



# Open Source - Ein Schlaraffenland für KI/NLP-Anwendungen

Fachtagung – Automatische Verfahren in der  
Erschließung

Deutsche Nationalbibliothek

18. Nov. 2021

Stefan Geißler

[www.kairntech.com](http://www.kairntech.com)

# Überblick

---

- Vorstellung Kairntech
- Open Source
- Beispiele: Kairntech bei der automatischen Erschließung
  - Thesaurus-basiertes Indexieren
  - Pflege von Vokabularen/Thesauri
  - Metadatengenerierung
  - (Relationenerkennung)
- Ausblick

- Software + Dienstleitung zu KI + NLP (natürlicher Sprachverarbeitung)
- Französisch/Deutsch, Hauptsitz Grenoble
- Viele Jahre Erfahrung aus NLP-Kontexten bei Xerox, IBM, TEMIS
- Fokus: KI + NLP Anwendungen für Unternehmen + Organisationen.
- Anwendungen für *Domänenexperten*, nicht nur für Data Scientists (keine Programmierung nötig)



# Open source?

---

## Ein wenig Geschichte

- Bis 70er Jahre: Computer Hardware generiert große Einnahmen Cash-Cow, Software oft nur die Dreingabe
- Ab 80er: Aufstieg von Microsoft und z.B. SAP zu Weltkonzernen mit proprietärer Software
- Ab 2000er: Open source (Richard Stallman bevorzugt den Term „Freie Software“)
  - Nutzung
  - Zugang
  - Weiterentwicklung/Anpassung/(Korrektur)
  - Weitergabesollen ohne Einschränkung erlaubt sein mit der die Vorgabe, dass Software, die aus freier Software abgeleitet wurde, ihrerseits wieder frei sein muss (und der impliziten Einladung, dass alle etwas beitragen können/sollen)

