

Schlussbericht

HARBSAFE2 (Verbundvorhaben)

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages



Verbundvorhaben: Softwareassistierte Harmonisierung

Zuwendungsempfänger Technische Universität Braunschweig Institut für Intermodale Transport- und Logistiksysteme (ITL) Hermann-Blenk-Straße 42, 38108 Braunschweig	Förderkennzeichen 03TN0018A-C
DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE Merianstraße 28, 63069 Offenbach am Main	
INOSOFT GmbH Im Rudert 15, 35043 Marburg	
Projektleiter Pannek, Prof. Dr. Jürgen	Tel.: +49 531/391-66300 Fax: +49 531/391-66399 Email: itl@tu-braunschweig.de
Laufzeit des Vorhabens von: 01.10.2020	bis: 30.09.2022
Berichtszeitraum von: 01.10.2020	Datum bis: 30.09.2022
	31.03.2023

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Wirtschaft und Energie unter dem Förderkennzeichen 03TN0018A-C gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Inhalt

I	Kurzdarstellung.....	4
I.1	Aufgabenstellung.....	4
I.2	Voraussetzungen, unter denen das Vorhaben durchgeführt wurde	4
I.3	Planung und Ablauf des Vorhabens	6
I.4	Wissenschaftlicher und technischer Stand, an den angeknüpft wurde.....	11
I.5	Zusammenarbeit mit anderen Stellen.....	12
II	Eingehende Darstellung	13
II.....	13
II.1	Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele	13
II.2	Wichtigste Positionen des zahlenmäßigen Nachweises.....	41
II.3	Notwendigkeit und Angemessenheit der geleisteten Arbeit.....	41
II.4	Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen.....	42
II.5	Erfolgte oder geplante Veröffentlichungen der Ergebnisse.....	42
III	Literaturverzeichnis	44

Abbildungsverzeichnis

Abbildung 1 Logische Architektur mit Datenflüssen in der Synchronisationsfunktion.	10
Abbildung 2 Ein geclustertes Bedeutungsspektrum mit beispielhaft vergebenen Cluster-Labeln für einen sinnvollen Schwellenwert t.....	23
Abbildung 3 Gesamtsystem.....	35
Abbildung 4 Browser-Ansicht der entwickelten Web-Plattform http://e-glossary.dke.de	38

Tabellenverzeichnis

Tabelle 1: Bezeichnungen mit großer Definitionsvielfalt im „Clause 3, Terms and Definitions“- Korpus.....	5
Tabelle 2: Aus den Arbeitspaketen ergibt sich die folgende Planung personeller Ressourcen (in Stunden).	7
Tabelle 3: Arbeits- und Zeitgrobplanung.....	8
Tabelle 4: Geplante und aufgewendete personelle Ressourcen.	13
Tabelle 5: Hinweistypen zu detektierbaren Anforderungsverstößen.....	20
Tabelle 6: Qualitätsmetriken.....	26
Tabelle 7: Service-Endpunkte des TermAnalyzers	29
Tabelle 8: Entwicklungsablauf der Web-Applikation	37
Tabelle 9: Projektpräsentationen und Workshops	40

I Kurzdarstellung

I.1 Aufgabenstellung

Harbsafe 2 projektierte die Prüfbarkeit und Qualität des vollständigen Terminologiebestands der IEC als Normungsorganisation zu verbessern. Ziel war es, eine moderne Datenbankanwendung für Normungsexpert*innen zu entwickeln. Gemäß Aufgabenstellung wurde ein prototypische Web-Anwendung umgesetzt und in Pilotanwendungen eingesetzt. Der Prototyp ermöglicht, anwenderfreundlich sämtliche Begriffe zu recherchieren und stellt ein niedrigschwelliges Angebot an die Normenexperten und das breitere technische Publikum dar. Zusätzlich werden drei innovative automatische Verfahren angewendet:

- **HintAn:** konkrete Harmonisierungshinweise, einerseits bezüglich definitorischer Inkonsistenzen zu anderen Einträgen, andererseits bezüglich der Konformität der Definitionen gemäß IEC Direktiven und Terminologienormen.
- **Bedeutungsspektren:** eine Übersicht über Definitionsvarianten durch Sortierung und Gruppierung aller Einträge zu einer Benennung im Suchergebnis.
- **SemAn:** eine semantische Analysemethode, die eine schnellere Lesbarkeit der einzelnen Definitionen durch die Zerlegung in Bedeutungskomponenten und deren hierarchisierte Darstellung erreicht.

Diese harmonisierungsrelevanten Annotationen identifizieren Inkonsistenzen, geben einen komfortablen Überblick über die Vielzahl definitorischer Varianten und schlüsseln die Bedeutungsmerkmal jeder einzelnen Definition plausibel auf. Zusätzlich waren Metriken zu entwickeln, die den Harmonisierungsgrad quantitativ abbilden, um Harmonisierungsaktivitäten zu konzentrieren und steuerbar zu machen.

Um bei den Problemstellungen in den Themenfeldern: Homonymie, Definitionsqualität und Definitionslesbarkeit, zu assistieren, kommen jeweils unterschiedliche Verfahren zum Einsatz. Ihnen liegt ein Natural-Language-Processing-Modell zugrunde, das Definitionen auswertet und annotiert, um Wortstämme, Wortarten und grammatische Abhängigkeitsstrukturen zu identifizieren [1]. Um homonyme Definitionen auszuwerten, kommt ein state-of-the-art Machine-Learning-Modell von Reimer & Gurevych zum Einsatz (Sentence-BERT) [2]. Mittels Semantic Textual Similarity (STS) werden Texte inhaltlich vergleichbar und können algorithmisch vorstrukturiert werden.

