Inhaltsverzeichnis

[Die Graphentheorie 2](#_Toc202173539)

[Semantik Web 4](#_Toc202173540)

[Normdaten 5](#_Toc202173541)

[XML 6](#_Toc202173542)

[RDF 7](#_Toc202173543)

[Wikidata 8](#_Toc202173544)

## Semantic Web

Das menschliche Gedächtnis setzt sich aus zwei zentralen Komponenten zusammen: einem semantischen Teil für allgemeines, kulturell geteiltes Wissen und einem episodischen Teil für persönlich Erlebtes. Diese Kombination ermöglicht es uns, Bedeutung zu erkenn und Erinnerung einzuordnen. Wörter und Phrasen können je nach Kontext verschiedene Bedeutungen haben und ohne Kontext können die Bedeutung vieler Ausdrücke unklar bleiben. Selbst strukturierte Informationen wie Zeitangaben entfalten ihre Bedeutung nur, wenn man die dahinterliegende Syntax kennt und weiß, wie diese Information zu lesen ist. Syntax gibt die Regeln vor, nach denen Zeichen geordnet sind, während Semantik ihre inhaltliche Bedeutung beschreibt. Pragmatik geht über die Bedeutungszuweisung hinaus und fragt nach dem Zweck und der Absicht sprachlicher Äußerungen. Sie ist entscheidend für eine vollständige Interpretation von Sprache im Handlungszusammenhang. Während Menschen über kognitive Fähigkeiten verfügen, um sprachliche Bedeutung intuitiv zu erfassen, muss diese Fähigkeit bei Maschinen künstlich nachgebildet oder kompensiert werden (vgl. Dengel 2012: 10–11, 13). Damit Computer Bedeutungen erkennen und verarbeiten können, muss Wissen formalisiert und in eine strukturierte Form gebracht werden. Nur so ist es möglich, Objekte maschinell zu kategorisieren, Relationen herzustellen und Kontexte zu berücksichtigen (vgl. ebd. 21).

Wissen soll nicht mehr als statische Ansammlung von Fakten verstanden werden, sondern als ein dynamisches Netz miteinander verbundener Information. In unterschiedlichen wissenschaftlichen Disziplinen wird zunehmend anerkannt, dass Objekte viele Bedeutungen und Verbindungen aufweisen (vgl. Werner 2020: 247). Dies macht deutlich, dass Daten ohne ihren Kontext kaum aussagekräftig sind (vgl. Dengel 2012: 5). Durch technologische Entwicklungen ist es möglich, Metadaten systematisch mit externem Wissen zu vernetzen (vgl. Werner 2020: 247). Dabei entsteht ein semantischer Raum, in dem Daten nicht isoliert, sondern relational verstanden und genutzt werden (vgl. Dengel 2012: 5). Diese Anforderung bildet die Grundlage für Konzepte wie Semantic Web.

Das Konzept Semantic Web geht auf Tim Berners-Lee zurück und zielt darauf ab, Informationen im World Wide Web automatisch und maschinell interpretierbar zu machen (vgl. Dörpinghaus 2022). Mit den Technologien des Semantic Web können Nutzer:innen Datenspeicher im Web aufbauen, Vokabulare entwickeln und Regeln zur Datenverarbeitung festlegen (vgl. Kamzelak 2018: 424). Es erweitert das bestehende Web, nutzt dessen globale Indexierungs- und Benennungssysteme und ist sowohl mit herkömmlichen Browsern als auch mit semantisch-fähiger Software zugänglich (Domingue/Fensel/Hendler 2011: 5).

In der Informatik wird unter dem Begriff *Semantik* die Verknüpfung von Informationen verstanden, wobei der Fokus auf den Regeln zur Strukturierung und Interpretation dieser Daten liegt. Ziel ist es, gleichartige Informationen, die an unterschiedlichen Orten gespeichert sind, miteinander in Beziehung zu setzen (vgl. Kamzelak 2018: 424). Im Kontext des Semantic Web soll jeder Begriff eine eindeutige Kennung erhalten, um Identitäten klar aufzulösen (Domingue/Fensel/ Hendler 2011: 5). Ein zentrales Element ist dabei die Repräsentation von Zusatzinformationen (Metadaten) in einer Form, die von Maschinen interpretiert werden kann (vgl. Grütter 2008: 75). Semantische Metainformationen dienen dabei als Verknüpfung der Daten, die idealerweise öffentlich zugänglich sind. Dies bildet die Grundlage für ein maschinenlesbares Wissensnetzwerk (vgl. Dörpinghaus 2022). Solche semantischen Netze organisieren Wissen durch strukturierte, graphische Darstellung, meist in Form von Knoten und Kanten sog. Graphen (vgl. Dengel 2012: 34).

