービスを利用していない。

知識源

Argus の自動コンポーネントは、MiTAP 時（2004 年以前）からさらに洗練されており、公衆衛生用の Argus オントロジーから様々な語彙および重み付け知識が組み込まれている。オントロジーは、過去、多くの場合においてインターネットが普及する前にまで遡って、アウトブレイクの報告を詳細かつ継続的に検証した回顧的調査の結果である。

オントロジーには、ニュース中のイベントを特定し、それらを段階モデルに関連付けるために手動で開発された約 200 のインジケータおよび警告が含まれる[5]。Wilson らは、この段階的アプローチは米国気象局が採用している方法に類似していると述べている。インジケータは、社会的混乱の基礎的な社会学の考えに基づいて開発され[6]、過去の流

行の回顧的解析に対して適応され、また、新しい予測的データを使用して検証されている。Wilson らが解説したイベントの段階化[5]を以下に示す。

P. Preparatory posture（準備姿勢） – 国が、生物学的イベントを予想して予備姿勢をとる。

0. Environmental conditions favorable（環境条件による示唆） –大規模洪水などの疾病リスクの高まりがあるイベント前の状態。

1. Unifocal event（単焦点イベント） – インジケータが単一の医療施設における症例の報告を示す生物学的アウトブレイクイベントの始まり。

2. Multifocal event（多焦点イベント） – インジケータが複数の医療施設における症例の報告を示すが、流行は抑制されていると考えられる

3. Severe infrastructure strain（重度の基盤負荷） – インジケータが、医療基盤が耐えられないほどの負荷となる抑制されていない複数の多焦点イベントの発生を示している。

4. Social collapse（社会的崩壊） – インジケータが抑制されていない多焦点イベントの結果としてある程度の社会崩壊が発生していることを示している。アウトブレイクを抑制しようとする努力にかかわらず起こることがある。

R - Recovery（復旧） – 生物学的アウトブレイクがもはや存在しないことを示す。

分析者によって確認されたイベントが段階化される。その後、各国の報告書は、1つ以上の警告基準を満たすイベントの概要としてまとめられ、システム内において、イベントの重要度に応じて赤、オレンジ、あるいは黄色で示される。

警告基準は、ユーザコミュニティにおける現在の懸念を反映するように、米国連邦政府機関のユーザグループとともに継続して検討されている。

人間による分析

システムでは、約 40 名の人間の分析者が地域チームに分かれて業務を担当している。現在、経験豊かな分析者 2 名が、分析者のスキルの訓練および開発を担当している。地域チームは、通常とは異なるイベントの報告書をまとめる。この報告書には 2 番目の優先度が割り当てられ、各国の報告書に記載される前に、上級分析者によって詳細な調査が行われる。これらについては、ユーザコミュニティに公表される前に、主席分析者によって最終的な検証が行われる。

分析者の雇用では、地域に関する専門知識、言語能力、および公衆衛生などの専門領域におけるスキルを重視する。公衆衛生上の緊急事態に対する正常あるいは異常な反応について理解する能力の文化的な側面は、社会経済ストレスの間接的な信号を介した早期警報インジケータの検出と密接に関連する。これらのスキルは、疾病アウトブレイクの直接的なインジケータとともに、訓練および共有された協議を介して磨かれる。

ハードウエア

ウェブから加工される規模の情報を扱うために、技術スタッフによってクラスタコンピュータが内部に維持されている。

アルゴリズム

基本的に Argus には 2 つの段階があるが、

明らかに人間による分析を重視している。

第一段階：自動分析。複数の言語の未処理テキストがニュースサイトから短いサイクルで収集される。特定の 13 言語に関しては、商用 MT ソフトウエアを使用して翻訳される。不要データの削除は、それぞれに合わせて作成されたスクリプトを使用して行われる。これらのスクリプトは、技術スタッフが使いやすいインタフェースを使用してオフラインで保守している。これにより、頻繁に行われるニュースサイトのフォーマットの変更に前処理ソフトウエアを適応させるための負担が軽減される。ダウンロードおよび整形されたテキストは、その後、オントロジーからの重み付けされたキーワードについてスキャンされて自動的に分類され（上記参照）、文書には公衆衛生イベントの重要度に基づくスコアが付けられる。文書は、スコアに基づいて要警報または警報不要に分類される。

第二段階：人間による分析。人間の分析者は、2 つの方法でニュースデータにアクセスすることができる。(a) 第一段階で自動的に検出された項目、および (b) 個々に作成されたクエリを使用して新しいデータベースから得られた順位付けされた検索結果。(b) では、複数の言語を使える人間の分析者が独自の検索をブールクエリとして定義する。これらのクエリは予定された時間にクラスタコンピュータ上でバッチ処理される。分析者は、カスタマイズされたインタフェースを使用することで、クエリ検索の結果を順位付けされた見出しと他補足情報のリストとして表示することができる。

分析者は、1 つ以上の外国語に堪能である。人間の分析者は地域チームとして働きながら、有意な記事を特定し、インジケータおよび警告を考慮しながら潜在的な生物学的脅威について評価する。警告レベルについては上級分析者に報告され、ユーザに公開される毎日の国別報告書への掲載について判断される。さらに、すべての分析者が地域責任者、訓練担当者、および副主任分析者との会議に毎日参加して情報を共有している。

2.2.4 考察

Argus チームは、回顧的解析を使用して生物学的イベントに伴う社会的混乱に関する分析を継続的に深め、既存のオントロジー、語彙項目と段階モデルの関係を精製している。維持や開発を手がける他のグループ同様に、オントロジーは、新しいケースを迅速に入力および分類する新技術により、恩恵を受けている。また、チームはこの分野において積極的に活動している。

参考文献・資料

[1] Wilson, J. (2007), "Argus: A Global Detection and Tracking System for Biological Events", Advances in Disease Surveillance, v.4, pp. 21.

