Forschungsdatenmanagement: die Web Science Perspektive

Impulsvortrag

Dr. Elena Demidova

Workshop „Zukunft des Forschungsdatenmanagements für Ingenieurinnen und Ingenieure“

08.03.2018

Technische Universität Darmstadt

Web Science @ L3S

„Preserving, understanding and shaping the Web”

Informatik- und interdisziplinäre Forschung zu Internet und Web

- Internet: Wie sieht das Netz von morgen aus?
- Information: Wie bekomme ich die Informationen, die ich brauche?
- Gemeinschaft: Wie nutzen Gruppen das Web?
- Gesellschaft: Welche Anforderungen stellen wir an das Web?

Ausgewählte Projekte

ALEXANDRIA

25 Jahre Web - Archivierung, Suche, Analyse (ERC)

Data4UrbanMobility

Datenbasierte Mobilitätdienstleistungen für die Stadt der Zukunft

ForgetIT

ForgetIT: Archivierung und Vergessen

QualiMaster

Echtzeit-Datenverarbeitung für Finanzmärkte

SoBigData

EU-Forschungsinfrastruktur für Big Data und soziale Studien

MAPPING

Datenschutz, Copyright und Internet Governance

Datenquellen für die Web Science Forschung

Webinhalte

- Webseiten
  - ALEXANDRIA: 30 TB, über 4 Milliarden URL Snapshots im .de Domain
- Metadaten (Web Markup)
  - Web Data Commons (66 Terabyte)

Daten aus sozialen Medien

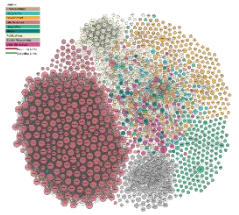
- Twitterdaten
  - @L3S: Seit 06/2017: 2.5 M Deutsche Tweets

Web-Aktivitäten

- Suchmaschinenanfragen

Semantische Daten

- Linked Open Data
- Wissensgraphen
- Ontologien



http://lod-cloud.net/

Domänenspezifische Daten

- Wetterdaten
- Fahrplan Auskünfte
- FCD

Data4UrbanMobility:

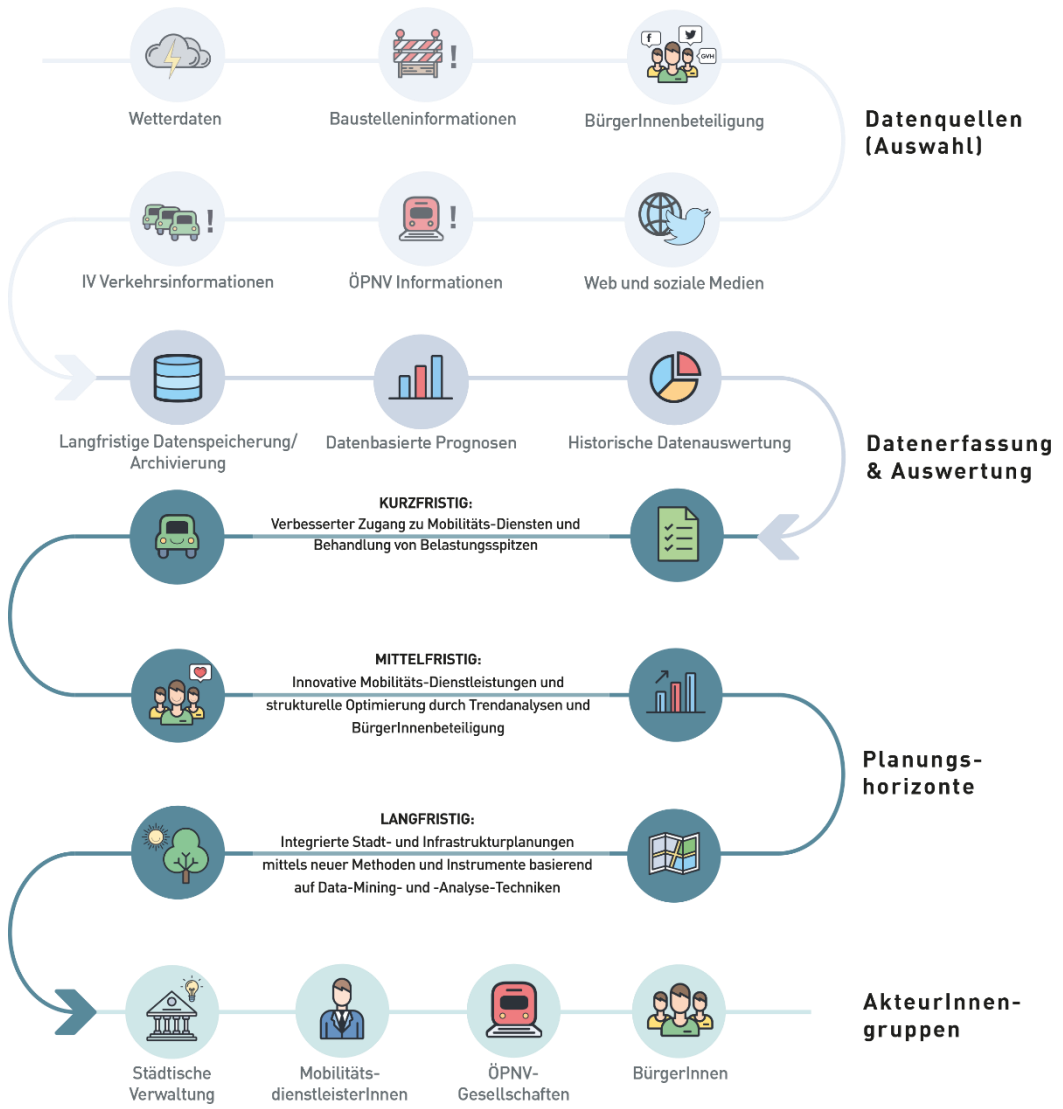
Datenbasierte Mobilitätsdienstleistungen für die Stadt der Zukunft

