Open Source - Ein Schlaraffenland für KI/NLP-Anwendungen

Fachtagung – Automatische Verfahren in der Erschließung
Deutsche Nationalbibliothek
18. Nov. 2021

Stefan Geißler
www.kairntech.com

Überblick

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

Kairntech

- 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)
  - Weitergabe

  sollen 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

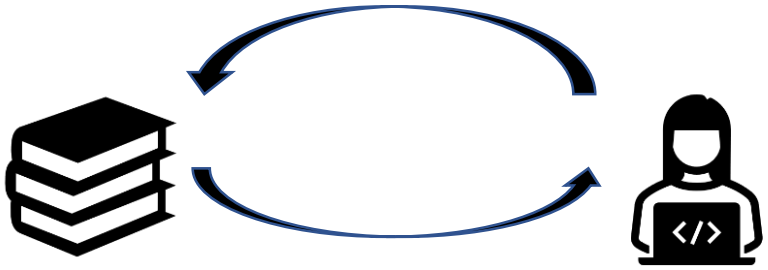
- Geschäftsmodelle?
  - Code ist frei zugänglich
  - Anpassungen / Wartung /Support etc kostenpflichtig

- Große Unternehmen sind längst eingestiegen
  - Microsoft kauft GitHub (Pattform 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)