I.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Die klassische auf Sektoren bezogene Organisationsstruktur der Normung [3] und die dokumentbezogene Terminologiearbeit führen zu einer Situation, in der dieselben Gegenstände und Konzepte auf teils widersprüchliche, teils irreführende Weise mehrfach definiert sind. Allein die Expertise der Normungs-Community kann die begrifflichen Herausforderungen lösen. Das Problem ist: Angesichts einer sechsstelligen Anzahl von Begriffseinträgen, verstreut in tausende Normen und Standards der IEC, bekommen Expert*innen die Divergenz und Bedeutungsvielfalt kaum gänzlich ins Blickfeld. Terminologie-Harmonisierung bedeutet, bestehende Definitionen prüfen, angeleichen und wiederverwenden. Aber der Aufwand allein für die Sichtung des Bestandes ist prohibitiv hoch.

Wenn dieselbe Bezeichnung für zwei oder mehrere unterschiedliche Konzepte stehen kann, sprechen wir von Homonymie. Hierin besteht eine wichtige Quelle von Ambiguitäten und Widersprüchen. Ursachen sind versehentliche Ad-hoc-Formulierung im Gremium, sowie konkurrierende Konzepte in der Praxis. Im Projekt wurden ca. 64.000 Begriffseinträge aus den 3.000 aktuellen und digital verfügbaren Terminologieteilen des IEC-Normenschatzes ausgewertet. Darin finden sich ca. 43.000 unterschiedliche Bezeichnungen. Von diesen wurden knapp 40% in unterschiedlichen Normen mindestens einmal unterschiedlich definiert. Zu 2.250 (370) Bezeichnungen finden sich mehr als fünf, zu 370 mehr als 15 nicht-identische Definitionen. Zu den Homonymen mit größter Diversität in der Definition gehören Grundkonzepte wie „system“, „device“, „node“, aber auch recht spezifische wie „rated current“ oder „rated voltage“ (siehe Tabelle 1: Bezeichnungen mit großer Definitionsvielfalt im „Clause 3, Terms and Definitions“-Korpus).

Der hohe Anteil an Mehrfachdefinitionen ist ein deutliches Zeichen für den Harmonisierungsbedarf. Viele aber bei weitem nicht alle unterschiedlichen Definitionen deuten auf konzeptionelle Widersprüche hin. Häufig unterscheiden sich zwei Definitionen nur in Syntax und ihren grammatischen Funktionswörtern. Gelegentlich werden überflüssigerweise inhaltlich äquivalente Phrasen formuliert. Andere werden in verschiedenen Domänen unterschiedlich verwendet.

Tabelle 1: Bezeichnungen mit großer Definitionsvielfalt im „Clause 3, Terms and Definitions“-Korpus

Bezeichnung	Einträge	formal unterschiedliche Definitionsvarianten	inhaltlich unterschiedliche Definitionen (automatisch prädiziert in Harbsafe2)
rated voltage	90	72	40
user	49	43	35
rated current	47	44	35
device	116	44	34
component	54	44	33
system	71	42	30
response time	50	43	29
manufacturer	53	42	28
normal operation	38	35	27
operator	39	29	26
service	69	34	26
channel	48	35	25
frame	89	30	25
node	70	37	23

Um die Problematik zu adressieren, pflegt das IEC das International Electrotechnical Vocabulary (IEC 60050, IEV), sowie die Web-Datenbank [4]. Dieses stellt konsolidierte, einheitliche und widerspruchsfreie Referenzdefinitionen bereit. Das IEV umfasst rund 20.000 Einträge aus unterschiedlichsten Domänen, wobei einige wichtige Terminologiebestände (zum Beispiel IEC TC 65) noch nicht berücksichtigt sind. Die Übereinstimmung von gültigen Norm-Definitionen mit IEV-Referenzeinträgen ist allerdings nicht immer gegeben. Und die Weiterentwicklung von IEC 60050 ist massiv vom dargestellten Problem betroffen.

Die Qualität der Normungsterminologie ist eine wichtige Voraussetzung für die Nutzbarkeit von Normen und Standards. Widersprüche und Inkonsistenzen innerhalb von Normenreihen und Einzelpublikationen stellen ihre Anwendbarkeit vor erhebliche Schwierigkeiten. Bei Widersprüchen zwischen benachbarten Normen und Standards sind Ambiguitäten in der Anwendung und höhere Aufwände in Schulung und Weiterbildung kaum zu vermeiden. Bei anhaltendem Wachstum der Publikationszahl und zunehmender interdisziplinärer Vernetzung der Anwendungen werden die genannten Herausforderungen stetig dringlicher. Daher ist die Harmonisierung der begrifflichen Grundlagen des technischen und institutionellen Wissens als Erfolgsfaktor mit Querschnittswirkung zu betrachten. Denn ohne diese werden sich neue Geschäftsmodelle der Digitalisierung nicht umsetzen lassen, wenn Maschinen Normen und Standards interpretieren und anwenden sollen. Insbesondere für die Normung und Standardisierung auf jungen, innovativen Marktsegmenten (z. B. die verschiedenen Formen der Elektromobilität, Smart Home, künstliche Intelligenz oder auch Industrie 4.0) ist eine gezielte Terminologie-Harmonisierung notwendig.

