Entwicklung eines cloudbasierten Systems zur Autorsuche anhand von Themen-Schlüsselwörtern unter Verwendung von Vektoreinbettung, Retrieval, LLMs und KI-Agenten

Betreuer: M. Sc. Gladys Kouakam, Dr.-Ing. Philippe Tamla, Prof. Dr. Matthias Hemmje

Forschungsbezug

Habilitationsvorhaben von Dr.-Ing. Philippe Tamla sowie Dissertation von M. Sc. Gladys Kouakam mit dem Arbeitstitel: *Supporting Cloud-based Information Extraction Life Cycle from MICE Content as a Basis for Managing Attendee Engagement and Satisfaction as well as Speaker Performance and Reputation in Professional Management of Meetings, Congresses and Events*

Voraussetzungen

Für eine erfolgreiche Bearbeitung der Aufgabenstellungen sind Vorkenntnisse in folgenden Bereichen notwendig:

- Container
- Information Retrieval
- Information Extraction
- Machine Learning
- Python
- Generative AI
- Natural Language Processing
- Unified Modeling Language
- Large Language Model (LLM)

Präambel

Diese Arbeit ist an die Habilitation von Philippe Tamla und die Dissertation von Gladys Kouakam angelehnt. Die Habilitation untersucht die Entwicklung domänenübergreifender, cloud-basierter Referenzmodelle für KI-gestützte Informationsextraktion und Wissensmanagement. Die Dissertation von Gladys Kouakam überträgt diesen Ansatz auf die Domäne der **Meetings, Incentives, Conferences and Exhibitions (MICE)** und untersucht, wie ein cloud-basierter **Information Extraction Life Cycle (IELC)** zur Verarbeitung veranstaltungsbezogener Daten und Inhalte ausgestaltet werden kann. Ziel ist die Entwicklung domänenspezifischer Referenzmodelle, mit denen unstrukturierte Veranstaltungsdaten systematisch gesammelt, aufbereitet, analysiert, in strukturierte Wissensrepräsentationen überführt und für datenbasierte Entscheidungen im professionellen Veranstaltungsmanagement zugänglich gemacht werden können.

Die MICE-Branche umfasst Meetings, Incentive-Veranstaltungen, Kongresse, Konferenzen und Ausstellungen. Sie spielt eine wichtige Rolle bei der Förderung von Zusammenarbeit, Wissensaustausch und professioneller Vernetzung zwischen Wissenschaft, Wirtschaft, öffentlichen Einrichtungen und Fachgesellschaften. In den letzten Jahren haben verschiedene technologische und gesellschaftliche Entwicklungen die MICE-Branche beeinflusst. Die COVID-19-Pandemie, Klimawandel, digitale Kommunikationstechnologien und der Aufstieg der künstlichen Intelligenz haben die Art und Weise, wie Konferenzen

organisiert und durchgeführt werden, erheblich verändert [1–3]. Insbesondere Fortschritte im Bereich der maschinellen Lernverfahren, die auf Transformer-basierten Modellen basieren, ermöglichen es Systemen, komplexe Textdaten zu analysieren, natürliche Sprache zu verstehen und Entscheidungsprozesse zu unterstützen [4]. Die Organisation von Konferenzen wird in der Regel von einer Vielzahl von Akteuren übernommen, zu denen Veranstalter, Autoren, Gutachter, Referenten und Teilnehmer zählen. Konferenzen werden in mehreren Phasen organisiert, die im Folgenden näher definiert werden. Zunächst werden Abstracts eingereicht [5]. Im Anschluss erfolgt eine Begutachtung [6]. Darauf aufbauend wird das Programm gestaltet. Schließlich wird die Veranstaltung durchgeführt. Während dieser Prozesse fallen große Mengen heterogener Daten an, darunter Einreichungen, Abstracts, Autorenmetadaten und Präsentationsmaterial. Diese Daten liegen in strukturierten, semistrukturierten und unstrukturierten Formaten vor. Hierzu gehören relationale Datenbanken, CSV-, JSON- und XML-Dateien, Programmierschnittstellen, Webseiten, HTML- und PDF-Dokumente, Freitexte sowie audiovisuelle Aufzeichnungen [7,8]. Zusätzlich sind die Daten häufig auf verschiedene Systeme und Veranstaltungseditionen verteilt und weisen unterschiedliche Schemata, Terminologien, Sprachen, Qualitätsniveaus und Zeitbezüge auf. Die generierten Daten, können mithilfe von Techniken der **Information Extraction (IE)** analysiert werden [9] . IE bezeichnet Methoden, die automatisch strukturierte Informationen aus unstrukturiertem Text extrahieren [10]. Techniken der **Natural Language Processing (NLP)** ermöglichen es Computern, Textdaten zu verstehen, zu interpretieren und zu analysieren. Moderne Ansätze stützen sich häufig auf Transformer-Architekturen und Deep-Learning-Modelle, die in der Lage sind, große Mengen an Textdaten zu verarbeiten [4]. Ein Schlüsselkonzept der modernen Informationsgewinnung ist die **Vektoreinbettung**, bei der Textinformationen in hochdimensionale Vektoren umgewandelt werden, die die semantische Bedeutung darstellen [11]. Mithilfe der vektorbasierten Suche können Systeme semantisch ähnliche Dokumente oder Autoren auf der Grundlage von Ähnlichkeitsmaßen statt durch den Abgleich von Schlüsselwörtern identifizieren.