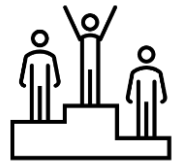




Problem → Literatursuche → Implementierung (oder Kauf) → Lösung

Lösung in Tagen/Wochen/Monaten

- Heute


- Paper in arxiv
- Implementierung +Daten+Modell in Github



„Lösung“ manchmal innerhalb eines Tages

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

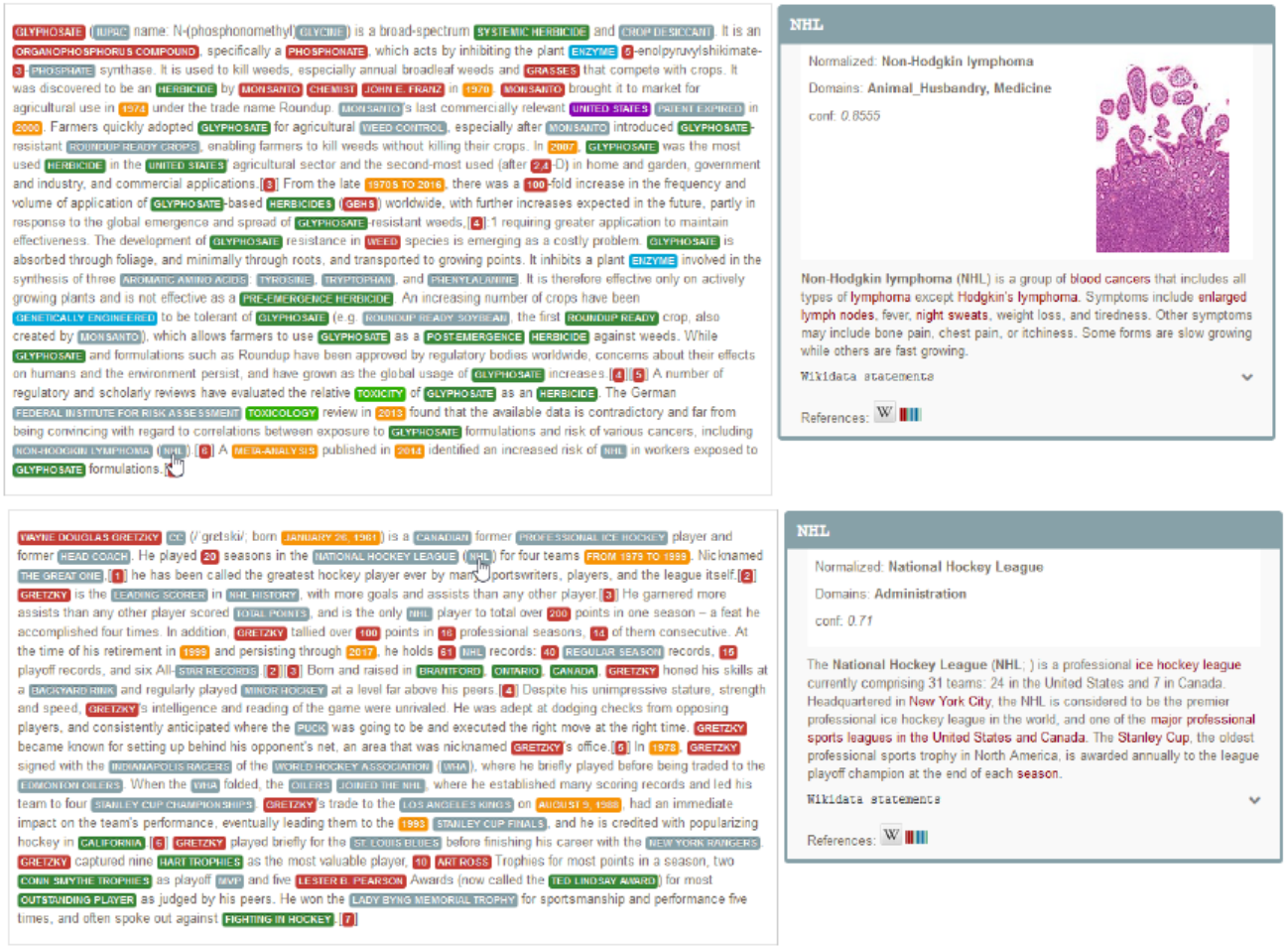
Anwendungsbeispiele