Hier schließt die Graphentheorie als methodisches Fundament an: Sie ist ein junges Teilgebiet der Mathematik, das ursprünglich durch praktische Problemstellungen inspiriert wurde. Ihre Relevanz für die Informatik ergibt sich aus der Schnittstelle zwischen mathematischer Struktur und algorithmischer Verarbeitung. Während sich die Mathematik vor allem mit der Struktur und den Eigenschaften von Graphen befasst, konzentriert sich die Informatik auf deren Berechnung und Verarbeitung. Die Ideen der Graphentheorie gehen auf Leonhard Euler (1736) zurück, wobei der Begriff *Graph* erst 1878 eingefügt wurde. Mit der digitalen Entwicklung erhielt die Graphentheorie neue Impulse, insbesondere durch: (a) Reflexion wissenschaftlicher Methoden im Netzwerkumgang; (b) Übernahme von Linked Data-Ansätzen aus der Datenwissenschaft und (c) kritische Methodendiskussion in der Netzwerkanalyse (vgl. Dörpinghaus 2022). Die Graphentheorie eröffnet vielfältige Perspektive auf gesellschaftlichen Strukturen und Prozesse. Darüber hinaus lässt sich *Gesellschaft* selbst als ein komplexes, dynamisches Netzwerk modellieren, das durch graphentheoretische Modelle systematisch beschrieben und untersucht werden kann (vgl. Schmidt 2023: 5).

Grundlage dieser Analysen ist das abstrakte Konzept des Graphen, der als mathematische Modelle zur formalen Beschreibung netzartiger Strukturen dient. Solche Systeme zeichnen sich typischerweise durch Zusammensetzung von zwei strukturell unterschiedlichen Objektklassen. In der Terminologie der Graphentheorie werden diese beiden Elementen als Knoten (Objekte) und Kanten (Verbindungen bzw. Beziehungen) bezeichnet (vgl. Tittmann 2019: 11), wobei die Knoten Entitäten eines bestimmten Bereichs wie bspw. eine Person oder ein Produkt repräsentieren. Die Kanten verknüpfen diese Entitäten miteinander und geben die Art der Verbindung zwischen ihnen an (vgl. Barrasa/ Webber 2023: [Introducing Knowlege Graphs](https://learning.oreilly.com/library/view/building-knowledge-graphs/9781098127091/ch01.html#idm45681067535008)).

In der Graphentheorie werden zunächst die rein topologischen Eigenschaften der Netzstruktur untersucht. Ein Graph G = (V, E) wird formal durch die Knotenmenge (V) und Kantenmenge (E) definiert. Zwei Knoten u, v ∈ V gelten als benachbart oder als Nachbarn, wenn es eine Kante {u, v} ∈ E gibt, die sie miteinander verbindet. Man unterscheidet grundsätzlich zwischen *gerichteten* (Kanten haben eine Richtung haben (u, v) ≠ (v, u)) und *ungerichteten* (Kanten haben keine Richtung (u, v) ∈ E ⇔ (v, u) ∈ E) Graphen (vgl. Tittmann 2019: 11; vgl. Dörpinghaus 2022).

Die Verbindungen haben immer eine definierte Richtung, die von einem Knoten zu einem anderen verläuft und können ebenfalls keine oder mehrere Eigenschaften besitzen. Dabei hat jede Beziehung immer einen Start- und einen Endknoten (vgl. Barrasa/ Webber 2023: [Introducing Knowlege Graphs](https://learning.oreilly.com/library/view/building-knowledge-graphs/9781098127091/ch01.html#idm45681067535008)). Die Beziehungen sind oft bidirektional, werden aber häufig in einer bestimmten Richtung benötigt (vgl. Dengel 2012: 34).