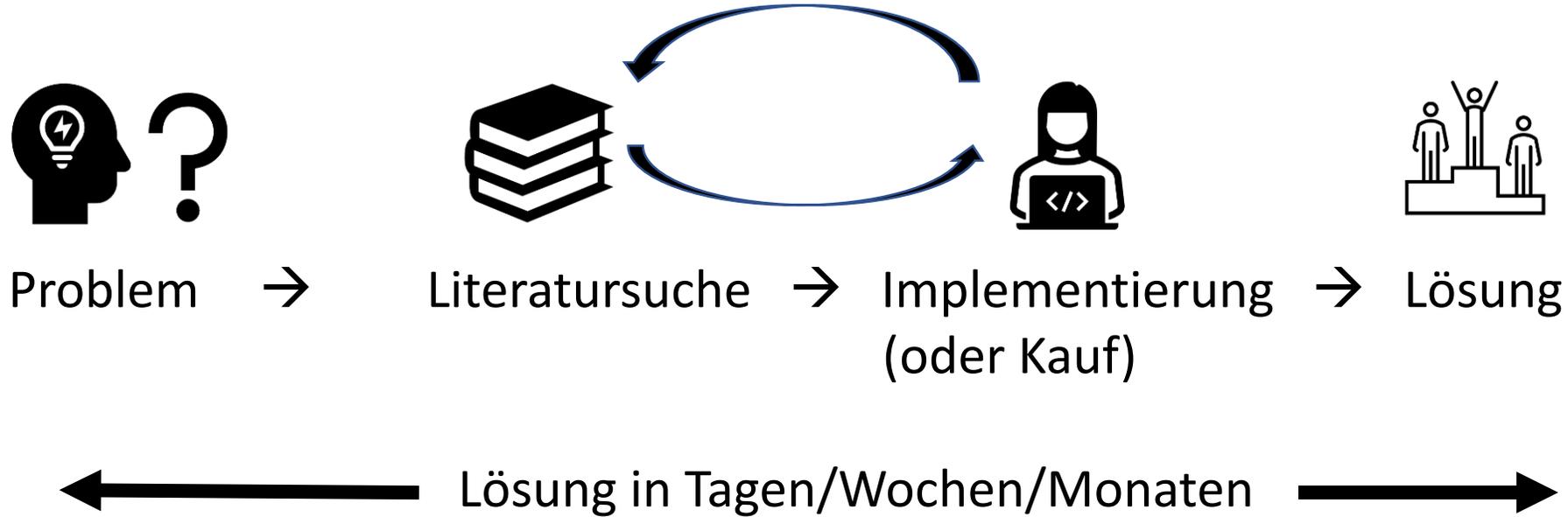
# Open source

---

- Geschäftsmodelle?
  - Code ist frei zugänglich
  - Anpassungen / Wartung /Support etc kostenpflichtig
- Große Unternehmen sind längst eingestiegen
  - Microsoft kauft GitHub (Plattform zum Verwalten und Teilen von open source Projekten)
  - IBM kauft für >30 Mrd\$ RedHat (erfolgreiche Distribution von Linux)
- Android ist open source (>2,5 Mrd Smartphones weltweit)
- Ein Tesla läuft heute mit >2,5mio Zeilen open source Code (mehr als ein modernes Kampfflugzeug)

# Open source in NLP

- Früher (< 2010)



- Heute



„Lösung“ manchmal innerhalb eines Tages [www.kairntech.com](http://www.kairntech.com) – 18.11.2021

# Open source in NLP

---

- Früher (<2010)
  - Allgemeiner Thesaurus (wie WordNet für englisch)? Im deutschen “GermaNet”, proprietär, hohe Lizenzkosten
  - Parser? Stanford, Brill, Marcus, ... proprietär
- Heute: frei verfügbar, gut dokumentiert, lebendige Community, ...
  - NLP-Basispakete: Spacy, NLTK, Tensorflow, ...
  - Wortschätze: Wikidata ...
  - Modelle: BERT, ...
- Open Source in unserem Team: (Patrice Lopez):
  - DELFT (Deep Learning Bibliothek): <https://github.com/kermitt2/delft>
  - Entity Fishing (Thesaurus-basiertes Indexieren): <https://github.com/kermitt2/entity-fishing>
  - Grobid (Metadatengenerierung): <https://github.com/kermitt2/grobid>

# Anwendungsbeispiele

---

- Thesaurus-basiertes Indexieren
  - u.a. “Entity Fishing”
- Pflege von Vokabularen/Thesauri
  - u.a. “Delft”
- Metadatengenerierung
  - u.a. “Grobid”
- (Relationenerkennung)

# Szenario: Thesaurus-basiertes Indexieren

- Grundlage: Wikidata ([www.wikidata.org](http://www.wikidata.org))
- > 90 mio Konzepte zu nahezu allen Themengebieten
- Viele fach-spezifische Thesauri sind Teil von Wikidata
  - MeSH, NCBI, Drugbank, Geonames, ...
- Taxonomie: Terme sind Typen zugeordnet
  - Leopard → Panthera → Felidae → Carnivora → Mammal ...
- Bei Kairntech:
  - Desambiguierung
  - Scoring
  - Linking zu Hintergrundinformation, Bild, Definition
  - Fortlaufend aktualisiert
  - Als fertige Annotationskomponente mit REST-API

