

XML und RDF im Informationsmanagement

Markus Stocker

18. Juni 2018

Rekapitulation

- Was ist eine Ontologie (in der Informatik)
- Was sind die wesentlichen Bestandteile einer Ontologie?
- Wie ist der Aufbau einer Ontologie (TBox, ABox)
- Was ist bei der Entwicklung von Ontologien besonders schwierig?
- Welche Wissensquellen gibt es?
- Was ist Protégé?

Übersicht

- Informationsmanagement, -infrastruktur, -system
- Was ist Information?
- XML und RDF: Rollen im Informationsmanagement
- Information in Forschungsinfrastrukturen
- Und zum Schluss ein Projektvorschlag

Informationsmanagement

- Steht für das Verwalten von Information
- Planen, Gestalten, Überwachen und Steuern von Information
- Versorgung des Unternehmens mit Information
- Insb. notwendige Information zur Erreichung strategischer Ziele
- Information als “dritten Produktionsfaktor” (nach Kapital und Arbeit)
- Information als unternehmerische Ressource
- Entwickelt und nutzt eine entsprechende Informationsinfrastruktur

<https://de.wikipedia.org/wiki/Informationsmanagement>

Informationsinfrastruktur

- Ist eine “Ermöglichungsstruktur”
- Für die Erzeugung, Verarbeitung und Verwendung von Information
- Informationsinfrastruktur bietet Basisleistungen
- Dient nicht unmittelbar der Produktion
- Infrastruktur ist “unsichtbar”, “verborgen”
- Man bemerkt Infrastruktur nur dann, wenn diese ausfällt
- Telefon-, Breitband-, Daten-, Funknetze bilden die Grundlage
- Eine Gesamtheit an interdependenten Informationssysteme

<https://de.wikipedia.org/wiki/Informationsinfrastruktur>

Informationssystem

- Ein soziotechnisches System (Mensch/Aufgabe/Technik-System)
- Deckung der Informationsnachfrage durch Informationsproduktion
- Mensch ist der Anwender der Aufgaben erfüllen möchte
- Entwickler ebenfalls dem Strukturelement Mensch zugeordnet
- Aufgabe ist das Problem, das mit dem System gelöst werden soll
- Technik besteht aus der Soft- und Hardware des Systems
- Erfüllung Verarbeitung-, Verteilungs- und Speicherungsprozessen
- Informationssysteme sind Ressourcen einer Informationsinfrastruktur

<https://de.wikipedia.org/wiki/Informationssystem>

Was ist Information?

DIKW Modell

- Daten: Symbole/Zeichen, keine Bedeutung
- Information: Ist nutzbar, bedeutungsvoll
- Wissen: Organisierte Information

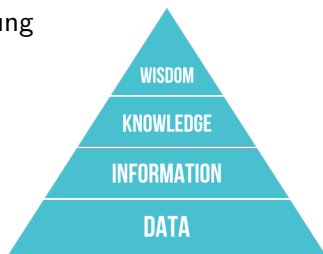


Image by Longlivetheux - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=37705247>

https://en.wikipedia.org/wiki/DIKW_pyramid

Von Daten zu Wissen: Prozesse

A. Aamodt, M. Nygård / *Data & Knowledge Engineering* 16 (1995) 191–222

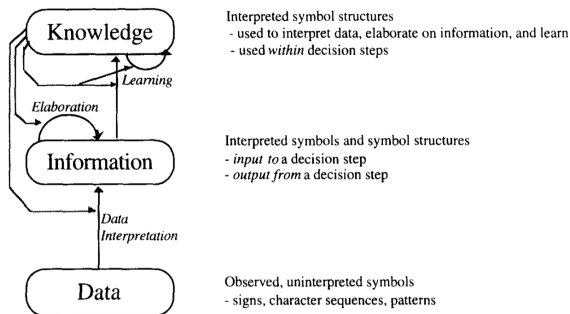


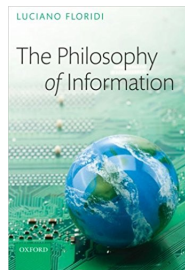
Fig. 2. The Data-Information-Knowledge model.