Die graphische Darstellung eines Graphs dient der Veranschaulichung und kann je nach Zielsetzung variieren. Unterschiedliche Visualisierungen desselben Graphen können dabei verschiedene strukturelle Merkmale sichtbar machen oder in den Vordergrund rücken. Bei der Darstellung gerichteter Graphen wird die Richtung der Relationen zwischen den Knoten üblicherweise durch Pfeile gekennzeichnet (vgl. Dörpinghaus 2022). Diese Struktur dient dazu, Objekte und deren Eigenschaften mit verschiedenen Formen von Relationen zu verbinden, um sie besser zu verstehen (vgl. Dengel 2012: 35), wie es Abb. 1 zeigt.

In Abbildung 1 werden die komplexen Beziehungen zwischen Personen und Orten durch ein einfaches Graphenmodell dargestellt. Die Knoten sind mit Labels wie Person und Ort gekennzeichnet, was eine klare Identifikation der Entitäten ermöglicht und sie erhalten spezifische Eigenschaften wie Name oder Gender. Diese Struktur macht es leicht, die Beziehungen der realen Welt übersichtlich abzubilden (vgl. Barrasa/ Webber 2023: [Introducing Knowlege Graphs](https://learning.oreilly.com/library/view/building-knowledge-graphs/9781098127091/ch01.html#idm45681067535008)).

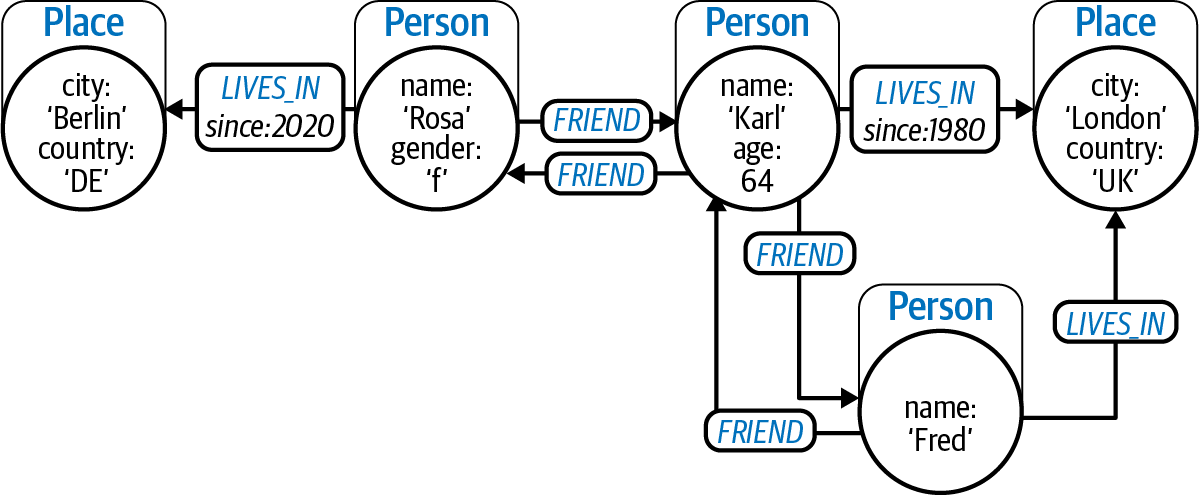


Abbildung 1. Eine Darstellung von Personen, ihren Freundschaften und ihren Standorten in Form eines Graphen (Barrasa/ Webber 2023: [Introducing Knowlege Graphs](https://learning.oreilly.com/library/view/building-knowledge-graphs/9781098127091/ch01.html#idm45681067535008)).

Durch ihre flexible Struktur können Graphen verschiedenartige Datenquellen verknüpfen und unterschiedliche Beziehungstypen und Anwendungsbereiche abzudecken (Heist et al. 2021: 128). Technologien wie RDF (Resource Description Framework), OWL (Web Ontology Language) oder SPARQL (Protocol and RDF Query Language) ermöglichen es, komplexe Sachverhalte durch semantisch angereichte Beziehungen zwischen Daten dazustellen. Aus diesen Daten bildet sich ein Informationsnetz, der durch seine offene Struktur eine vielseitige Verknüpfung erlaubt. Jede Information kann über eine URI (Uniform Resource Identifier) eindeutig angesprochen werden, wodurch sie mit anderen externen, offenen und öffentlich zugänglichen Datenquellen im Web verbunden werden kann. Die semantisch angereichten Daten werden in sog. RDF-Tripeln modelliert. In der Praxis werden solche strukturierte Daten als LOD (Linked Open Data) bezeichnet (vgl. Andraschke/Wagner 2020: 118–119).