**GLYPHOSATE** (IUPAC name: N-(phosphonomethyl) **GLYCINE**) is a broad-spectrum **SYSTEMIC HERBICIDE** and **CROP DESICCANT**. It is an **ORGANOPHOSPHORUS COMPOUND**, specifically a **PHOSPHONATE**, which acts by inhibiting the plant **ENZYME 5-enolpyruvylshikimate-3-PHOSPHATE synthase**. It is used to kill weeds, especially annual broadleaf weeds and **GRASSES** that compete with crops. It was discovered to be an **HERBICIDE** by **MONSANTO CHEMIST JOHN E. FRANZ** in **1970**. **MONSANTO** brought it to market for agricultural use in **1974** under the trade name Roundup. **MONSANTO'S** last commercially relevant **UNITED STATES PATENT EXPIRED** in **2000**. Farmers quickly adopted **GLYPHOSATE** for agricultural **WEED CONTROL**, especially after **MONSANTO** introduced **GLYPHOSATE-resistant ROUNDUP READY CROPS**, enabling farmers to kill weeds without killing their crops. In **2007**, **GLYPHOSATE** was the most used **HERBICIDE** in the **UNITED STATES** agricultural sector and the second-most used (after **2,4-D**) in home and garden, government and industry, and commercial applications. **[3]** From the late **[1970S TO 2016]**, there was a **100-fold** increase in the frequency and volume of application of **GLYPHOSATE-based HERBICIDES (GBS)** worldwide, with further increases expected in the future, partly in response to the global emergence and spread of **GLYPHOSATE-resistant weeds**, **[4]** requiring greater application to maintain effectiveness. The development of **GLYPHOSATE** resistance in **WEED** species is emerging as a costly problem. **GLYPHOSATE** is absorbed through foliage, and minimally through roots, and transported to growing points. It inhibits a plant **ENZYME** involved in the synthesis of three **AROMATIC AMINO ACIDS**: **TYROSINE**, **TRYPTOPHAN**, and **PHENYLALANINE**. It is therefore effective only on actively growing plants and is not effective as a **PRE-EMERGENCE HERBICIDE**. An increasing number of crops have been **GENETICALLY ENGINEERED** to be tolerant of **GLYPHOSATE** (e.g. **ROUNDUP READY SOYBEAN**), the first **ROUNDUP READY** crop, also created by **MONSANTO**), which allows farmers to use **GLYPHOSATE** as a **POST-EMERGENCE HERBICIDE** against weeds. While **GLYPHOSATE** and formulations such as Roundup have been approved by regulatory bodies worldwide, concerns about their effects on humans and the environment persist, and have grown as the global usage of **GLYPHOSATE** increases. **[4]** **[5]** A number of regulatory and scholarly reviews have evaluated the relative **TOXICITY** of **GLYPHOSATE** as an **HERBICIDE**. The German **FEDERAL INSTITUTE FOR RISK ASSESSMENT TOXICOLOGY** review in **2013** found that the available data is contradictory and far from being convincing with regard to correlations between exposure to **GLYPHOSATE** formulations and risk of various cancers, including **NON-HODGKIN LYMPHOMA (NHL)**. **[6]** A **META-ANALYSIS** published in **2014** identified an increased risk of **NHL** in workers exposed to **GLYPHOSATE** formulations. **[7]**

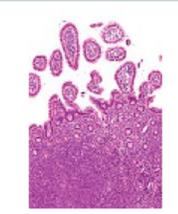
**WAYNE DOUGLAS GRETZKY** (/ˈɡrɛtski/; born **JANUARY 26, 1961**) is a **CANADIAN** former **PROFESSIONAL ICE HOCKEY** player and former **HEAD COACH**. He played **20** seasons in the **NATIONAL HOCKEY LEAGUE (NHL)** for four teams **FROM 1978 TO 1999**. Nicknamed **THE GREAT ONE**, **[1]** he has been called the greatest hockey player ever by many sportswriters, players, and the league itself. **[2]** **GRETZKY** is the **LEADING SCORER** in **NHL HISTORY**, with more goals and assists than any other player. **[3]** He garnered more assists than any other player scored **TOTAL POINTS**, and is the only **NHL** player to total over **200** points in one season – a feat he accomplished four times. In addition, **GRETZKY** tallied over **100** points in **16** professional seasons, **14** of them consecutive. At the time of his retirement in **1999** and persisting through **2017**, he holds **61** **NHL** records: **40** **REGULAR SEASON** records, **16** playoff records, and six **All-STAR RECORDS**. **[2]** **[3]** Born and raised in **BRANTFORD, ONTARIO, CANADA**, **GRETZKY** honed his skills at a **BACKYARD RINK** and regularly played **MINOR HOCKEY** at a level far above his peers. **[4]** Despite his unimpressive stature, strength and speed, **GRETZKY'S** intelligence and reading of the game were unrivaled. He was adept at dodging checks from opposing players, and consistently anticipated where the **PUCK** was going to be and executed the right move at the right time. **GRETZKY** became known for setting up behind his opponent's net, an area that was nicknamed **GRETZKY'S OFFICE**. **[5]** In **1978**, **GRETZKY** signed with the **MINNAPOLIS RANGERS** of the **WORLD HOCKEY ASSOCIATION (WHA)**, where he briefly played before being traded to the **EDMONTON OILERS**. When the **WHA** folded, the **OILERS** **JOINED THE NHL**, where he established many scoring records and led his team to four **STANLEY CUP CHAMPIONSHIPS**. **GRETZKY** is trade to the **(LOS ANGELES KINGS)** on **AUGUST 5, 1988**, had an immediate impact on the team's performance, eventually leading them to the **1993 STANLEY CUP FINALS**, and he is credited with popularizing hockey in **CALIFORNIA**. **[6]** **GRETZKY** played briefly for the **ST. LOUIS BLUES** before finishing his career with the **NEW YORK RANGERS**. **GRETZKY** captured nine **HART TROPHIES** as the most valuable player, **40** **ART ROSS** Trophies for most points in a season, two **CONN SMYTHE TROPHIES** as playoff **MVP** and five **LESTER PEARSON** Awards (now called the **TED LINDSAY AWARD**) for most **OUTSTANDING PLAYER** as judged by his peers. He won the **LADY BYNG MEMORIAL TROPHY** for sportsmanship and performance five times, and often spoke out against **FIGHTING IN HOCKEY**. **[7]**