Bei der Digitalisierung der Normung ist eine digitale Schnittstelle für die Normungsdaten (XML-Bereitstellung) NISO-STS entstanden, die einen strukturierten Zugriff auf die erforderlichen Inhalte ermöglicht. Der Zugriff erfolgt auf den kompletten Normenbestand der IEC und erlaubt die automatisierte Überführung in die verschiedenen Anwendungen. Unter dieser Grundvoraussetzung kann nun die automatische Auswertung von Normungsinhalten datentechnisch dargestellt werden.

I.3 Planung und Ablauf des Vorhabens

Die Planung erfolgte unter der Führung der TU Braunschweig und konnte weitgehend eingehalten werden. Die in der Antragsphase erarbeiteten Arbeitspakete wurden weitgehend termingerecht erledigt.

Geplanter Personalaufwand

Tabelle 2: Aus den Arbeitspaketen ergibt sich die folgende Planung personeller Ressourcen (in Stunden).

AP1: Abstimmung institutioneller Schnittstellen und prädizierbarer Harmonisierungstasks (Leitung: ITL)			
Partner:	ITL	DKE	INOSOFT
	480	390	40
AP2: Auswahl und Annotation von Terminologieelementen als Trainings- und Testdatensets (Leitung: ITL)			
Partner:	ITL	DKE	INOSOFT
	520	160	20
AP3: Modellierung und Klassifikation der Terminologiedaten mit Machine-Learning-Verfahren (Leitung: ITL)			
Partner:	ITL	DKE	INOSOFT
	600	150	20
AP4: Globale Konsistenzmetriken entwickeln (Leitung: ITL)			
Partner:	ITL	DKE	INOSOFT
	320	80	20
AP5: Entwicklung der Datenservices (Leitung: ITL)			
Partner:	ITL	DKE	INOSOFT
	420	80	20
AP6: Technische Integration der Komponenten (Leitung: DKE)			
Partner:	ITL	DKE	INOSOFT
	100	250	400
AP7: Design und Implementierung der Webkomponente und Proposalgenerierung (Leitung: INOSOFT)			
Partner:	ITL	DKE	INOSOFT
	200	320	1404
AP8: Öffentlichkeitsarbeit, Präsentation und Schulung (Leitung: DKE)			
Partner:	ITL	DKE	INOSOFT
	240	680	40

Das Projekt gliederte sich grob in drei Arbeitsfelder.

- 1) In AP 1 und AP 8 wurden Stakeholder aus der Normungscommunity adressiert. AP 1 diente der Detailabstimmung von Projektergebnissen anhand der Anforderungen aus der Normungspraxis, der Erprobung von Verfahren anhand konkreter Aufgabenstellungen, sowie der Entwicklung von ergänzenden Normungsdokumenten (VDE Spec) und Handlungsempfehlungen. AP 8 diente der Dissemination der Projektergebnisse durch Publikationen und Vorträge.
- 2) AP 2, AP 3 und AP 4 diente der Erarbeitung, Modellierung und Evaluation der computerlinguistischen und Machine-Learning-basierten harmonisierungsrelevanten Annotationen: HintAn, SemAn und Bedeutungsspektren.
- 3) AP 5, AP 6 und AP 7 beinhalteten ein komplexes Software-Entwicklungsprojekt in verteilten Teams mit Beiträgen von TU Braunschweig, DKE und INOSOFT AG. Hier wurde die Daten-Pipeline von den Quelldaten, über die Anreicherung mit harmonisierungsrelevanten Annotationen bis hin zur Auslieferung und Präsentation beim Web-Anwender geplant, entwickelt, integriert und getestet.

Die Meilensteine dienten der zeitkritischen Koordinierung zwischen den Tätigkeitsfeldern. Bis zum Meilenstein in Projektmonat 13 mussten die Abstimmungen der Inhalte und Nutzerschnittstelle mit den Normungsexperten auf der einen Seite und die Erarbeitung der harmonisierungsrelevanten Annotationen auf der anderen Seite strukturell abgeschlossen und inhaltlich soweit fortgeschritten sein, dass die Implementierungen auf dem Arbeitsfeld der Software-Entwicklung risikofrei anlaufen konnten. Dieses Vorgehen konnte umgesetzt

werden und hat sich als zielführend erwiesen. Im strukturell festen Rahmen konnten verbindliche Datenbank- und Software- und UI-Modelle erzeugt werden, die die Entwicklung auf ein sicheres Fundament stellten, aber genügend Flexibilität für Feinarbeiten an Anzahl, Inhalt und Performanz der prädizierten Annotationen boten.

Zeitgrobplanung

Tabelle 3: Arbeits- und Zeitgrobplanung

Monate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
AP1																								
AP2																								
AP3																								
AP4																								
AP5																								
AP6																								
AP7																								
AP8																								



Auf dem **Arbeitsfeld 1**) konnte bereits früh im Projekt Zustimmung von Normungsexperten zu den Entwürfen der Benutzerschnittstelle und den geplanten Funktionen generiert werden (DKE K111, DKE UK 931.2). Aufbauend auf das Vorgängerprojekt „Harbsafe“ standen Modelle der Nutzerschnittstelle am Projektbeginn zur Verfügung und blieben über die gesamte Projektlaufzeit im Kern stabil und konnten schrittweise verfeinert und iterativ verbessert werden.

In der ersten Projekthälfte wurden Struktur und Inhalt des Begriffseintragskorpus, normative Anforderungen an Quelldaten und Datenintegration erhoben und Bedienkonzepte abgestimmt.

Zur Sicherstellung der Adäquatheit wurden Entwürfe mit projektinternen Normungsexperten validiert, sowie Präsentationen in diversen Gremien durchgeführt. Das betrifft die UI-Entwürfe und die automatischen, harmonisierungs-relevanten Annotationen. Letztere wurden vor allem aus den Anforderungen aus IEC DIR1:2019 abgeleitet, auf Umsetzbarkeit überprüft und mit Normungsexperten abgestimmt.

