④共通語彙やデータ連携標準の整備

データのオープン化に関しては、分野間、組織間のデータ連携を容易にするため、データ記述を正確に行うための「共通語彙」など「データ連携標準」を定める必要がある。2005年に米国は、NIEM（National Information Exchange Model）[※50]と呼ばれる情報交換フレームワークの策定を開始し、情報共有の自動化を目指した。NIEMは、米国司法省のGlobal Justice XMLデータモデル（GJXDM）がベースとなり、公共安全や災害管理などの分野で米国の全てのレベルの政府組織や省庁、民間企業などを対象としている。

またEU圏では、2011年にEU圏の行政機関間でのデータ交換のためにSEMIC（Semantic Interoperability Community）[※51]を開始した。SEMICでは、EU加盟国、EU行政機関でのセマンティック層での共通の定義と仕様の策定を行う。

日本では、2018年にIMI（Infrastructure for Multilayer Interoperability；情報共有基盤）[※52]の策定を開始している。これは、政府の『未来投資戦略2018 ―「Society 5.0」「データ駆動型社会」への変革 ―』（2018年6月15日閣議決定）を受け開始されたものである。3年以内の整備、5年以内の本格稼働を目指す。様々なテータモデル記述（Data Model Description；DMD）を策定し、データ交換時に利用することで、様々な組織が持つ情報の共有を可能にする。また、行政で用いられる人名漢字など6万文字の漢字を整備した「文字情報基盤[※53]」も策定する。

データ処理基盤技術

「データ処理基盤技術」は、大量データを実際に扱うための技術である。大量のデータやリアルタイムに発生するデータの処理を従来のように行っていたのでは、実用的な時間内に処理が完了する可能性が低い。このため、「分散並列処理」、「圧縮データ処理」、「ストリーム処理」などが利用される。「分散並列処理」では、複数箇所で処理を独立・並列に行う。例えば、中央（クラウド）でカメラ画像の認識と判断をまとめて行うと過大な処理が発生する。しかし、複数のカメラ画像をとりまとめるノード（エッジ）で認識処理を行うことで、中央では認識結果を利用した判断を行えばよくなるため、負荷を抑えることができる。

「圧縮データ処理」とは、データを圧縮したまま処理する方式一般をいう。外部記憶に蓄積されたデータをそのままメモリに読み込むことができれば、伸張されたデータを扱う場合に比べて読込み完了までの時間を短縮できる。書込みについても同様であり、また、データを送受信する場合においても圧縮状態のデータにはメリットがある。

「ストリーム処理」は、データを蓄積することなく、リアルタイム時間で処理し、そのまま次の処理へ手渡す方式である。外部記憶への書き込み、読み出しが入らないことで、処理時間を短縮できる。反面、リアルタイムに処理を行えるだけの性能がハードウェアに求められる。しかし、大量のデータを受け取るための外部記憶が不要になることもあり、必ずしもシステム全体のコスト上昇につながるとは限らない。前述の分散並列処理と組み合わせるなどして、個々のノードのストリーム処理の負荷を一定以下に抑えることも可能である。

※50 https://www.niem.gov/
※51 https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic
※52 https://imi.go.jp/
※53 https://mojikiban.ipa.go.jp/

データ保護技術

データ保護技術では、個人の特定を不可能にする「プライバシー保護」、データの「漏洩対策」などのセキュリティ面と「ビッグデータに対する処理効率」を両立させる必要がある。

プライバシー保護上の一般的な方法として、データに対して、個人を特定できるような情報を秘匿する「データ匿名化」があるが、データ中の電話番号など、直接個人を特定できる情報（識別子）を秘匿したとしても、例えば希少疾患の病歴、高額所得といったデータと年齢、性別、地域などの情報（準識別子）を組み合わせることで、個人を特定できてしまう可能性がある。

こうした場合を防ぐ手法としてk-匿名性（k-anonymity）[1]がある。k-匿名性は、データ中の準識別子の任意の組み合わせによる検索結果がk件以上になるという匿名性の指標であり、データを公開するときに、k-匿名性を満たすように準識別子を加工することで、匿名性が確保されるという考えである。しかし実際には、より匿名性を高めるため、l-多様性（l-diversity；同じ準識別子の組み合わせデータのグループに対してl個以上のセンシティブ情報が入る[2]）、t-近接性（t-closeness；同じ準識別子の組み合わせデータのグループ内のセンシティブ情報の分布とデータ全体のセンシティブ情報の分布の差がt以下[3]）といった指標を用いた匿名化が行われることが多い。

このほか、データへの問い合わせに対する回答に特定個人の情報が寄与する「敏感度」を求め、ノイズなどを回答に含めることで、全てのユーザーの敏感度を下げた回答を生成する「差分プライバシー（Differential privacy）」[4]といった手法もある。

また、異なる組織が管理するデータを用いて統計処理を行う場合、暗号化されたデータを複号することで、互いのデータが見れてしまう。こうした問題を解決する手法として、データを暗号化したまま計算する「秘密計算」という手法がある。秘密計算では、暗号化されたデータをそのまま計算に使うことで、計算経過も秘匿することが可能となる。秘密計算では、暗号化により数学的構造が変わらない「準同型暗号方式」を使うもの、データを分割して管理し、その管理ノード同士が協調して統計処理などを行う“Secure multi-party computation”方式（秘密分散方式）などがある。

テキストからの知識の獲得

知識のためのビッグデータとしては、インターネットに存在するWebページなどが使われることも多いが、内容に誤りも多く、量は豊富にあるものの、質としては高いとはいえない。これに対して、誤りが少ないと考えられるのが「科学技術論文」である。最近では、様々な分野でネットでの発刊が可能になり、ビッグデータ化された。しかし、研究者個人が読むことができる量は限られており、論文量が可読量を超えつつある。このため、大量の論文からテキストマイニングを行うなど、大量の論文そのものをビッグデータとして扱ったり、知識構築のソースとしたりすることで、知識として獲得することが考えられている。すでに医療分野では、医学／医療論文データベースであるMEDLINEでこうした試みが行われている。

自然言語からの知識の獲得に関しては、ディープラーニングにより、End-to-Endのニューラル自然言語処理が見出され、従来手法よりも高い精度を上げているものの、文脈の理解や常識に基づいた推論といった、従来から困難とされてきた問題については、ディープラーニングを利用しても困難なままである。常識推論は、自然言語処理の初期段階から取り組まれてきた問題ではあるが、含意関係認識（文Aが正しいときに文Bも正しい関係にあるかどうかを判定する）、ス