**NHL**

Normalized: Non-Hodgkin lymphoma

Domains: Animal\_Husbandry, Medicine

conf: 0.8555



Non-Hodgkin lymphoma (NHL) is a group of **blood cancers** that includes all types of **lymphoma** except **Hodgkin's lymphoma**. Symptoms include **enlarged lymph nodes**, fever, **night sweats**, weight loss, and tiredness. Other symptoms may include bone pain, chest pain, or itchiness. Some forms are slow growing while others are fast growing.

Wikidata statementa

References:  

**NHL**

Normalized: National Hockey League

Domains: Administration

conf: 0.71

The **National Hockey League (NHL)** is a professional **ice hockey league** currently comprising 31 teams: 24 in the United States and 7 in Canada. Headquartered in **New York City**, the NHL is considered to be the premier professional ice hockey league in the world, and one of the **major professional sports leagues in the United States and Canada**. The **Stanley Cup**, the oldest professional sports trophy in North America, is awarded annually to the league playoff champion at the end of each **season**.

Wikidata statementa

References:  

# Szenario: Kairntech und Thesauruspflege

- Thesauri entwickeln sich parallel zum jeweiligen Thema weiter
- Manuelle Pflege von großen Thesauri jedoch kann große Aufwände erfordern
- Automatische Unterstützung:
  - System lernt zugrundeliegende Konzepte
  - Schlägt neue Kandidaterme nach automatischer Analyse neuer Inhalte vor
  - Manuelle Entscheidung erforderlich
- Zeitersparnis, Ausweitung der betrachteten Quellen
- Fallbeispiel: Kairntech im Einsatz im “Technologie-Monitoring” bei TecIntelli & Univ Stuttgart auf sich rasch verändernden Technologiefeldern
- Cf. <https://kairntech.com/finding-new-needles-in-content-haystacks-vocabulary-maintenance-with-ai/>

The image displays three screenshots of the Kairntech Sherpa software interface, illustrating its functionality in managing a thesaurus.

**Top Screenshot: Lexicons**  
This view shows a grid of term categories: Redox (11 terms), Solidstate (2 terms), Sufforbased (12 terms), Airbased (6 terms), and Energy density (3 terms). An 'Import term in test lexicon' dialog box is open, offering file formats: csv, excel, and text.

**Middle Screenshot: Advanced Features**  
This view shows a sidebar with options: ANNOTATIONS (Remove all annotations), SUGGESTIONS (Remove all suggestions), MODELS (Remove all models), INDEX (Reindex project), and LABEL CORPUS (Generate labels for documents in corpus). A 'Generate labels for documents in corpus' dialog box is also visible.

**Bottom Screenshot: Test**  
This view shows a document snippet with an abstract. The text contains several terms: 'redox flow battery' (highlighted in red), 'redox flow battery' (highlighted in red), and 'MV12 flow battery' (highlighted in red). Below the text, two boxes identify extracted terms: 'Known term extracted' (pointing to 'redox flow battery') and 'New term candidate extracted' (pointing to 'MV12 flow battery'). A code editor shows the JSON representation of the extracted terms.

```
..[
  identifier: "redox flow battery",
  label: "redox",
  preference: "redox flow battery"
]
```

# Szenario: Metadatengenerierung

- Erzeugung von Metadaten aus PDF
  - Author?
  - Titel?
  - Datum?
  - Affiliations?
  - Abstract?
  - Literatur? (Dereferenzierung mit Crossref)
- Maschinelle Lernsoftware: PDF→TEI XML
  - Full open source: <https://github.com/kermitt2/grobid>
  - Derzeitige Modelle optimiert für Dokumententyp „wissenschaftlicher Aufsatz“
  - Weitere Modelle möglich (Erstellung von Trainingskorpora, Training, Evaluierung, ...)