[2] Wilson, J. (2007), "Global Monitoring for Disease Outbreaks: Project Argus", Published as a National Institute of Health VideoCast, 29th November. Available from http://videocast.nih.gov/Summary.asp?file=14175 (last accessed 11th December 2008).

[3] Damianos, L., Zarrella, G. and Hirschman, L. (2004), "The MiTAP System for Monitoring Reports of Disease Outbreaks", MITRE technical paper, August. Available from http://www.mitre.org/work/tech_papers/tech_papers_04/04_0804/index.html (last accessed 11th December 2008).

[4] Damianos, L., Ponte, J. Wohlever, S., Reeder, F., Day, D., Wilson, G. and Hirschman, L. (2002), "MiTAP for Bio-Security: A Case Study", AI

Magazine, 23(4), pp. 13-29.

[5] Wilson, J., Polyak, M., Blake, J. and Collmann, J. (2008), "A heuristic indication and warning staging model for detection and assessment of biological events", J. American Medical Informatics Association, 15:158-171, DOI:10.1197/jamia.M2558, pp. 158-171.

[6] Strauss, A., Corbin, J. (1990), "Basics of qualitative research: grounded theory procedures and techniques", Published by Sage, Newbury Park, Ca.

2.3 MedISys／PULS（Joint Research Centre, Italy および Helsinki University, Finland）

2.3.1 背景

システムの歴史： MedISys

欧州委員会の（EC）Medical Information System（MedISys）[1-5]は、イタリアのイスプラにある Institute for the Protection and Security of the Citizen（IPSC）の Joint Research Centre（JRC）によって運用される年中無休で稼働する全自動公衆衛生サーベイランスシステムである。Directorate General for Health and Consumer Protection（DG SANCO）の Health Threats Unit との密接な協力関係にある。システムの使命は、感染症サーベイランスからの早期警戒およびバイオテロの検出であり、公開ニュースメディアにおいて報道される自然発生的および故意の化学的、生物学的な放射線、および核（CBRN）の脅威が対象に含まれる。人間に対する脅威に加え、生物学的脅威には様々な動物の病気にも拡大されている。2004 年 8 月に起動された MedISys は、JRC が開発したメディアモニタリングアプリケーションの 1 つであり、Europe Media Monitor（EMM）によって収集されたニュースを処理している[6]。EMM の最初のアプリケーション（NewsBrief、2002 年以降オンライン稼働中）は、EU 機関および EU 加盟国による様々な地域における報道のモニタリングを EC の Directorate General Communication（DG COMM）が支援するために開発された。このソフトウエアは、DG SANCO のイニシアチブにより公衆衛生領域において採用され、MedISys の創設に至った。MedISys は技術の大半がメディアのモニタリング用の姉妹アプリケーションと共有されるため、1 つのアプリケーションの開発が他のアプリケーションに直接恩恵をもたらすこととなる。

これまでに検証した米国およびカナダにおける政府システムとの比較において、MedISys の特徴的な違いの 1 つは、分析プロセスに分析者が介在しないことである。システムは、欧州委員会および衛生行政機関における下流ユーザにこの作業を委譲するように設計されている。そのようなユーザは、「Rapid News System」（RNS）と呼ばれるサービスアプリケーションにアクセスする。RNS では、ユーザは手動で、ユーザ定義のカテゴリに記事の選択、記事へのコメント、フォーマット化されたニュースレターの作成、そして、ユーザ定義グループへの送信が可能である。ユーザは、カテゴリ定義を改善することもできる。メインシステムにおける早期警戒機能は、特定のトピックに関する情報の突発的な増加を自動的に検出することによって達成される。

MedISys は、欧州委員会が提供する JRC 組織の予算から資金を得ている。追加的な財源は、DG SANCO によって提供されている。

2007 年以降、MedISys は、University of Helsinki が開発した文書内容の正確な言語学的分析を行うことで症例数や地理的位置などの生物学的イベントに関する事実を判断するテキストマイニング システム Pattern-based Understanding and Learning System（PULS）と連携している。両システムからの結果は、IPSC によって運用される共通のポータルサイトにおいて報告される。2 つのシステムの関係を図 6 に、また、ポータルの画像を図 7 に示す。両システムについては以下で詳細に取り上げる。2 つのシステムにおいて技術的に大きく異なる点については、それぞれのセクションに分けて取り上げる。

MedISys のコアチームは、研究開発職員を含む 5 名ほどのメンバーで構成され、メン