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

Szenario: Thesaurus-basiertes Indexieren

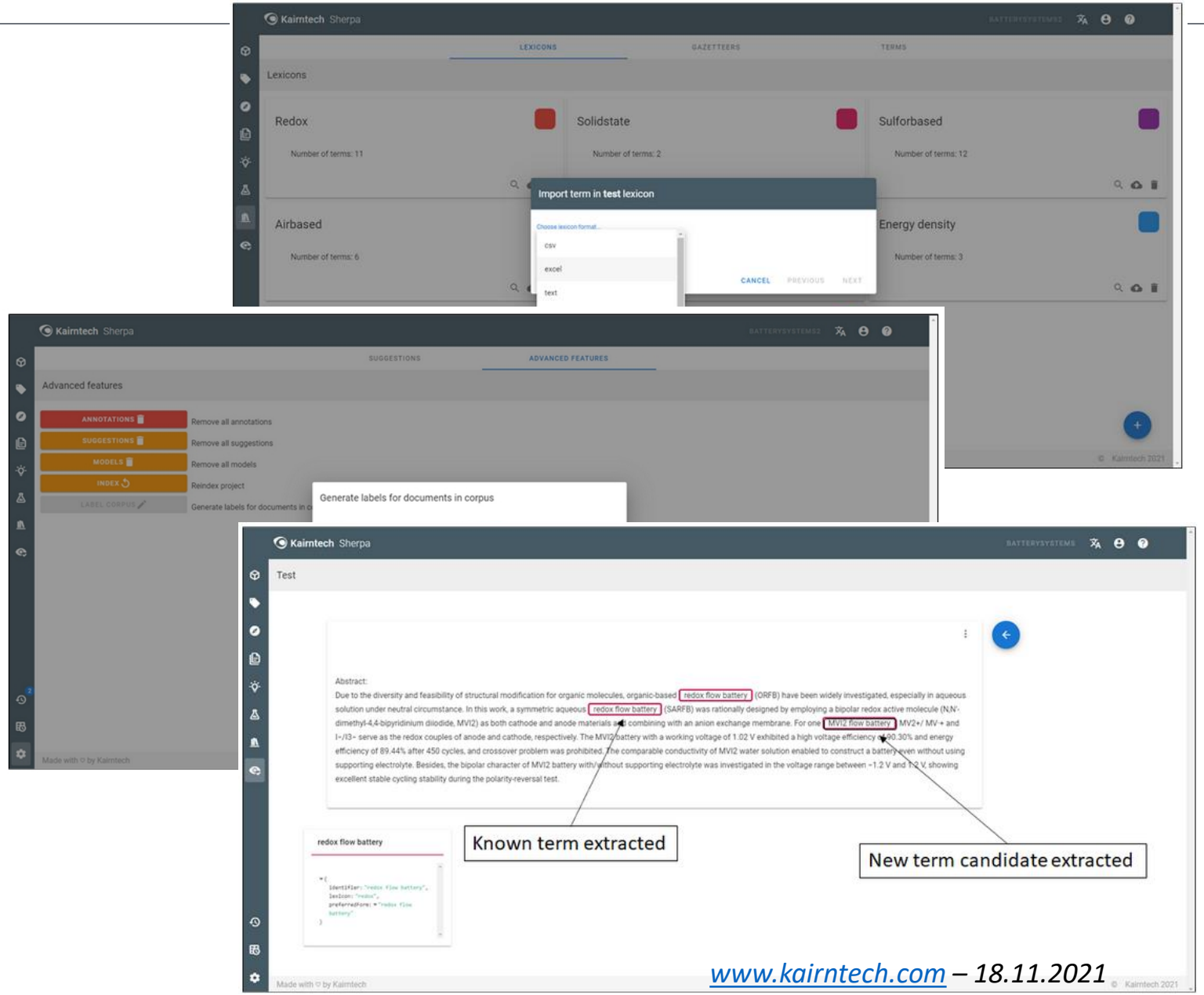
- Grundlage: Wikidata (www.wikidata.org)
- > 90 mio Konzepte zu nahezu allen Themengebieten
- Viele fach-spezifische Thesauri sind Teil von Wikidata
  - MeSH, NCBI, Drugbank, Geonames, …
- Taxonomie: Terme sind Typen zugeordnet
  - Leopard → Panthera → Felidae → Carnivora → Mammal …