Als konzeptionelle Grundlage für die Verarbeitung der heterogenen Daten dient das von Tamla vorgeschlagene IELC-Referenzmodell [12]. Das Modell beschreibt die schrittweise Transformation heterogener Rohdaten in strukturierte und zugängliche Informationen. Es umfasst die Phasen *Data Collection*, *Data Preparation*, *Information Extraction*, *Information & Knowledge Organization* sowie *Information Access & Retrieval*.

Dieses Modell unterstützt die Extraktion und strukturierte Repräsentation von Autorenwissen aus heterogenen Datenquellen, darunter Konferenzbeiträge, Abstracts und wissenschaftliche Publikationen. Im MICE-Kontext können solche Ansätze Veranstaltungsorganisatoren dabei unterstützen, anhand thematischer Suchanfragen relevante Autorinnen und Autoren zu identifizieren, deren Forschungsschwerpunkte zu einem bestimmten Konferenzthema oder eingereichten Beitrag passen. Hierzu werden Quelldaten wie PDF-Dokumente und bibliografische Metadaten verarbeitet und in semantische Repräsentationen überführt. Diese bilden die Grundlage für eine themenbasierte Autorensuche, beispielsweise zur Auswahl geeigneter Gutachter, Vortragender oder weiterer fachlich passender Autorinnen und Autoren.

Die Übertragbarkeit dieses lebenszyklusorientierten Ansatzes auf unterschiedliche Anwendungsdomänen wird durch verwandte Forschungs- und Entwicklungsvorhaben verdeutlicht. Das Kooperationsvorhaben **Knowledge-Management Ecosystem-Portal for the National Centre of Excellence for Green Ports and Shipping (KMEP4ECOGPS)** [13] untersucht die Integration und Vermittlung von Informationen über nachhaltige maritime Betriebssysteme und Aktivitäten in grünen Häfen und Schifffahrtsnetzwerken. Es basiert auf dem **Knowledge Management System (KMS) Content and Knowledge Management Ecosystem Portal (KM-EP)**, das vom Lehrgebiet Multimedia und Internetanwendungen der FernUniversität in Hagen gemeinsam mit dem FTK e.V. entwickelt wurde [14,15].

Ein weiteres verwandtes Forschungsprojekt ist **Artificial Intelligence for Hospitals, Healthcare & Humanity (AI4H3)** [16]. Es untersucht die Unterstützung transparenter und erklärbarer medizinischer Entscheidungen und baut auf den Ergebnissen von **Recommendation Rationalisation (RecomRatio)** auf [17]. Dabei werden medizinische Fachpublikationen analysiert, um Argumente für oder gegen bestimmte Behandlungsmöglichkeiten zu extrahieren, strukturiert bereitzustellen und für medizinische Entscheidungsprozesse nutzbar zu machen. Als technische Grundlage dient ebenfalls das KM-EP [18]. Die zugehörige AI4H3-Hub-Architektur bildet einen Bezugspunkt für die Habilitation von Philippe Tamla und untersucht wiederverwendbare, cloud-basierte Informationsextraktionsprozesse für unterschiedliche Wissensdomänen.