Auf Basis der Ergebnisse in Arbeitsfeld 2) wurden in der zweiten Projekthälfte eine Pilotanwendung für den Common-Data-Dictionary-Korpus aus IEC/SC 65E, WG 2 durchgeführt. Der vorgesehene praktische Einsatz in der Gremienarbeit war im Rahmen von IEC/TC 65, UK 931.2 (vormals IEC/TC 65, UK 921.1) zur Erarbeitung der IEV-Section "Automatisation" vorgesehen, musste jedoch aufgrund von krankheitsbedingter Inaktivität des Gremiums auf den Verwertungszeitraum verschoben werden. Wie geplant, konnten nach Erreichung des letzten Meilensteins aus den Befunden Normungsprodukte erstellt werden. Diese wurden in Form einer VDE SPEC und im Rahmen der Normungsroadmap 4.0 entworfen und zur Abstimmung und Edition an die DKE übergeben.

Bei der Dissemination der Ergebnisse in AP 8 konnte auf Zwischenergebnisse aus den übrigen Arbeitsfeldern aufgebaut werden. Das betrifft die Projektergebnisse in AP8, die von Beispiel-Implementationen und explorativen Prädiktionen getragen. Veröffentlicht wurden u.a. zwei Fachartikel in DIN-Mitteilungen, eine Projektdarstellung auf der VDE Homepage, diverse Vorträge in Gremien und ein Abschlussworkshop zum Projektende.

Das **Arbeitsfeld 2**) umfasste die Modellierung der Machine-Learning- und Natural-Language-Processing-basierten Funktionen: HintAn, SemAn und Bedeutungsspektren.

Hier wurden zunächst in AP 3 Datensets erhoben und manuell für Training- und Validierung der automatischen harmonisierungsrelevanten Annotationsalgorithmen ausgewertet. Zum Projektbeginn wurden zunächst diverse Semantic-Textual-Similarity(STS)-Modelle auf Eignung zur Bewertung von Inkonsistenzen und für die automatische Definitionsgruppierung der Bedeutungsspektren überprüft. Anschließend wurde das Language-Modell mit der besten Performance ausgewählt und weitere Optimierungsoptionen eruiert. Das Trainings- und Validierungsset umfasst ca. 7.000 Definitionsvergleiche, deren manuell bewertete Ähnlichkeiten mit dem Modelloutput korreliert werden konnten, um STS-Modelle zu evaluieren und zu trainieren. Die Evaluation konnte in AP4 erfolgreich durchgeführt werden. Auf Basis des STS-Metrik konnte mittels hierarchischem Clustering eine sinnvolle Gruppierung und Sortierung der Bedeutungsspektren jedes Homonyms, sowie eine intuitive Klassifikation der Inkonsistenzen zwischen Definitionen eines Homonyms gefunden werden.

Anschließend wurden in AP 3 für die HintAn-Konformitätsprüfung der Definitionen verschiedene Normungsanforderungen aus ISO/IEC DIR1:2019 abgeleitet, auf automatische Prüfbarkeit getestet und auf Plausibilität bei der manuellen Bewertung überprüft. Dann wurde ein repräsentatives Validierungsset (ca. 2.000) von zufällig selektierten Definitionen erhoben und bereinigt, die in umfangreichen Rating-Tasks jeweils auf Anforderungsverstöße überprüft wurden. Die Entwicklung der Trainingssets konnte bis zum Meilenstein in Projektmonat 13 strukturell abgeschlossen werden. Ergänzende Bewertungen zur Detailschärfung wurden im Anschluss bedarfsoorientiert hinzugefügt. Anhand des Validierungssets konnte in AP4 die HintAn-Konformitätsprüfung umgesetzt werden, indem mittels NLP-Pipeline regelbasiertes Matching umgesetzt wurde. Dabei wurden aus den Beispielen Indikationen und Kontraindikationen ein.

Das im Projekt neu entwickelte SemAn-Verfahren strukturiert einzelne Definitionen in seine Bedeutungskomponenten und stellt sie in einem logisch-hierarchischen Format dar, um Lesbarkeit zu verbessern und die Konzeptualisierung visuell zu erleichtern. Es handelt sich um eine verlustfreie Darstellung mit Strukturelementen auf Zeichenebene (“Pretty Print”). Die Entwicklung des Formats erfolgte iterativ in Kollaboration mit Causalis an einem kleineren Definitionsset (ca. 150), indem grammatische Regeln für die Elementarisierung, Subordination und Koordination von Bedeutungskomponenten gefunden und am Kleinkorpus erprobt wurden. Im Anschluss ist es gelungen, den so entstandenen Goldstandard algorithmisch durch Auswertung von NLP-Annotationen aus Dependency Parsing der Definition mit hoher Genauigkeit zu reproduzieren.

In AP 4 wurden Qualitätsmetriken erarbeitet, die den Harmonisierungsstand und Fortschritt

und die entsprechenden Qualitätsziele der Messbarkeit zuführen. Hierbei war zu zeigen und zu dokumentieren, welche messbaren Eigenschaften des Terminologiebestandes sich durch automatisch generierte Maßnahmenvorschläge im Sinne einer Harmonisierung beeinflussen lassen. Eingehen konnte neben einer Dublettenrate zum Beispiel ein Maß für die semantische Divergenz von Definitionen derselben Benennung (Mehrfachdefinitionen), sowie Maße zur Benennungskonsistenz und diverse weitere.