- Bei Kairntech:
  - Desambiguierung
  - Scoring
  - Linking zu Hintergrundinformation, Bild, Definition
  - Fortlaufend aktualisiert
  - Als fertige Annotationskomponente mit REST-Api



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 Kandidatenterme 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/

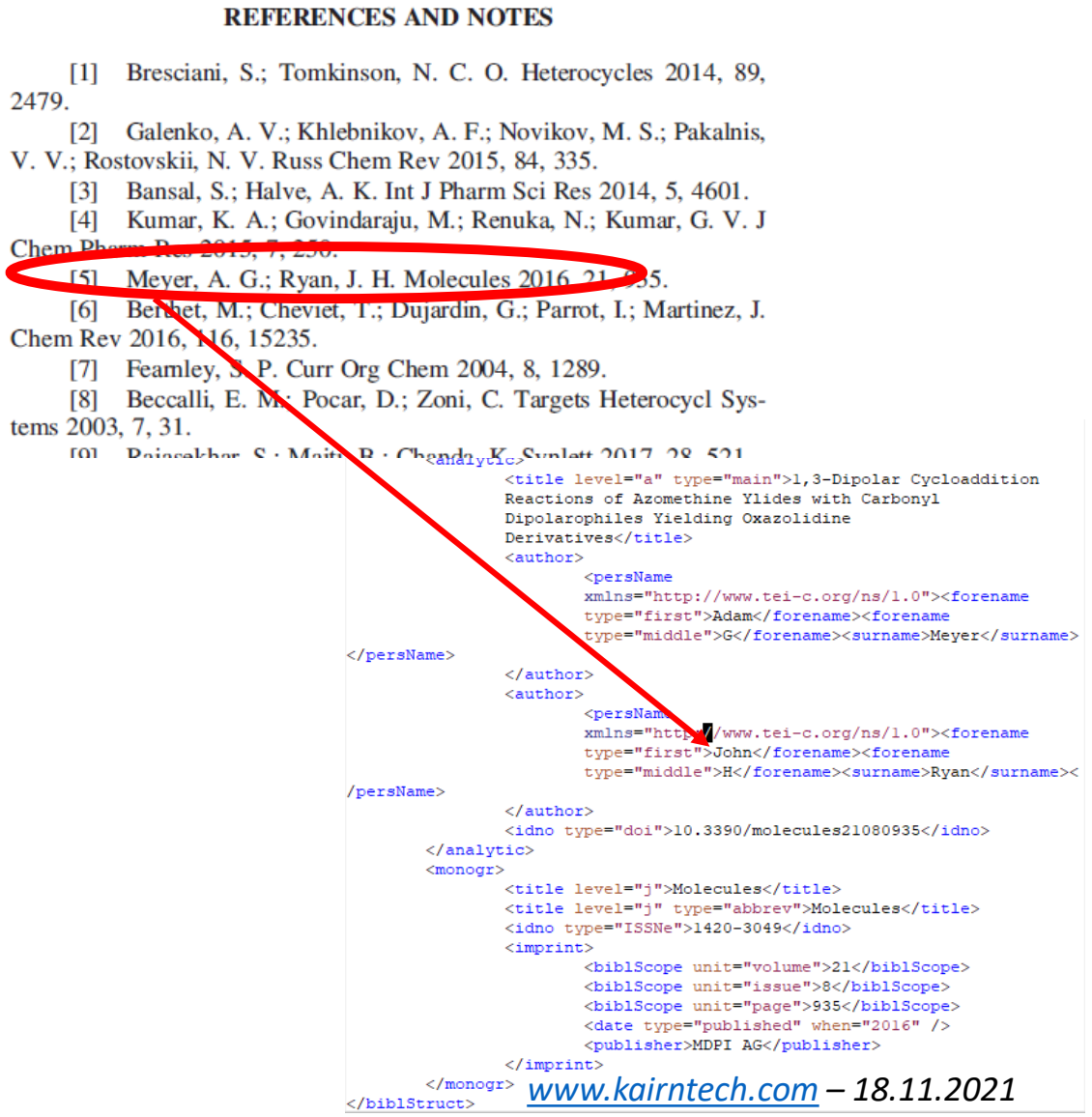


www.kairntech.com – 18.11.2021

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] Berthet, M.; Cheviet, T.; Dujardin, G.; Parrot, I.; Martinez, J. Chem Rev 2016, 116, 15235.

[7] Fearnley, S. P. Curr Org Chem 2004, 8, 1289.

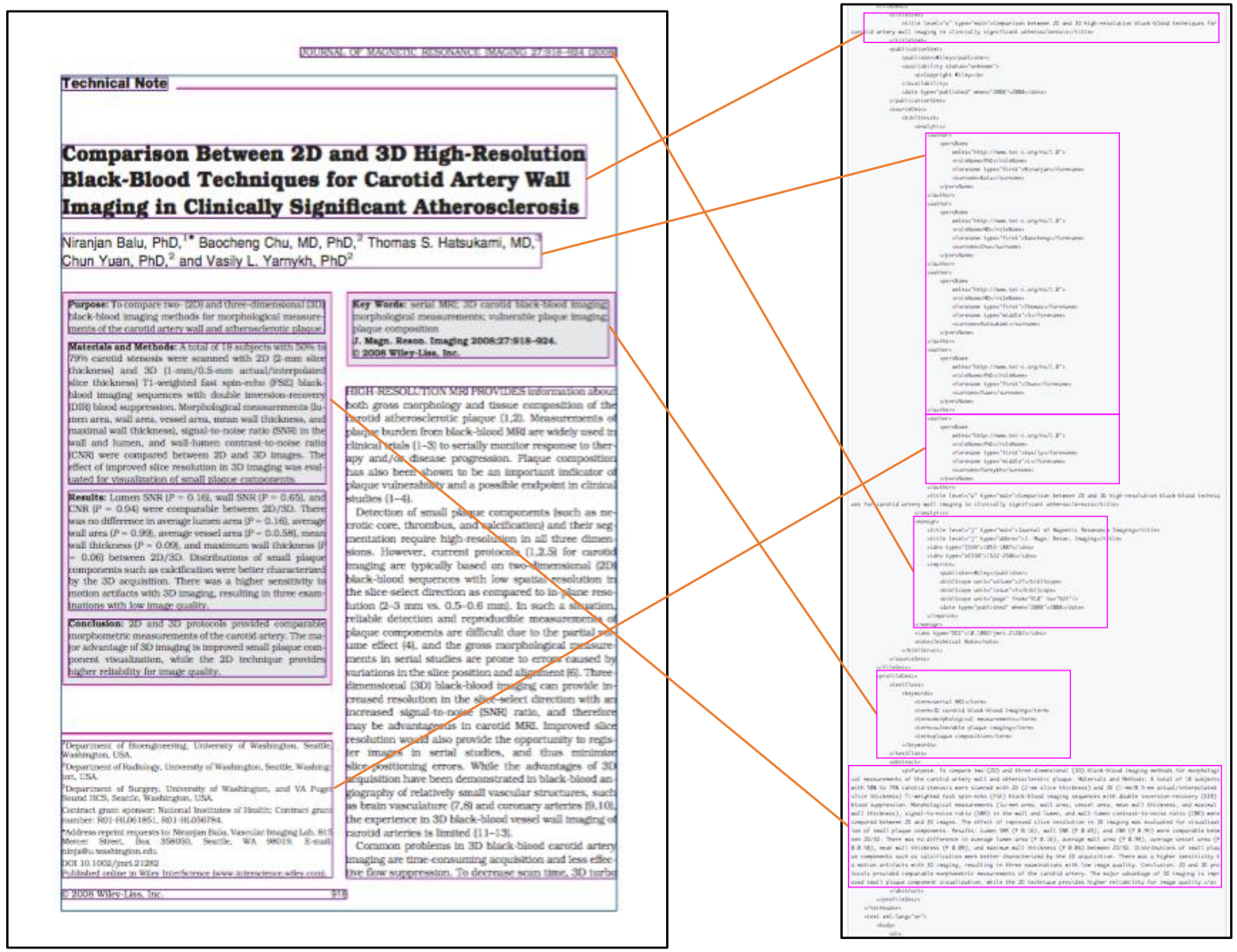
[8] Beccalli, E. M.; Pocar, D.; Zoni, C. Targets Heterocycl Systems 2003, 7, 31.