### 1.1

### 1.2

## W3C Standards

# Normdaten

Da die Maschinen keine indirekten Bedeutungen erkennen können, daher müssen Relationen zwischen Informationen eindeutig sein. Normdaten wie [GND-Nummer](https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html) (Gemeinsame Normdaten) ermöglichen diese Eindeutigkeit. Wird eine Person über GND-Nummer identifiziert, können Inhalte automatisch mit bspw. [Wikipedia](https://www.wikipedia.de/)-Eintrag oder [Deutschen Biographie](https://www.deutsche-biographie.de/) verknüpft werden. Trotzt der Bemühungen der Deutschen Nationalbibliothek kann die Eindeutigkeit von Normdaten nicht in allen Fällen gewährgeleistet werden. Regelmäßig auftretende Probleme, wie etwa fehlerhafte oder doppelt vorhandene GND-Einträge, lassen sich häufig auf die dezentrale Struktur der Datenerfassung zurückführen, die im Rahmen kooperativer Arbeitsprozesse. In der Praxis wird aus Effizienzgründen oder unter Zeitdruck nicht selten darauf verzichtet, bereits bestehende Datensätze auf ihre Richtigkeit und Vollständigkeit hin zu prüfen. Stattdessen wird ein neuer Eintrag erzeugt, was wiederum zu Redundanzen sowie zu einer Verringerung der Datenkonsistenz und -qualität im Gesamtsystem führen kann. Die Erstellung und Pflege von Normdaten stellt auch die Herausforderung im Bereich von Körperschaften, geographische Orte, Werte, Periodika usw. Ihre eindeutige Identifikation erweist sich oftmals als deutlich komplexer. Insbesondere bei Orten treten spezifische Schwierigkeiten auf: Während georeferenzierende Systeme (GIS) in der Regel lediglich punktuelle Koordinaten erfassen, bleiben flächenhafte Ausdehnungen sowie historische Veränderungen unberücksichtigt. Die historische Dimension eines Ortes – etwa sich verändernde politische Zugehörigkeiten, Ortsnamen oder Grenzverläufe – kann daher nur unzureichend über bestehende textbasierte Normdaten erfasst werden und bedarf ergänzender, oft nicht-textueller Quellen (vgl. Kamzelak 2018: 426-427).

# XML

Eine Grundvoraussetzung für die Erstellung des Semantic Web sind strukturierte und maschinenlesbare Daten. Um diesen Anforderungen gerecht zu werden, wurde XML als Teil der Standardized Generalized Markup Language (SGML) entwickelt (vgl. Jeckle 2002: 11). XML ermöglicht es, eine individuelle Auszeichnungssprache zu erstellen und Elemente sowie Attribute nach Bedarf zu definieren (vgl. Ray 2001: 157). Diese Flexibilität erlaubt es, Daten in einem einheitlichen, maschinenlesbaren Format zu erstellen (vgl. Jeckle 2002: 11).

Obwohl XML als universelle Meta-Sprache gilt, bietet es keine Unterstützung zur Beschreibung der Semantik der Daten. Die Bedeutung der Beziehungen zwischen geschachtelten Elementen bleibt meist implizit und hängt von der jeweiligen Anwendung ab. Dieses Defizit wird durch Einsatz von RDF kompensiert (vgl. Baumeister/ Seipel 2005: 40).

# RDF

RDF ist eine domänenunabhängige Sprache zur semantischen Beschreibung von Datenmodellen (vgl. Baumeister/ Seipel 2005: 40). Damit lassen sich Informationen über *Dinge* (Aussagen) darstellen, die im Web eindeutig identifiziert werden können. Jede Aussage wird in Form eines Tripels dargestellt: Subjekt, Prädikat und Objekt. Diese Tripel bilden zusammen einen Graphen aus Knoten und Kanten (vgl. Dengel 2012: 110).