## REFERENCES AND NOTES

- [1] Bresciani, S.; Tomkinson, N. C. O. *Heterocycles* 2014, 89, 2479.
- [2] Galenko, A. V.; Khlebnikov, A. F.; Novikov, M. S.; Pakalnis, V. V.; Rostovskii, N. V. *Russ Chem Rev* 2015, 84, 335.
- [3] Bansal, S.; Halve, A. K. *Int J Pharm Sci Res* 2014, 5, 4601.
- [4] Kumar, K. A.; Govindaraju, M.; Renuka, N.; Kumar, G. V. *J Chem Pharm Res* 2015, 7, 250.
- [5] Meyer, A. G.; Ryan, J. H. *Molecules* 2016, 21, 935.
- [6] Berchet, M.; Chevret, T.; Dujardin, G.; Parrot, I.; Martinez, J. *Chem Rev* 2016, 116, 15235.
- [7] Feamley, S. P. *Curr Org Chem* 2004, 8, 1289.
- [8] Beccalli, E. M.; Pocar, D.; Zoni, C. *Targets Heterocycl Systems* 2003, 7, 31.
- [9] Deivasalhar, S.; Meite, P.; Choudh, K. *Synlett* 2017, 28, 521.

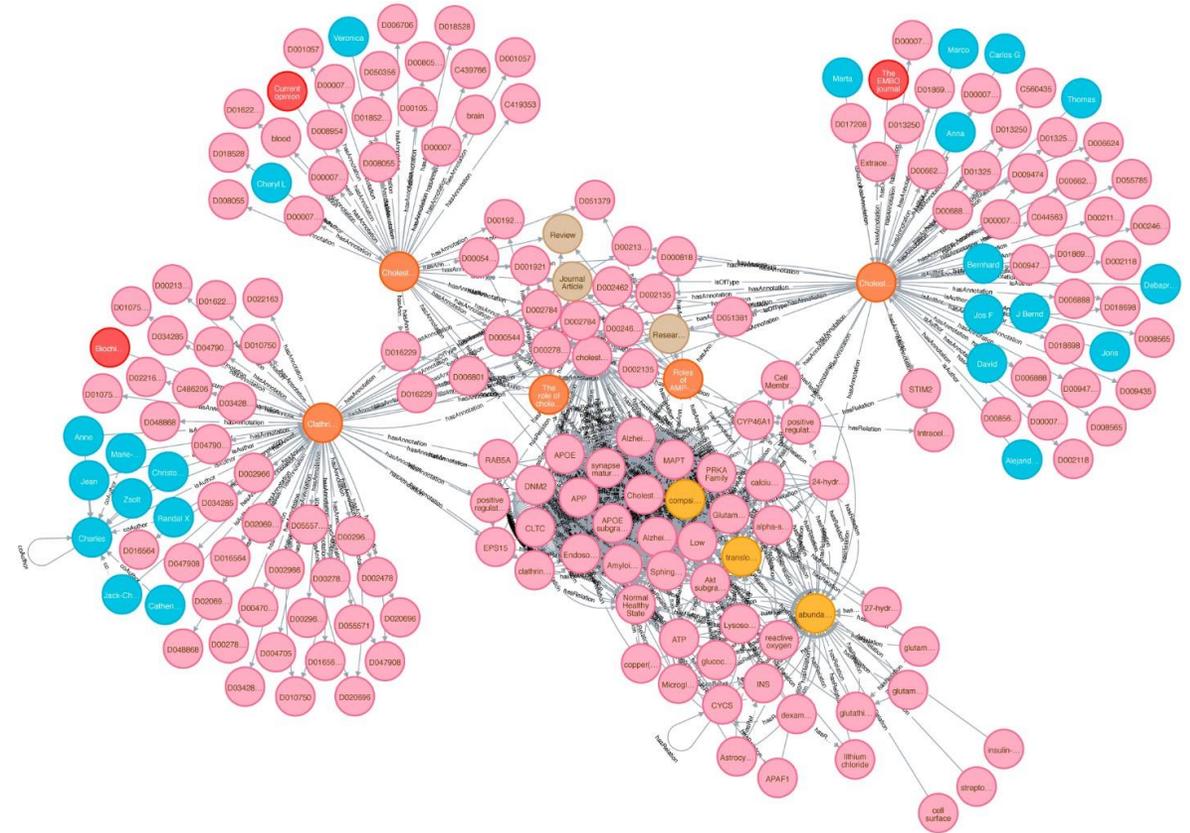
```
<analytic>
  <monogr>
    <title level="a" type="main">1,3-Dipolar Cycloaddition
    Reactions of Azomethine Ylides with Carbonyl
    Dipolarophiles Yielding Oxazolidine
    Derivatives</title>
    <author>
      <persName
        xmlns="http://www.tei-c.org/ns/1.0"><forename
        type="first">Adam</forename><forename
        type="middle">G</forename><surname>Meyer</surname>
      </persName>
      </author>
      <author>
        <persName
          xmlns="http://www.tei-c.org/ns/1.0"><forename
          type="first">John</forename><forename
          type="middle">H</forename><surname>Ryan</surname>
        </persName>
      </author>
      <idno type="doi">10.3390/molecules21080935</idno>
    </analytic>
  </monogr>
  <title level="j">Molecules</title>
  <title level="j" type="abbrev">Molecules</title>
  <idno type="ISSNe">1420-3049</idno>
  <imprint>
    <biblScope unit="volume">21</biblScope>
    <biblScope unit="issue">8</biblScope>
    <biblScope unit="page">935</biblScope>
    <date type="published" when="2016" />
    <publisher>MDPI AG</publisher>
  </imprint>
</monogr>
</biblStruct>
```