ITL entwickelt statistische Verfahren zur Erhebung zentraler Kennzahlen für die Beschreibung des globalen Harmonisierungszustandes der DKE-Terminologie. Die Ergebnisse können als Teilerfolg auf dem Weg zu globalen KPIs zum Harmonisierungszustand gelten.

Die Entwicklungen zu SemAn, HintAn und zu den Semantic Spectrums konnten weitgehend abgeschlossen, positiv evaluiert und im Pilotprojekt sinnvoll eingesetzt und in Arbeitsfeld 3 integriert werden. Nicht gelungen ist dies für die Qualitätsmetriken, die den Harmonisierungsstand und -fortschritt gemessen an plausiblen Qualitätszielen der Messbarkeit zuführen sollten. Die im Projekt erarbeiteten Metriken (TU Braunschweig) erschienen bei Projektabschluss zu rudimentär und unvollständig, um die Harmonisierungsqualität realistisch abzubilden.

Das **Arbeitsfeld 3**) umfasste sämtliche Aktivitäten des Software-Projekts für den Einsatz der erarbeiteten Verfahren und Annotationen im Rahmen einer anwenderfreundlichen Web-Applikation. Die Anforderungen zur Nutzerschnittstelle aus AP 1 sowie die harmonisierungsrelevanten Annotationsverfahren aus AP 3 konnten zum Meilenstein im Monat 13 freigegeben werden. In dem Zeitraum wurde die grobe Software-Architektur abgestimmt, System-Anforderungen erarbeitet und seitens der DKE eine Serverumgebung geplant und beschafft.

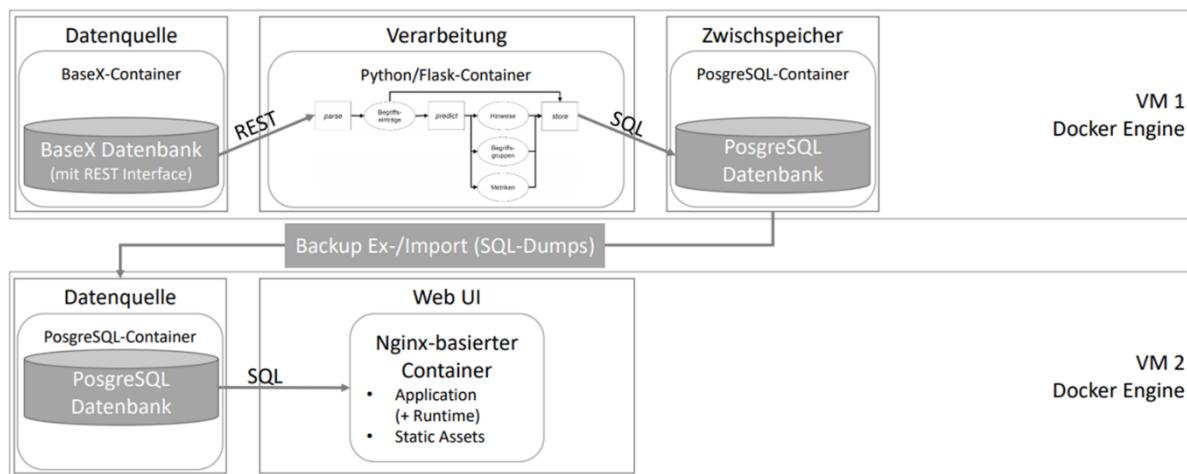


Abbildung 1 Logische Architektur mit Datenflüssen in der Synchronisationsfunktion.

In AP 5 wurden seitens der TU Braunschweig die erarbeiteten Machine-Learning-Modelle und NLP-Verfahren aus AP 3 in eine modulare Pipeline aus Bereinigung der Quelldaten (Data-Cleaning) und Annotationsanreicherung (Data-Augmentation and ML-Prediction) implementiert, in einem selbstdokumentierenden REST-Service gekapselt, mit der DKE abgestimmt und ausgeliefert. Neben der Synchronisationsfunktion zur Verarbeitung und Annotation der Quelldaten für die Ablage im Zwischenspeicher sind auch On-Demand-Funktionen entstanden, die beliebige Terminologiedaten auf harmonisierungsrelevante

Annotationen auswerten.

In AP 6 wurde von der DKE ein Backend-System (VM) abgestimmt, aufgebaut und darin der REST-Service (AP 5) und die Modelle aus AP 3 containerisiert integriert. Das Backend-System stellt das Gesamtsystem ohne Ausgabekomponente (AP7) dar und wurde in der IT-Umgebung der DKE verankert. Die Datenbereitstellung der Quelldaten wurden, basierend auf den IEC Normdokumenten im NISO-STS-Format, in eine BaseX-Datenbank integriert und in einen REST-Service gekapselt, der Terminologie-Einträge gemäß Korpus-Definition extrahiert und zur Synchronisation freigibt. Um den hohen Anforderungen an Datensicherheit bezüglich der Norm-Dokumente zu entsprechen, wurde eine physische Trennung zwischen Backend-System und der Web-offenen VM 2 eingerichtet und ein manueller Synchronisationsprozess der verarbeiteten Ergebnisse aus dem Zwischenspeicher mit dem Serversystem der Webapplikation "Enhanced Glossary" (e-glossary.vde.com) für Normungsexperten abgestimmt (AP 7).

Die Webapplikation wurde in AP 7 entwickelt. Im Zusammenspiel zwischen TU Braunschweig und INOSOFT AG wurden Anforderungen in User Stories und Wireframes schrittweise verfeinert und von INOSOFT agil umgesetzt. System-Anforderungen wurden mit der DKE abgestimmt. Nutzer- und Abnahmetests erfolgten bei der TU Braunschweig.