Ziele

- Effiziente Infrastrukturplanungen, innovative Mobilitäts-Dienstleistungen und strukturelle Optimierung, verbesserter Zugang zu Mobilitäts-Diensten
- Berücksichtigung vielfältiger Einflussfaktoren auf urbane Mobilität (Wetter, Veranstaltungen, Trends wie E-Mobilität, etc.)

Datenbasierter Ansatz

- Ermöglicht, diese vielfältigen Faktoren zu untersuchen
- Heterogene Datenquellen: ÖPNV, IV, Web, soziale Medien, BürgerInnenbeteiligung



GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Webdaten und soziale Medien

Das Web: Das durch Suchmaschinen indizierte Web wird auf > 4,2 Milliarden Webseiten geschätzt (März 2018, www.worldwidewebsize.com)

The Internet Archive (IA)

- Eine digitale Bibliothek von Webseiten in der USA
- 20+ Jahre der Webgeschichte seit 1996
- 279 Milliarden Seiten (Snapshots)
- Weitere Inhalte (digitalisierte Bücher, Texte, Audio, Video, Bilder, Software)
- Eine Kopie der Inhalte: 30+ Petabytes

https://archive.org/about/

Twittersammlung bei der Library of Congress

- 2010 - 2017 Library of Congress: alle Tweets
- 2013: 170 Milliarden Tweets, halbe Milliarde pro Tag (nur Text).
- 2017: Sammlung eingestellt

Library Of Congress Will No Longer Archive Every Tweet

December 26, 2017 · 5:56 PM ET

https://www.npr.org/sections/thetwo-way/2017/12/26/573609499/library-of-congress-will-no-longer-archive-every-tweet

Herausforderungen für Forschungsdatenmanagement im Web Science

- **Interdisziplinarität***
  - Informatiker
  - Digital Humanities-Forscher
  - Juristen
  - Architekten
  - …

***Anforderungen an die Daten [1]**

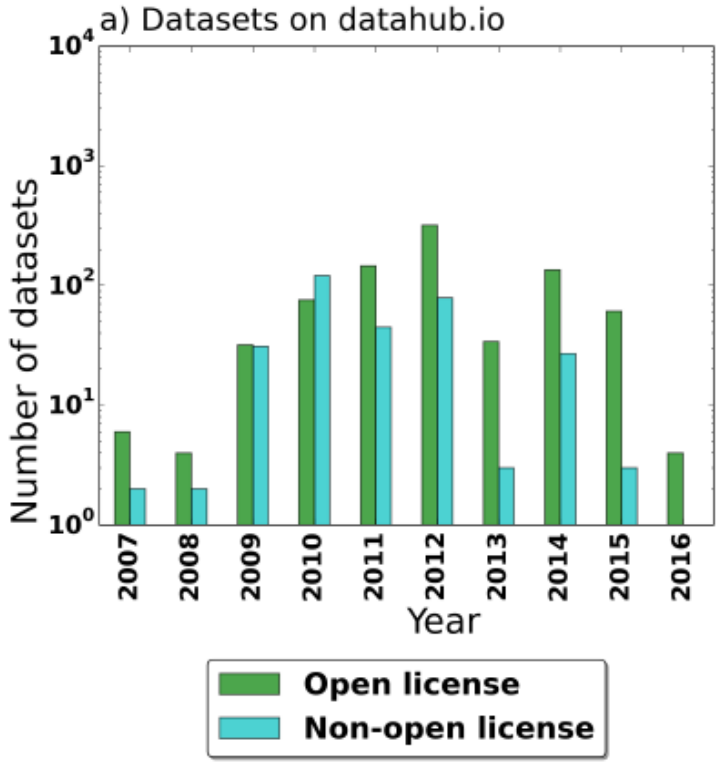
  - Herkunft
  - Authentizität
  - Kontext