Aamodt and Nygård (1995). Different roles and mutual dependencies of data, information, and knowledge - An AI perspective on their integration. *Data & Knowledge Engineering*, 16(3):191-222.
[https://doi.org/10.1016/0169-023X\(95\)00017-M](https://doi.org/10.1016/0169-023X(95)00017-M)

$$E = mc^2$$

Floridi über Information

- Luciano Floridi, Prof. Oxford Internet Institute
- Philosophie der Information und Informationsethik
- Autor von "The Philosophy of Information" (2013)



<https://global.oup.com/academic/product/the-philosophy-of-information-9780199232383>

Floridi über Information

- Information besteht aus Daten
- Daten müssen wohlgeformt sein
- Wohlgeformte Daten müssen Bedeutungsvoll sein
- Wohlgeformte und Bedeutungsvolle Daten müssen wahrhaftig sein

Floridi über Information

- Wohlgeformte und Bedeutungsvolle Daten sind semantischer Inhalt
- Nur wahrer semantischer Inhalt ist semantische Information
- Falschinformation und Desinformation keine semantische Information

Floridi über Information

- Semantische Information unterscheidet sich von Umweltinformation
- Umweltinformation sind Signale der Umgebung (“natürliche Daten”)
- Z.B. Jahresringe, Fingerabdrücke, flüchtige organische Verbindungen
- Umweltinformation muss nicht zwingend Bedeutung enthalten
- Auch Pflanzen/Tiere können diese für praktische Zwecke nutzen

Information im Informationsmanagement

- Informationsmanagement verwaltet Information
- Möglichst semantische Information
- Sprich wohlgeformte, bedeutungsvolle und wahrhaftige Daten
- Aber auch nur semantische Inhalte
- Es wird sicherlich auch Falschinformation verwaltet
- Erreichung strategischer Ziele benötigt semantische Information

Information im Informationsmanagement

- Information wird von Informationssystemen verwaltet
- Die interdependenten Systeme einer Informationsinfrastruktur
- Insbesondere die Technik, die Soft- und Hardware des Systems
- Zentral für Informationsverarbeitung, -verteilung, -speicherung

Information im Informationsmanagement

- Die Software eines Informationssystems spielt eine wichtige Rolle
- Diese repräsentiert, verarbeitet, verteilt, speichert Information
- Als wohlgeformte und bedeutungsvolle Daten

Wohlgeformte Daten: Die Rolle von XML

- XML ist eine *meta* Auszeichnungssprache
- Eignet sich für die Entwicklung spezifischer Auszeichnungssprachen
- Zur Markierung von Inhalten
- Wobei sich dadurch strukturierte Daten ergeben
- Die entsprechend den XML Sprachregeln *wohlgeformt* sind

Wohlgeformte Daten: Die Rolle von XML

Unstrukturiert

Der Planet Erde hat einen Radius von 6371 km.

Semistrukturiert

Der `<planet>`Planet `<name>`Erde`</name>` hat
`<property>`einen Radius von `<radius>`6371`</radius>`
`<unit>`km`</unit>``</property>``</planet>`.

Strukturiert

```
<planet>
  <name>Erde</name>
  <property>
    <radius>6371</radius>
    <unit>km</unit>
  </property>
</planet>
```

Wohlgeformte Daten: Die Rolle von XML

- Mit DTD / XML Schema kann man zusätzlich auf Gültigkeit prüfen
- Entsprechen XML Daten einem Schema, insb. Struktur, Datentypen
- In einem Informationssystem unterstützt XML somit die
 - ▶ Repräsentation
 - ▶ Verarbeitung
 - ▶ Kommunikation
 - ▶ Speicherung
- von Information, als wohlgeformte Daten

Wohlgeformte Daten: Die Rolle von RDF

- RDF Daten sind wohlgeformt entsprechend einer RDF Syntax
- Informationssystem prüft Wohlgeformtheit beim Lesen und Schreiben
- Wie XML unterstützt auch RDF die Verwaltung von Information
- Insbesondere Information als wohlgeformte Daten

Bedeutungsvolle Daten: Die Rolle von XML

- XML Daten sind “selbstbeschreibend”
- Die Tags sind in der Tat meist Bedeutungsvoll
- Allerdings nur für Menschen eines Informationssystems
- Der Technik bedeutet nur die Baumstruktur etwas
- Und die Datentypen, da diese Interpretation von Werten einschränken
- Baumstruktur kann traversiert und navigiert werden (XPath)

Bedeutungsvolle Daten: Die Rolle von RDF

- Die
 - ▶ Repräsentation
 - ▶ Verarbeitung
 - ▶ Kommunikation
 - ▶ Speicherung
- von Information als wohlgeformte *und bedeutungsvolle* Daten
- wird von RDF und Ontologiesprachen effektiver unterstützt
- Ontologien formalisieren die Bedeutung verwendeter Terminologie
- Gegeben sind nicht nur Symbole, z.B. `ex:Planet`, `ex:12bv`
- Auch deren Interpretation, "Klasse der Dinge mit Eigenschaften ..."
- Wobei die spezifizierte Interpretation maschinenverarbeitbar ist