Eine zentrale empirische Grundlage der Dissertation von Gladys Kouakam bildet die Zusammenarbeit mit der GLOBIT Global Information Technology GmbH [19]. GLOBIT fungiert als Praxis- und Datenpartner und stellt für Forschungszwecke veranstaltungsbezogene Realweltdaten aus dem von GLOBIT betriebenen Legacy-Konferenzmanagementsystem bereit. Die Daten liegen als kongress- beziehungsweise

veranstaltungsspezifische JSON-Dateien vor. Als primäre Datenquelle wird der Datensatz GLOBIT-Kongressprogrammdaten 2015–2025 verwendet. Nach aktuellem Exportstand umfasst er mehr als 17 594 Beitragsdatensätze und 3 404 Sitzungsdatensätze aus 11 Veranstaltungsausgaben beziehungsweise Konferenzdatensätzen. Die Daten enthalten strukturierte Informationen zu eingereichten wissenschaftlichen Beiträgen und zu den Programmen verschiedener wissenschaftlicher und medizinischer Fachkongresse.

Zu den von GLOBIT betreuten Kongressreihen gehören unter anderem der Deutsche Pflegetag (DPT), der Deutsche Kongress für Parkinson und Bewegungsstörungen (DPG beziehungsweise DPGHD), der European Congress of Pathology (ECP beziehungsweise ESP), die European Conference on Schizophrenia Research (ECSR), der World Congress on ADHD, der Kongress der Deutschen Gesellschaft für Psychiatrie und Psychotherapie, Psychosomatik und Nervenheilkunde (DGPPN), der European Congress of Child and Adolescent Psychiatry (ESCAP), der Deutsche Suchtkongress (SUCHT), der World Congress of Biological Psychiatry (WFSBP) und der World Congress of Psychiatry (WPA). Welche dieser Kongressreihen in die konkrete Untersuchung einbezogen werden, wird auf Grundlage der verfügbaren JSON-Dateien, der zeitlichen Abdeckung und der Qualität der Textdaten festgelegt. Für die Untersuchung werden insbesondere folgende Attribute aus den JSON-Dateien verwendet:

- eindeutige beziehungsweise pseudonymisierte Identifikatoren der Sitzungen und Beiträge,
- die Zuordnung zur jeweiligen Kongressreihe und Veranstaltungsausgabe,
- das Veranstaltungsjahr sowie weitere verfügbare zeitliche Metadaten,
- Titel, Beginn und Ende der jeweiligen Sitzung,
- Sessiontyp, Sprache, Raum- und Track-Zuordnung, soweit vorhanden,
- Titel und Abstracttext des jeweiligen Beitrags,
- vorhandene Topic- oder Keyword-Zuordnungen,
- öffentliche Beitrags- beziehungsweise Programmnummern.

Ergänzend können öffentlich zugängliche Veranstaltungsprogramme, Conference Proceedings und bibliografische Metadaten verwendet werden. CrossRef [20] kann zur Ergänzung bibliografischer Angaben über DOIs eingesetzt werden. GROBID [21] dient nicht als eigenständige Datenquelle, sondern als Werkzeug zur Extraktion und Strukturierung von Metadaten und Textinhalten aus PDF-Dokumenten.

Aufbauend auf dem IELC, den beschriebenen Vorarbeiten und der Zusammenarbeit mit GLOBIT untersucht die Dissertation von Gladys Kouakam die Entwicklung domänenspezifischer Referenzmodelle zur Unterstützung von Veranstaltungsorganisatoren bei der datengetriebenen Planung, Durchführung und Auswertung von MICE-Veranstaltungen. Im Rahmen der Dissertation wird hierzu die **Information Extraction Lifecycle for Attendee Satisfaction & Speaker Performance (IELC4AS2P)**-Architektur entwickelt (vgl. Abbildung 1). Sie beschreibt den Verarbeitungspfad von veranstaltungsbezogenen Rohdaten über die Informationsextraktion und Wissensorganisation bis hin zu Informationszugangs- und Entscheidungsunterstützungsdiensten.

Die vorgeschlagene Architektur unterscheidet drei zentrale Kategorien von Eingabedaten:

- *Event Metadata*, beispielsweise Veranstaltungsprogramme, Veranstaltungsausgaben, Sessions, Tracks, Themen, Abstracts sowie Angaben zu Teilnehmenden und Vortragenden,
- *Event Content*, beispielsweise wissenschaftliche Publikationen, Präsentationen, Workshop-Zusammenfassungen, Conference Proceedings, Veranstaltungsberichte und Aufzeichnungen,
- *Event Execution Data*, beispielsweise Echtzeit-Feedback, Umfrageantworten, Fragen und Antworten, Abstimmungen sowie Anwesenheits- und Interaktionsdaten.