[9] Rajasekhar, S.; Maity, B.; Chanda, K. Synlett 2017, 28, 521.

```
<analytic>
        <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>
```

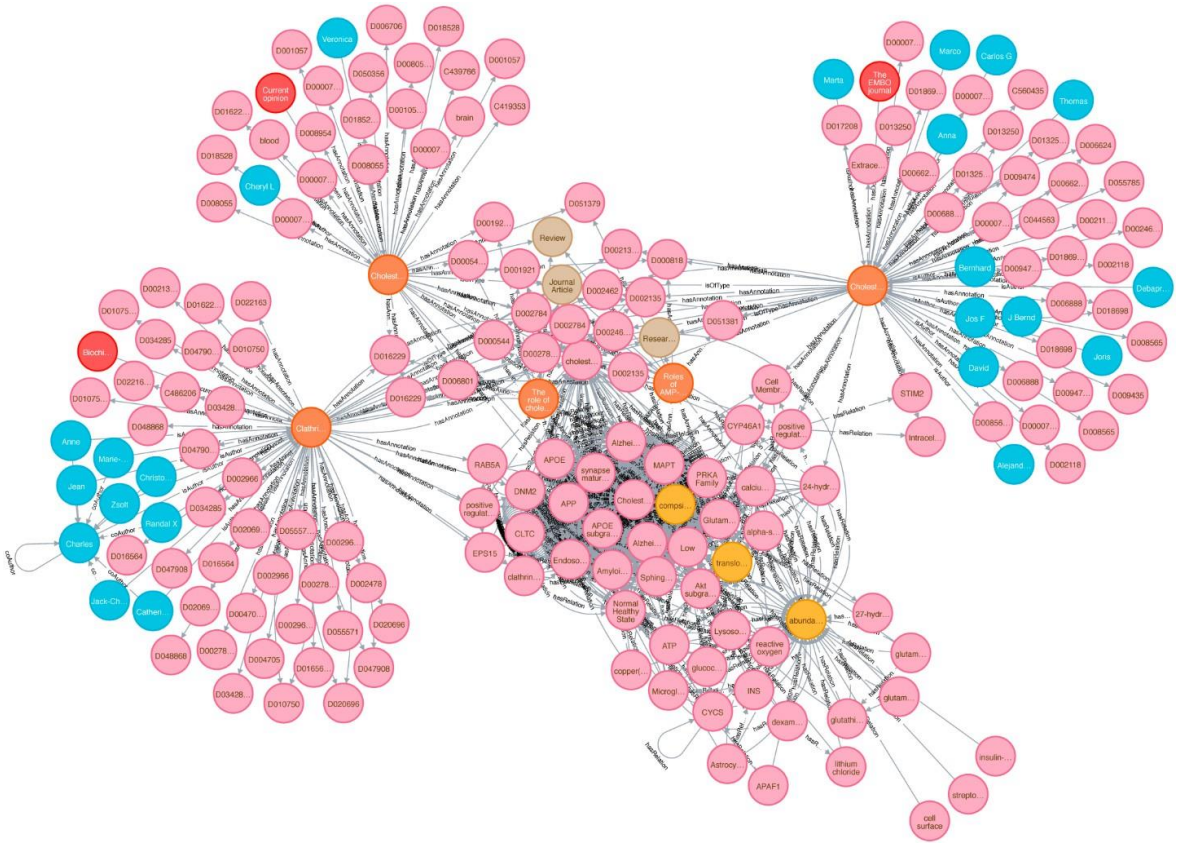
Szenario: Metadatengenerierung

- Dokumentenstrukturierung
- Cascadierte Machine Learning Modelle verwenden unterschiedliche Hinweise (clues) wie Fontgröße, Positionen im Dokument, Fettdruck, lexikalische Muster etc.
- Im Einsatz bei ResearchGate, CERN, Mendely, European Patent Office, CNRS



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 des Daseinsvorsorge (wie Gesundheit, Straßen, Feuerwehr)
- Sicherheit von öffentlicher Infrastuktur
- 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)

Ausblick

- 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, www.Kairntech.com
- Danke für Ihre Aufmerksamkeit