- **Große Datenmengen**
  - 10-1000 TB
- **Heterogenität**
  - Formate
  - Quellen
  - ..
- **Multilingualität**
  - Z.B. Wikipedia in 295 Sprachversionen
- **Rechtliche Aspekte**

Datenschutzaspekte / Lizenzen / Nutzungsbedingungen

- Personenbezogenen Daten
  - Einwilligung für bestimmte Forschungsfragen
  - Probleme bei der Nachnutzung
- Kommerzielle Daten
- Vertrauliche Daten
- Komplexe Nutzerbedingungen
- Urheberechte
- …

Wiederverwendung von Forschungsdaten in wiss. Publikationen [2]



Lösungsansätze @L3S: SoBigData

- **SoBigData**
  - Eine EU-Forschungsinfrastruktur für Big Data. 7 EU-Länder, > 100 Forscher
- **Virtuelle Forschungsumgebung**
  - Web-basierte Umgebung mit Applikationen für Datenzugang, Diensten und Algorithmen.
- **Transnationaler Zugang**
  - Forschungsaufenthalte mit dem Zugang zu nicht-öffentlichen Forschungsdaten
  - Explorative Projekte zu festgelegten Themen
  - Open Call

www.sobigdata.eu

Lösungsansätze @L3S: ForgetIT

- Verwaltung von Archivierung und Vergessen
  - Auswahl der Ressourcen für Archivierung
    - Z.B. basierend auf Popularität und Signifikanz
  - Optionen für den Archivierungsgrad
    - Z.B. volle Archivierung, Redundanzminimierung und digitales Vergessen
- Synergetische Archivierung
- Archivierungsprozesse als Part vom Lebenszyklus der Informationsmanagement
- Kontextualisiertes Erinnern
- Archivierung von Informationen im Kontext

www.forgetit-project.eu

Lösungsansätze: Datenprofile für (RDF) Daten, Vokabulare und Werkzeuge

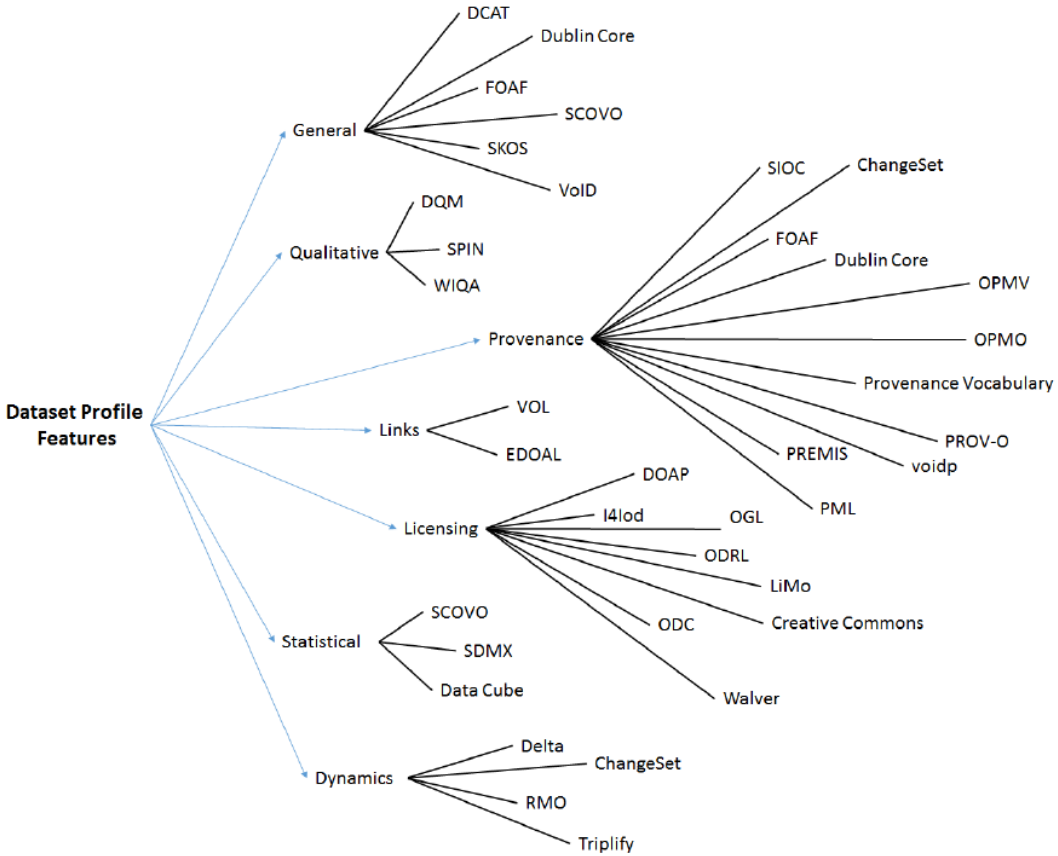
Das World Wide Web Consortium (W3C) ist das Gremium zur Standardisierung der Techniken im World Wide Web seit 1994.