Nach ihrer Sammlung und Aufbereitung werden diese Daten mit unterschiedlichen Information-Extraction-Verfahren untersucht. **Named Entity Recognition (NER)** unterstützt die Extraktion eines **Event Vocabulary (EV)**. **Sentiment Analysis (SA)** und quantitative Messverfahren dienen der Erzeugung eines **Attendee Interaction and Impact Vocabulary (AIIV)**. **Topic Modeling (TM)** wird auf zeitlich strukturierte Event-Inhalte angewendet, um ein **Emerging Topic Vocabulary (ETV)** und ein **Trending Topic Vocabulary (TTV)** zu erzeugen. **Predictive Analytics (PA)**-Methoden ergänzen diese Vokabulare um Bewertungen und Metriken zu Teilnehmerengagement, Teilnehmerzufriedenheit, Teilnehmerinteresse, Rednerleistung und Rednerreputation. Ergänzend

werden mithilfe von Verfahren zur Personal-, Referenz- und Evaluations-**IE** Autoreninformationen aus den verfügbaren Veranstaltungs- und Publikationsdaten gewonnen.

In der nachfolgenden Phase der Informations- und Wissensorganisation sollen die extrahierten Vokabulare in hierarchische Strukturen überführt werden. Hierzu gehören insbesondere eine **Event Taxonomy (ET)**, eine**Attendee Interaction and Impact Taxonomy (AIIT)**, eine **Emerging Topic Taxonomy (ETT)**, eine **Trending Topic Taxonomy (TTT)** sowie eine **Expert Ontology (EO)** beziehungsweise ein **Expert Knowledge Graph (EKG)**. Die daraus resultierenden Wissensstrukturen bilden gemeinsam das Professional Congress Organizer Knowledge und werden über Such-, Browsing- und Navigationsfunktionen zugänglich gemacht. Für die Autorensuche sind insbesondere die EO beziehungsweise ein EKG von Bedeutung. Darin werden Informationen zu Autorinnen und Autoren, ihren Publikationen, Forschungsthemen und fachlichen Profilen strukturiert verknüpft. Auf dieser Grundlage können thematische Suchanfragen verarbeitet und relevante Autorinnen und Autoren identifiziert werden, deren wissenschaftliche Expertise zu einem bestimmten Konferenzthema, Beitrag oder Fachgebiet passt. Die strukturierten Wissensbestände unterstützen damit Anwendungsfälle wie die Auswahl geeigneter Gutachter, Vortragender oder weiterer fachlich passender Autorinnen und Autoren sowie ergänzend die Empfehlung und Planung von Sessions.

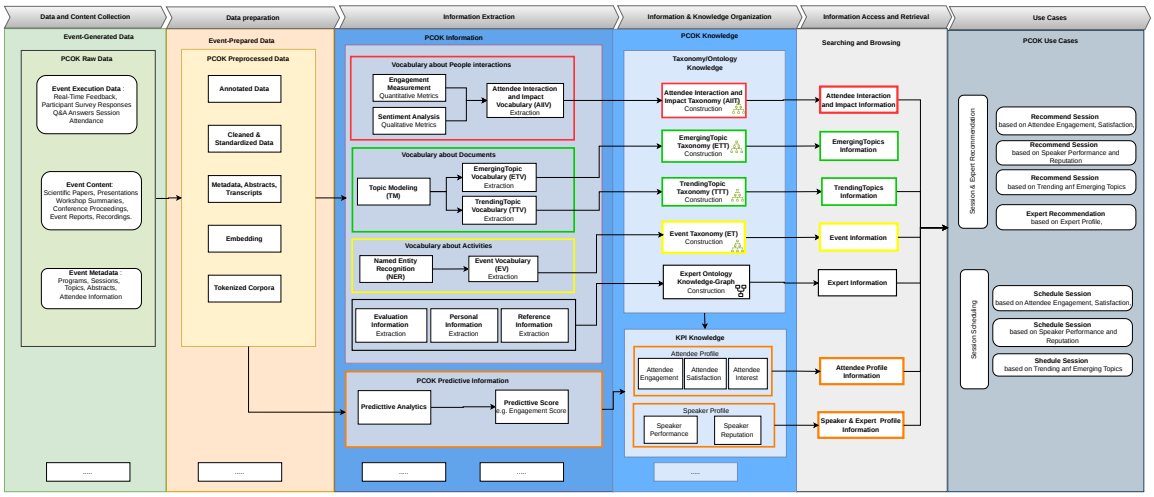

Figure 1: Vorgeschlagene IELC4AS2P-Architektur [22]