[www.kairntech.com](http://www.kairntech.com) – 18.11.2021



# Szenario: Relationenerkennung für Fraunhofer SCAI

- Erkennung von Relationen zwischen domänen-spezifischen Entitäten (Proteinen, Genen, chem. Substanzen, ...)
- Kodierung in maschinenlesbares Format BEL (“biological expression language”) → Anreicherung von großem Wissensgraphen → Inferenzen über den gespeicherten Zusammenhängen
- SCAI&Kairntech planen Ausweitung der Zusammenarbeit
- Cf. <https://www.scai.fraunhofer.de/en/press-releases/news-26-10-2021.html>



# Szenario: Relationenerkennung für Fraunhofer SCAI

## Das Projekt:

- Trainingsdaten: In BEL kodierte Relation aus der Literatur + jeweilige Stelle im Dokument

p(HGNC:"METRNL") decreases path(MESH:"Parkinson Disease")

- Kairntech erzeugt 90000+ corpus über “Schizophrenie & Bipolare Störungen”
- Kairntech wendet vorliegende Entitätenextraktion an (Gene, Proteine, Krankheiten, chem. Substanzen, biol. Prozesse, ...)
- Erzeugung von Modell zur Relationsextraktion
- Manuelle Evaluation von 200 Samples durch SCAI Expert:innen, 71 von 74 Beispielen (=95%) mit hoher Konfidenz sind “korrekt”, 32 von 40 (=80%) mit niedriger Konfidenz.

# Open source: Mehr als nur ein Geschäftsmodell...

---

- IT-Infrastruktur ist Teil der Daseinsvorsorge (wie Gesundheit, Straßen, Feuerwehr)
- Sicherheit von öffentlicher Infrastruktur
- Digitale Souveränität: Nutzer:innen (→ Menschen) erhalten Autonomie über die Technik, die sie verwenden
- “Public money → public code”
- Zum Weiterlesen: Open Knowledge Foundation ([www.okfn.de](http://www.okfn.de))

- Open source ist heute Grundlage für raschen Austausch und Fortschritt (nicht nur in) der NLP-Gemeinde
- Für Anwender: Open source vermindert die Gefahr des “Lock in” (Erzwungene fortgesetzte Bindung) an bestimmten Anbieter
- Für Anbieter: Open source bietet spannende Geschäftsmodelle
- Fragen? Anregungen? Eigene Anforderungen? [info@kairntech.com](mailto:info@kairntech.com), [www.Kairntech.com](http://www.Kairntech.com)
- Danke für Ihre Aufmerksamkeit