Umfangreiche Vokabulare zur Datenbeschreibung

z.B.

- Dublin Core
- VoID
- DCAT
- …

Viele Methoden und Werkzeuge zur automatischen Extraktion von Datenbeschreibungen (Datenprofilen) [3]



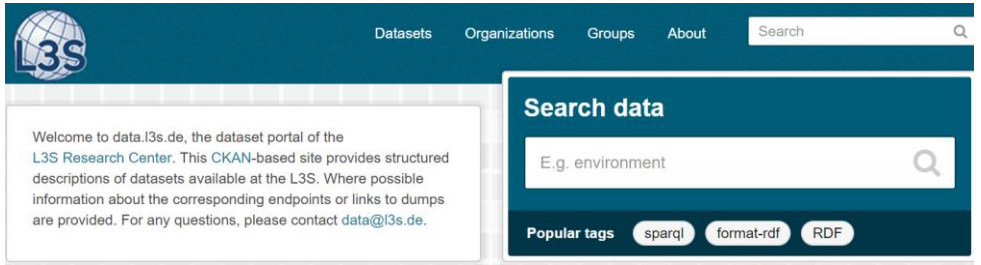
Lösungsansätze: Plattformen für Datenveröffentlichungen

- Webseiten der Forscher / Institutionen

[16] The datasets can be found at http://markup.l3s.de.
[17] http://www.dbpedia-spotlight.org/faq

- Dienste bei den Institutionen
  - z.B. CKAN (http://data.l3s.de/)

  webbasierte Datenkatalog-Software



- Webplattformen
  - z.B. Zenodo https://zenodo.org/

Infrastruktur für Forschungsdatenmanagement der Zukunft

- **Einfachere Veröffentlichung von Daten**
  - (Semi-)Automatische Erstellung / Extraktion von semantischen Datenbeschreibungen
  - Vereinfachung des rechtlichen Rahmens / Lizenzen

- **Einfachere Wiederverwendung von Daten**
  - Verfügbarkeit von offenen Daten für Forschungszwecke
  - Semantische Beschreibungen: Aussagekräftige Datenprofile
  - Semantische Datensuche [4]: Institutionenübergreifend
  - Kontext: Datenveröffentlichung + Publikationen + Software

PROFILES
data:search

- **Rückmeldung / Datenzitationen**
  - Dokumentation der Nachnutzung, Transparenz

Dr. Elena Demidova

Forschungszentrum L3S
Appelstr. 9a
30167 Hannover

Phone: +49 511 762 17776
E-mail: demidova@L3S.de
Web: http://demidova.wordpress.com

References

[1] Risse, T., Demidova, E., Gossen, G.: What do you want to collect from the web? In: Proceedings of the Building Web Observatories Workshop, BWOW 2014

[2] K. M. Endris, J. M. Giménez-García, H. Thakkar, E. Demidova, A. Zimmermann, C. Lange, E. Simperl. Dataset Reuse: An Analysis of References in Community Discussions, Publications and Data. In Proceedings of the Knowledge Capture Conference, K-CAP 2017.

[3] M. Ben Ellefi, Z. Bellahsene, J. Breslin, E. Demidova, S. Dietze, J. Szymanski, K. Todorov. (2017) RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications. Semantic Web Journal. IOS Press.

[4] PROFILES & DATA: SEARCH – International Workshop on Profiling & Searching Data on the Web. https://profiles-datasearch.github.io/2018/