Aufgabenbeschreibung

Ziel der Abschlussarbeit ist die Konzeption, prototypische Implementierung und Evaluation eines **cloudbasierten, lifecycle-orientierten Systems** zur **Autorensuche** anhand von Themen-Schlüsselwörtern unter Verwendung von **Vektor-Einbettungen und -Abfragen**, wobei der Schwerpunkt auf modernen **KI-Technologien** wie **Large Language Models (LLMs)** und **KI-Agenten** liegt Das System soll Abstracts und Autorenmetadaten verarbeiten, die aus einem Konferenzmanagementsystem bezogen werden. Diese Dokumente werden in semantische **Vektor-Einbettungen** umgewandelt und in einer Vektordatenbank gespeichert. Wenn Nutzer thematische Suchanfragen eingeben, soll das System relevante Dokumente abrufen und Autoren identifizieren, deren Forschungsschwerpunkte zum angegebenen Thema passen. Im Gegensatz zu herkömmlichen Suchsystemen soll die Architektur LLM-basiertes Suchanfragenverständnis und agentenbasierte Koordinationsmechanismen integrieren. LLMs können genutzt werden, um Nutzeranfragen zu interpretieren, Themen semantisch zu erweitern und Erklärungen für die abgerufenen Ergebnisse zu generieren. KI-Agenten können Systemkomponenten koordinieren, Abruf-Workflows verwalten und transparente Entscheidungswege bereitstellen. Das System sollte als cloudbasierter und reproduzierbarer Prototyp konzipiert werden, idealerweise unter Verwendung von Containerisierungstechnologien wie Docker. Die Verwendung von Open-Source-Software und -Modellen wird empfohlen, um Transparenz, Reproduzierbarkeit und Unabhängigkeit von proprietären Plattformen zu gewährleisten. Die Aufgabe orientiert sich am **IELC** [12] als Referenzmodell, das die Phasen *Data Collection*, *Data Preparation*, *Information Extraction (TM)*, *Information & Knowledge Organization (Taxonomie)* sowie *Information Access & Retrieval* umfasst.

Die Umsetzung der Abschlussarbeit folgt der Forschungsmethodik nach Nunamaker [23] . Die Arbeit gliedert sich in vier Phasen:

Beobachtung (Observation, OP): Untersuchung von NLP-Techniken, Transformer-Modellen, Vektor-Embeddings, Vektor-Retrieval und LLM-basierten Retrieval-Ansätzen. Durchführung einer Literaturrecherche zum aktuellen Stand der Technik und Vergleich verfügbarer Tools und Architekturen für das semantische Dokumenten-Retrieval und die themenbasierte Identifikation relevanter Autorinnen und Autoren.

Theoriebildung (Theory Building, TBP): Entwurf einer Architektur für ein vektorbasiertes Autorensuchsystem, das anhand thematischer Suchanfragen relevante Autorinnen und Autoren identifiziert. Modellierung der Datenstrukturen für Autoren, Abstracts und Metadaten sowie des Workflows für Einbettung und Retrieval. Anwendung nutzerzentrierter Designprinzipien [24] und Erstellung von UML-Diagrammen zur Beschreibung der Systemarchitektur und der Interaktionen zwischen den Komponenten [25].

Systementwicklung (System Development): Umsetzung eines cloudfähigen Prototyps unter Verwendung von Containerisierungstechnologien wie Docker [26]. Die Einbettungsgenerierung und die Vektorabfrage sind unter Verwendung geeigneter Modelle zu integrieren. Eine LLM-Komponente zur Interpretation von Suchanfragen und zur Erläuterung der Suchergebnisse ist zu integrieren, und ein KI-Agent, der für die Koordination der Suchaufgaben zuständig ist, ist zu implementieren.

Experimentierung und Evaluation (Experimentation): Bewertung des Systems anhand quantitativer Metriken wie Precision@k, Recall@k und Mean Reciprocal Rank. Durchführung qualitativer Bewertungsmethoden wie kognitiver Walkthroughs [27], um die Benutzerfreundlichkeit und Erklärbarkeit aus der Perspektive eines Organisators zu beurteilen. Zusammenfassung der Ergebnisse und Erörterung von Einschränkungen und zukünftigen Verbesserungsmöglichkeiten.