RDF besteht aus zwei Komponenten: dem RDF-Modell (einer Menge von Aussagen) und der RDF-Syntax zur Serialisierung und damit zum Austausch (vgl. ebd.: 118). Das anwendungsspezifische Vokabular eines RDF-Datenmodells wird im RDF-Schema (RDFs) und in einer Ontologie (OWL) festgelegt. Im Unterschied zu XML-Schema, das die Struktur von XML-Dokumenten durch Einschränkungen definiert, beschreibt das RDFs die Eigenschaften von Objekten sowie deren Beziehungen zueinander und zu ihren Wertebereichen (vgl. Baumeister/ Seipel 2005: 40–41).

# Wikidata

Wikidata ist eine freie und frei zugängliche Datenbank, die strukturierte Daten sammelt. Die Sammlung von Wikidata besteht aus Datenobjekten, die durch ein *Q* und eine Nummer, dem sog. eindeutigen Bezeichner, identifiziert werden (vgl. [Wikidata:Introduction](https://www.wikidata.org/wiki/Wikidata:Introduction/de)). Jedes Datenobjekt enthält eine Aussage (Statement), die die Eigenschaften (property, P) und die dazu gehörigen Werte (values, V) eines Objekts beschreibt. Für Personen können bspw. Information wie Ausbildung hinzugefügt werden (vgl. [Help:Statements](https://www.wikidata.org/wiki/Help:Statements/de)).

Die Datenobjekte im Graph-Format (Subjekt-Prädikat-Objekt) enthalten Informationen, die über sämtliche Wikimedia-Portale erfasst wurden. Dies erfolgt durch die Verknüpfung einer Eigenschaft mit mindestens einem Wert. Die Aussagen können weiterentwickelt, mit Anmerkungen ergänzt oder durch zusätzliche Werte und optionale Qualifikationen erweitert sowie in einen spezifischen Kontext eingeordnet werden. Zudem dienen sie dazu, Objekte miteinander zu verknüpfen. Dies führt zu einer vernetzten Datenstruktur (vgl. [Help:Statements](https://www.wikidata.org/wiki/Help:Statements/de)).

Visualisierung

Netzwerkanalyse biete neue Möglichkeit zur Visualisierung und Analyse von Sammlungswissen. Ihre zentralen Ziele sind die Identifikation wichtiger Verbindungen und Strukturen. (vgl. Werner 2020: 248–249).

Für geistes- und sozialwissenschaftliche Forschung ist die Verbindung zwischen Netzwerken und Visualisierung von großer Bedeutung. Die graphische Visualisierung von abstrakten oder komplexen Strukturen macht die Zusammenhänge, die sich auf rein sprachlichem Wege nur schwer oder unzureichend vermitteln lassen, greifbar. Digitale Werkzeuge bieten hier neue Möglichkeiten, große Datenbestände nicht nur zu verarbeiten, sondern auch visuell aussagekräftige Formen überführen können. Daraus ergibt sich einen Anknüpfungspunk für die Verwendung netzwerkbasierter Modelle. Für die biographische Forschung eröffnet der Netzwerkansatz die Perspektive, individuelle Lebensläufe nicht isoliert zu betrachten, sondern als Teil größer Beziehungsgeflechte (vgl. Dörpinghaus 2022).

Was soll geschrieben werden:

* Was ist NER, wie sie erkannt werden
* Erstellung von Vokabular (einheitlich, orientier an Wikidata) – Normdaten als Hilfe?
* FAIR Prinzipien
* Visualisierung? Sollen unterschiedliche Darstellungen gezeigt werden?
* Sprachmodelle?

Wissensgraphen gewinnen zunehmend an Bedeutung, da sie geeignet sind, komplexe Wissensstrukturen abzubilden (Heist et al. 2021: 128).

Hierbei erweist sich die graphbasierte Netzwerkanalyse als ein zentrales Instrument zur systematischen Verknüpfung von Datenobjekten und deren Kontextualisierung (vgl. Werner 2020: 249–250).

Diese methodischen Entwicklungen knüpfen an grundlegende Konzepte der Graphentheorie.

Analyse

* Herausforderungen bei der Normdatenverknüpfung von Personen/ Körperschaften/ Orten in autobiografischen Quellen: doppelte und fehlerhafte Daten
* Falsche Relationen ()
* Kontrolle
* Regeln für maschinell verständliche Aussagen: Kamzelak 2018: 431-434
  + Prädikate im **Präsens**.
  + Prädikate im **Aktiv**