I.4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Terminologieharmonisierungsprozesse im Normungswesen bildet die internationale Norm ISO 860:2007 [5] ab, die die Harmonisierung inkonsistenter Terminologieressourcen auf nationaler und multinationaler Ebene adressiert. Sie geht davon aus, dass Terminologieressourcen unabhängig voneinander erstellt werden. Einzelne Begriffe und die Begriffssystemstruktur sind nicht hundertprozentig äquivalent, was ihren Harmonisierungsbedarf begründet. Der verteilte dokumentbezogene Normerarbeitungs-Prozess führt beinahe zwangsläufig zu Mehrfach-Definitionen und inkongruenten Begriffssystemen. Technische Fortschritte im Bereich von neuronalem Machine-Learning und vor allem im Feld von Semantic-Textual-Similarity (STS) versprechen Lösungen bei der automatischen Identifikation und Bewertung von Inkonsistenzen in verteilten Konzeptsystemen.

Im Vorgängerprojekt Harbsafe [6] wurden daher erstmalig Concept Embeddings für Begriffseinträge aus Normen erstellt, um die inhaltlichen Beziehungen und Inkonsistenzen zwischen den Begriffssystemen mehrerer Normen maschinell abzubilden und beherrschbar zu machen. Dies beruhte auf Sprunginnovationen im Bereich des Natural-Language-Processing (NLP), Text-Mining und Information Retrieval (IR). Einen der einflussreichsten Beiträge lieferten Mikolov [7]: Die neuronale Trainingsmethode word2vec ermöglichte erstmals das effiziente unüberwachte Training von Maschinen-verarbeitbaren semantischen Wortrepräsentationen auf Grundlage von großen natürlichsprachlichen Textkorpora von bis zu einer Billion Wörtern. Die Aggregation von Wort-Vektoren zu semantischen Vektorrepräsentationen von ganzen Sätzen oder Texten (STS) wurde durch die Verfahren Quick-Thoughts [8], InferSent [9], MILA/MSR's General Purpose Sentence Representations [10], Google's Universal Sentence Encoder [11] u.a. begründet. In Harbsafe kamen Bag-of-Word-Ansätze aus gewichtet gemittelten Wortvektoren zum Einsatz [12].

Seither hat eine neue neuronale Netzarchitektur für weitere Fortschritte gesorgt [13]. Sogenannte “Transformer” haben einen neuen Stand der Technik unter anderem im semantischen Satzpaar-Vergleich definiert. Die umfangreichen neuronalen Netzwerke sorgen jedoch nicht nur in der Trainingsphase, sondern auch in der Anwendung für erheblichen rechnerischen Overhead. Sentence-BERT (SBERT) von Reimers (2019) [2] erweitert das BERT-Modell durch den Einsatz einer siamesischen Architektur, die semantisch bedeutungsvolle Satzembeddings generiert und effizient per Cosinusähnlichkeit vergleichbar macht, um die Rechenzeit um mehrere Größenordnungen zu senken. Durch die Performance-Gewinne in der Effizienz wurden die Prädiktions- und Klassifikationstask aus Harbsafe 2 technisch in der erreichten Qualität wirtschaftlich und praktisch umsetzbar.

Neben der Semantic-Textual-Similarity (STS) konnte in Harbsafe 2 auch an technische Fortschritte im Bereich des Natural-Language-Processing (NLP) angeknüpft werden. Der De-Facto-Industriestandard, das NLP-Framework “spaCy” [1] bietet state-of-the-art Tokenizing, Part-of-Speech-Tagging und Dependency-Parsing regelbasiert und auf Basis von neuronalen Netzen. Die Anwendung von NLP-Verfahren beruht methodisch ferner auf der Arbeit von Henning Femmer zur automatischen Detektion sogenannter “requirement smells” [14]. Analog zu Qualitätsmängeln von Anforderungen konnten in Harbsafe2 Qualitätsmängel von Definitionen aufgefunden werden.

I.5 Zusammenarbeit mit anderen Stellen

Durch die Mitwirkung in wissenschaftlichen und Normungsnetzwerken wurden Synergien geschaffen. In der Normung wurde zahlreiche Gremien angesprochen, Ergebnisse verbreitet und Anforderungen aufgenommen (siehe Abschnitt 2.1, AP 8). Hervorzuheben ist die Pilot Anwendung für IEC TC 65/SC 65E, WG 2 und das deutsche Spiegelgremium AK 931.0.1. Hier konnten die Common-Data-Dictionary-Daten des Gremiums evaluiert, der Nutzen des Harbsafe2-Ansatzes validiert und Anwendungs-bezogen weiterentwickelt werden. Ein fruchtbare Informationsaustausch fand zudem mit der Terminologiestelle DIN-TERMKONZEPT statt. Ferner wurden auf den methodischen Erkenntnissen aus Harbsafe2 Handlungsempfehlung für die Normungsbereiche Elektromobilität, Smart Home, künstliche Intelligenz und Industrie 4.0 für die Normungsroadmap I4.0 zusammen mit Vertretern aus AK STD_1941.0.1 erarbeitet.

Wissenschaftlich konnte von einem Austausch mit Prof. Dr. Henning Femmer, Professor für Wirtschaftsinformatik, insb. Requirements Engineering und IT-Qualitätsmanagement profitiert werden. Analogien zu Ansätzen der automatischen Qualitätsprüfung im Requirements Engineering wurden festgestellt und konnten fruchtbar gemacht werden. Zudem konnten Erfahrungen aus Harbsafe2 in das Projekt IDiS-NLP über Vertreter des Instituts für Technische Gebäudeausrüstung an der TH Köln einfließen.