Literatur

[1] Matthias Schultze Wilhelm Bauer, Stefan Rief. Ökosysteme im wandel–zukunftsszenarien für business events im zeitalter grenzenloser kommunikation. In *Future Meeting Space. Fraunhofer-Institut für Arbeitswissenschaft und Organisation IAO [Themenheft]*. Fraunhofer IAO, 2022.

[2] Hassnah Wee, Nurul Dafiqah Kamarulzaman, and Muhd Saufi Anas. Mice industry survival: A systematic literature review. *International Journal of Academic Research in Business and Social Sciences*, 2022.

[3] Edgar Göll and Michaela Evers-Wölk. Tagung und kongress der zukunft - management summary, 01 2013.

[4] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[5] Andre Miguel Japiassú. How to prepare and submit abstracts for scientific meetings. *Revista Brasileira de Terapia Intensiva*, 25:77–80, 2013.

[6] Yordan Kalmukov. Architecture of a conference management system providing advanced paper assignment features. *arXiv preprint arXiv:1111.6934*, 2011.

[7] J.-D. Kim, T. Ohta, Y. Tateisi, H. Mima, and J. Tsujii. Xml-based linguistic annotation of corpus. *In Proceedings of the first NLP and XML Workshop*, pages 47–53, 2001.

[8] Inmaculada Martín-Rojo and Ana Gaspar González. The impact of social changes on mice tourism management in the age of digitalization: a bibliometric review. *Review of Managerial Science*, pages 1–24, 02 2024.

[9] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearbook of Medical Informatics*, 17(01):128–144, August 2008.

[10] Ralph Grishman. Information extraction: Techniques and challenges. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology: International Summer School, SCIE-97 Frascati, Italy, July 14–18, 1997*, pages 10–27. Springer, 1997.

[11] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60, 2024.

[12] Philippe Tamla. Towards a reference model for an information extraction lifecycle supporting data collection, data preparation, information extraction, information and knowledge organization, and information access and retrieval. Technical report, FernUniversität in Hagen, 2023. Available via the institutional repository of FernUniversität in Hagen.

[13] FTK. Artificial Intelligence for Hospitals, Healthcare & Humanity (AI4H3). R&D White Paper, FTK e.V. Research Institute for Telecommunications and Cooperation, Dortmund, Germany, April 2025.

[14] Chair of Multimedia and Internet Applications. http://www.lgmmia.fernuni-hagen.de, 2023.

[15] FTK e.V. Research Institute for Telecommunications and Cooperation. http://www.ftk.de, 2023.

[16] FTK. Artificial Intelligence for Hospitals, Healthcare & Humanity (AI4H3). R&D White Paper, Dortmund, Germany, April 2020.

[17] Bielefeld University. RATIO: Rationalizing Recommendations (RecomRatio), 2017.

[18] Binh Vu, Yanxin Wu, Haithem Afli, Paul Mc Kevitt, Paul Walsh, Felix Engel, Michael Fuchs, and Matthias Hemmje. A metagenomic content and knowledge management ecosystem platform. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–8. IEEE, 2019.

[19] Globit – global information technology gmbh, 2025.

[20] Rachael Lammey. Using the crossref metadata api to explore publisher content. *Science Editing*, 3(2):109–111, 2016.

[21] Patrice Lopez. Grobid. `https://github.com/kermitt2/grobid`, 2008–2024.

[22] FTK. Cloud based event organization and management. R&D White Paper, Dortmund, Germany, April 2025.

[23] Jay F. Nunamaker Jr, Minder Chen, and Titus D. M. Purdin. Systems Development in Information Systems Research. *Journal of Management Information Systems*, 7(3):89–106, December 1990. Publisher: Routledge _eprint: https://doi.org/10.1080/07421222.1990.11517898.

[24] Roy D Pea. User centered system design: new perspectives on human-computer interaction. *Journal educational computing research*, 3:129–134, 1987.

[25] Martin Fowler. *UML Distilled: A Brief Guide to the Standard Object Modeling Language*. Addison-Wesley Professional, 2004. Google-Books-ID: nHZslSr1gJAC.

[26] Inc. Docker. Docker: Build safer, share wider, run faster, 2021.

[27] Peter G. Polson, Clayton Lewis, John Rieman, and Cathleen Wharton. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, 36(5):741–773, May 1992.