Interdependente Systeme: Die Rolle von XML und RDF

- Die Systeme einer Informationsinfrastruktur sind interdependent
- Sie produzieren, konsumieren, kommunizieren Information
- Interdependenz beruht auf Interoperabilität
- Die Fähigkeit eines Systems mit anderen zusammenzuarbeiten
- Interoperabilität wird erhöht wenn sich Systeme einigen
 - ▶ Auf eine gemeinsame Syntax der ausgetauschten Daten
 - ▶ Damit Daten auf wohlgeformtheit geprüft werden können
 - ▶ Wie auch auf eine gemeinsame Semantik der Symbole
 - ▶ Damit die Bedeutung der Symbole in der Infrastruktur gemeinsam ist
- XML und RDF unterstützen die Interoperabilität
- Konsum und Verarbeitung von Information wird vereinfacht
- Information als wohlgeformte und bedeutungsvolle Daten

XML und RDF im Informationsmanagement: Fazit

Diese Technologien unterstützen die Verwaltung von Information als wohlgeformte und bedeutungsvolle Daten (semantische Inhalte) innerhalb Informationsinfrastrukturen.

Und zum Schluss noch dies ...

Information in Forschungsinfrastrukturen

- Informationsmanagement zentral auch in Forschungsinfrastrukturen
- Forschungsinfrastrukturen sind
 - ▶ Einrichtungen, Ressourcen und Dienstleistungen
 - ▶ Die speziell für wissenschaftliche Zwecke errichtet werden
 - ▶ Und Forschung und Lehre ermöglichen oder erleichtern
- Beispiele
 - ▶ Laboratorien (z.B. CERN)
 - ▶ Wissenschaftliche Grossgeräte (z.B. Large Hadron Collider)
 - ▶ Archive, Bibliotheken, Datenbanken, Sammlungen
 - ▶ Informationstechnische Einrichtungen

<https://de.wikipedia.org/wiki/Forschungsinfrastruktur>

Information in Forschungsinfrastrukturen

- Interpretation von Daten ist eine fundamentale Aktivität
- Hauptsächlich von Wissenschaftler durchgeführt
- Primär mittels der Technik von Informationssystemen
- Insbesondere in der Datenanalyse mit entsprechender Software
- Daten werden dabei meist in tabellarische Datenstrukturen verwaltet
- Diese sind somit wohlgeformt

Information in Forschungsinfrastrukturen: Problem

- Bedeutung wird in solchen Datenstrukturen meist nicht verwaltet
- Die Daten sind somit wohlgeformt aber nicht bedeutungsvoll
- Technik eines Informationssystems verwaltet somit nicht Information
- Sondern nur wohlgeformte Daten

Projektvorschlag

- Verbesserte Repräsentation der Bedeutung von Daten
- Wissenschaftliche Daten die in der Datenanalyse erzeugt werden
- Zum Beispiel numerische Zeitreihen, statistische Kennzahlen, ...
- Konkreter
 - ▶ Entwicklung einer (bidirektionalen) Übersetzung von RDF Daten
 - ▶ In eine Python DataFrame Datenstruktur
 - ▶ Ohne Verlust an (formaler) Bedeutung der Daten
- Zum Beispiel als Bachelor- oder Masterarbeit
- Wer dazu Lust hat oder mehr wissen möchte darf sich gerne melden!

Beispiel: RDF Daten nach ...

```
[] rdf:type sosa:Observation ;  
   sosa:observedProperty :temperature ;  
   sosa:hasFeatureOfInterest :air ;  
   sosa:madeBySensor :thermometer ;  
   sosa:resultTime "2018-01-01T00:00:10Z"^^xsd:dateTime ;  
   sosa:hasSimpleResult "3.1 degC"^^cdt:ucum .
```

```
[] rdf:type sosa:Observation ;  
   sosa:observedProperty :temperature ;  
   sosa:hasFeatureOfInterest :air ;  
   sosa:madeBySensor :thermometer ;  
   sosa:resultTime "2018-01-01T00:00:20Z"^^xsd:dateTime ;  
   sosa:hasSimpleResult "2.7 degC"^^cdt:ucum .
```

Beispiel: ... Python DataFrame und wieder zurück

Time	Result
2018-01-01T00:00:10Z	3.1
2018-01-01T00:00:20Z	2.7

Zusammenfassung