Als Vertreter der Normungsexperten im Projekt, zur Abstimmung von Anforderungen und zur Erarbeitung und Beurteilung von fachspezifischen Modellen wurden Arbeiten an Prof. Dr. Peter Ladkin (Causalis) in Auftrag gegeben. Die langjährige Erfahrung als Normungsexperte im Safety-Bereich, sowie der fachliche Hintergrund in Logik und formalen Methoden konnte im Projekt gewinnbringend eingesetzt werden und methodische Ergebnisse in enger Abstimmung mit dem ITL erreicht werden. Des Weiteren wurde die Qualicen GmbH für die

Einrichtung und Implementierung der Cloud-Umgebung sowie für die Konzeption und Architekturarbeit in AP6 beauftragt.

II Eingehende Darstellung

II.1 Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

Die folgende Tabelle gibt einen Überblick über die APs und ihren zeitlichen Aufwendungen:

Tabelle 4: Geplante und aufgewendete personelle Ressourcen.

AP	ITL Gepl. Arbeits- Std.	ITL Benötigte Arbeits-Std	DKE Gepl. Arbeits- Std.	DKE Benötigte Arbeits.std	Inosoft Gepl. Arbeits- Std.	Inosoft Benötigte Arbeits.std
1	480	498	390	429,5	40	40
2	520	620	160	17,5	20	20
3	600	590	150	12	20	20
4	320	89	80	2	20	2
5	420	549	80	268	20	20
6	100	87	250	249,25	400	420
7	200	187	320	44,75	1404	1384
8	240	260	680	763,5	40	58

AP 1: Abstimmung institutioneller Schnittstellen und prädizierbarer Harmonisierungstasks

AP 2: Auswahl und Annotation von Terminologieelementen als Trainings- und Testdatensets

AP 3: Modellierung und Klassifikation der Terminologiedaten mit Machine-Learning-Verfahren

AP 4: Globale Konsistenzmetriken entwickeln

AP 5: Entwicklung der Datenservices

AP 6: Technische Integration der Komponenten (Leitung: DKE)

AP 7: Design und Implementierung der Ausgabekomponente und Proposalsgenerierung

AP 8: Öffentlichkeitsarbeit, Präsentation und Schulung

Die geplanten Ziele konnten weitgehend im geplanten Gesamtrahmen erreicht werden. Abweichungen gegenüber den Aufwandsschätzungen zu den einzelnen Arbeitspakete ergaben sich im Projektverlauf:

AP1: Die geplante Koordinierung und Validierung der Projektergebnisse mit

Normungsgremium UK 931.2 musste krankheitsbedingt entfallen und konnte nicht kurzfristig vollumfänglich ersetzt werden. Dieser Ausfall wurde durch Leistungen von Projektteilnehmern von TU und DKE und zusätzlichen Präsentationsterminen in Normungsgremien aufgefangen.

AP2: Durch die geringere Spezifität der terminologischen Inhalte infolge der Inaktivität von UK 931.2 konnte die Erstellung der Test- und Trainingsdaten weitgehend bei der TU Braunschweig ohne Zuarbeit von der DKE erfolgen.

AP3: Ebenso wurde die NLP-Modellentwicklung noch stärker als ursprünglich geplant auf Seiten der TU Braunschweig verankert.

AP4: Die Entwicklung von Qualitätsmetriken für den Harmonisierungszustand der Normungsterminologie wurde im letzten Projektquartal herab priorisiert und mit einem Zwischenergebnis abgeschlossen. Grund ist das ausdrücklich geringe Interesse der Normungsexperten an Qualitätsmetriken und der erhöhte Aufwands ihrer Entwicklung anhand der erarbeiteten harmonisierungsrelevanten Annotation.

AP5: Die frei gewordenen Ressourcen aus AP4 konnten sinnvoll der Ausentwicklung des prototypischen TermAnalyzers gewidmet werden. Hier wurde eine robuste modular wiederverwendbare Backend-Komponente erreicht. Diese wurde von der DKE getestet und zur Weiternutzung auch in anderen Zusammenhängen vorgesehen.

AP6: Die Softwareintegration der einzelnen Komponenten verlief weitgehend nach Plan.

AP7: Die Entwicklung der Web-Applikation konnte, wie geplant, zwischen TU Braunschweig und Inosoft AG durchgeführt werden. Aufgrund stärkerer Koordinationsarbeiten in AP1 und weitgehendem Konsens im Konsortium erforderte dies geringere Aufwände seitens der DKE.

AP8: Auf die Dissemination wurde infolge der überzeugenden Ergebnisse besonderes Gewicht gelegt.

Erläuterung der einzelnen APs

Es folgt eine detaillierte Darstellung der geplanten Arbeitspakete im Vergleich mit dem durchgeföhrten Ablauf und der Ergebnisse.

