て「知識」の表現の共通化といった問題がある。残りの2つは、大量データを実用時間内に処理するための「データ処理基盤技術」、そして、データに含まれる個人に関わる情報を保護しつつ、データを有効に活用する「データ保護技術」である。AIとデータ及び知識の関わりの歴史的発展については、『AI白書2019』「2.4 知識処理とデータ」を参照されたい。

(2)主な技術

ここでは、前述のオープンデータ技術、データ処理基盤技術、データ保護技術、ディープラーニングを活用したテキストからの知識獲得、グラフ構造に着目した知識グラフについて説明する。

オープンデータ技術

①セマンティックWeb

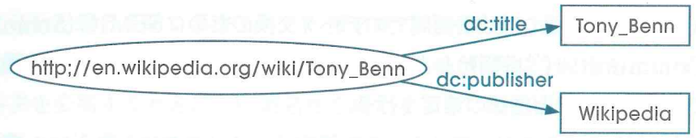
Webの提唱者であるTim Berners-Leeは1999年にセマンティックWeb構想を発表した[※43]。通常のWebは人が読む文書情報をHTML形式で表すのに対し、セマンティックWebは、「情報リソースに意味(セマンティック)を付与することで、人を介さずに、コンピューターが自律的に処理できるようにするための技術」である。セマンティックWebに必要なものは、Webページに対するメタ情報と後述のオントロジーである。メタ情報は「XML」ならびに「RDF」で定義することができ、オントロジーは「RDF Schema」ならびに「OWL(Web Ontology Language)」[※44]で定義することができる。

RDFは主語(subject)、述語(predicate)、目的語(object)の三つ組(triplet)でデータ/知識を表す。**図2-2-22**にRDFの例を示す。この例は、あるWebページに対するタイトル(title)と出版社(publisher)に関するメタ情報を定義している。図の下段がグラフ表示であり、上がNotation3(N3)というXMLをベースとするRDFのシリアル化記述形である。グラフの楕円ノードはURI(Uniform Resource Identifier)であり、これはWeb空間でユニークなIDとなる。上のNotation3記述の2行目、3行目はそれぞれrdfとdc(Dublin Coreという書誌情報メタデータ記述用の関係語彙セット)のXML名前空間(xmlns)が＝の右側のURIで定められることを記している。4行目の <rdf:Description rdf:about= " …… " > はRDF三つ組構造データの主語(subject)項目を記し、その下の2行で dc:title と dc:publisher (dc名前空間のtitleとpublisher)の述語(predicate)関係を持つ目的語(object)に当たるリテラル(文字列か数値)「Tony_Benn」と「Wikipedia」に関係づけている。

※43 Tim Berners-Lee, "The Semantic Web" (PDF). Scientific American. May 17,2001.
※44 https://www.w3.org/2001/sw/wiki/OWL

■図2-2-22 RDFの例ー下がグラフ表示で上がNotation3(N3)記述

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/l. l/">
  <rdf: Description rdf :about="http://en.wikipedia.org/wiki/Tony _Benn">
    <dc:title> Tony Benn</dc:title>
    <dc:publisher>Wikipedia</dc:publisher>
  </rdf :Description>
</rdf:RDF>
```



出典:Wikipedia記事「ファイル:Rdf-graph-example-TonyBenn.png」より作成

②オントロジー

AIにおけるオントロジーとは「概念化の明示的な仕様」(Tom Gruber)[※45]であり、共通の概念の体系("語彙"とその定義とそれらの関係)のことを指す。単に知識を集めるだけではなくて、それを活用すること(検索や推論など)に重点が置かれた。オントロジーには様々な構築方法が提案されているが、その一つにOWLがある。OWLは、W3C[※46]のセマンティックWebプロジェクトに由来する。同プロジェクトは、データのネットワークであるWWW[※47]に対してデータ同士の関係を示すメタデータを付与して、意味を解釈、処理することを目指した。このときに使われたのがOWLである。OWLは、DAML(DARPA Agent Markup Language;Webの機械可読表現を目的とした言語)+OIL(Ontology Inference Layer)をベースに開発された。OILに"Inference"(推論)が含まれていることから分かるように、オントロジーを利用した様々な推論ができる。

③LOD(Linked Open Data)

既存のデータに対して、メタデータ記述形式RDFを利用し、コンピューターがデータの意味を判読してWebのようにオープンにアクセスできるようにしたのが、LOD(Linked Open Data[※48])と呼ばれるデータのWebである。既存のWikipediaのページをLOD化したDBpediaがつくられるなど、LODによるオープンデータの公開は増えつつある。LODは2019年3月時点で1239データセットが公開されており、総体としてLinked Open Data (LOD) Cloudと呼ばれる。

LODが使うRDFに対する問い合わせ言語がSPARQL(SPARQL Protocol and RDF Query Language[※49])で、LODは標準検索APIとしてSPARQLエンドポイントを持ち、ここからSPARQLによる検索が可能になっている(例:作者Aの著作のうちノーベル賞作家と共著の著作はどれ?など)。

※45 http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
※46 The World Wide Web Consortium (W3C):Web技術の標準を策定する国際団体。
※47 World Wide Web(WWW)とは、インターネット上で提供されているハイパーテキストシステムである。Web、ウェブとも呼ばれる。
※48 https://www.w3.org/DesignIssues/LinkedData.html
※49 https://www.w3.org/2001/sw/wiki/SPARQL