Arbeitspaket 1: Zielstellung

AP-Nr. 1	Abstimmung institutioneller Schnittstellen und prädizierbarer Harmonisierungstasks
Zeitraum	Projektmonat 1 bis Projektmonat 23
AP-Leiter	ITL
	Ziele des Arbeitspaketes
	<p>In Zusammenarbeit mit relevanten Stakeholdern in interessierten Normungsgremien werden Werkzeuge und Methoden unter realen Bedingungen im Rahmen von Pilotanwendungen erprobt. Dabei werden Bedingungen ausgelotet, unter denen die Methoden und Softwarelösungen nachhaltig in den Normungsprozess integriert werden können.</p> <p>Prädiktionsergebnisse und Darstellungsformen werden auf Gebrauchstauglichkeit geprüft und optimiert. Die Qualität der Prädiktionsergebnisse wird überprüft und an den Anforderungen der Normungsexperten validiert. An die automatisch generierten Klassifikationen, relationalen Querverweise und semantischen Analysen schließen Fachanwender konzeptuelle Analysen (ConcAn) an und generieren software-assistierte Change Requests, die unmittelbar in die beteiligten Normungsprozesse einfließen (vgl. Anhang 1). Die in „Harbsafe“ entwickelten Methoden werden dabei weiterentwickelt. Vorgesehen ist die Zusammenarbeit mit DKE/K 111 (Spiegel zu IEC/TC 1), um Fragen des übergeordneten IEC-Terminologie-Harmonisierungsprozesses zu klären.</p> <p>Interesse an der Pilotierung der Werkzeuge und Methoden aus Harbsafe II zeigt zudem das IEC/TC 65, UK 921.1, das sich mit einer unüberschaubaren nicht-harmonisierten Terminologiemasse konfrontiert sieht. Unter anderem können auch die DKE-Gremien GK 914 und UK 931.1 angesprochen werden, um mit IEC/SC 65A/AHG 17 hinsichtlich der Terminologie um Safety und Cybersecurity zusammen zu arbeiten.</p> <p>ITL definiert Zielstellungen, Harmonisierungstasks und Strukturmerkmale der automatischen Analyse (Relationstype, Gruppenstruktur, Darstellungsmerkmale). ITL publiziert und präsentiert Analyseergebnisse und Terminologierecherchen in diversen Gremien und Workshops und leitet Änderungsanträge ab. ITL verarbeitet Rückmeldungen, leitet Anforderungen für das zu entwickelnde integrierten ML-Prädiktionsmodell ab, validiert Werkzeuge und Methoden und erhebt Softwarefehler und Kommentare zur Analysequalität. Auf Basis der Erfahrungen wird eine VDE-Anwendungsregel erstellt.</p> <p>In diesem Kontext wird vom ITL ein Unterauftragnehmer eingesetzt. Dieser unterstützt und berät die Arbeiten des ITL als projektinterner Fach- und Normungsexperte und vertritt das Projekt in diversen Normungsgremien. Der Auftragnehmer erstellt Fachentwürfe und Change Requests zur Normungsterminologie und unterstützt bei der Erstellung der VDE-Anwendungsregel.</p>

Output

1. In teil-automatisierten Pilotanwendungen erarbeiteten strukturierten Terminologierecherchen, Harmonisierungstasks und Analyseergebnisse sind direkter Projektoutput und fließen in die Normungsarbeit ein.

2. Aus Erfahrungen und Kommentaren zu den teil-automatisierten Pilotanwendungen werden Anforderungen abgeleitet. Analysebezogene Anforderungen fließen in **AP2**, **AP3** und **AP4** ein. Anforderungen bezüglich der weiteren Gebrauchstauglichkeit fließen in **AP7** ein.
3. In den Gremien validierte und verifizierte Ergebnisse, die von allgemeinem Interesse sind, werden der Öffentlichkeitsarbeit zur Publikation zugeführt (**AP8**).
4. Entwurf der VDE-Anwendungsregel „Methode zur Erkennung und Bearbeitung von Harmonisierungsbedarf von Begriffseinträgen“.

Arbeitspaket 1: Ablauf

Die Abstimmung und Validierung der innovativen Assistenzverfahren und sowie deren Darstellung im Enhanced Glossary stellten einen kritischen und schwer planbarer Teil des Projekts dar. Grundlegende User-Stories, UI-Modelle und Datenbeispiele wurden in Q4/2020 erarbeitet und standen bereits frühzeitig zur Verfügung. Sie wurden während der Projektlaufzeit in verschiedenen Normungsgremien vorgestellt und positiv aufgenommen. Durch diesen Austausch konnten diverse Anforderungen und Prioritäten im Projekt berücksichtigt werden. Analysebezogene Anforderungen wurden in **AP3** und **AP4** adressiert und Anforderungen bezüglich Gebrauchstauglichkeit sind in **AP7** eingeflossen.

Einschränkungen / Hinweise:

Die ursprüngliche Planung zur Sicherstellung der Adäquatheit der Projektergebnisse sah ein iteratives Vorgehen vor. Dabei sollten Ergebnisse (UI-Entwürfe, prädizierte harmonisierungs-relevante Annotation und Workflows) unmittelbar in der Gremienarbeit eingesetzt und validiert werden. Für diese prototypischen Anwendungen wurde im Vorfeld vorwiegend das Gremium IEC/TC 65, UK 931.2 (vormals IEC/TC 65, UK 921.1) ausgewählt. Das DKE/UK 931.2 "Begriffe der Automatisierung" als deutsches Spiegelgremium der IEC/TC 65/WG 1 "Industrial-process measurement, Control and Automation - Terms and Definitions" erarbeitet Normen für die Festlegung einheitlicher technischer Begriffe für den Technologiebereich "Automatisierung", einschließlich dem IEV (IEC 60050-ff). Das Gremium erschien besonders geeignet für den Piloteneinsatz von Harbsafe2-Anwendungen, sowohl in Software-Piloten als auch mittels vorbereiteter Zwischenergebnisse in mit Hinweisen angereicherter und strukturierter Rohdatenform. Das Ziel war die Erarbeitung eines Neuentwurfs von Section 351 im IEV, sowie die Überprüfung von Begriffen der Automatisierung in Normen benachbarter technischer Gebiete auf Widerspruchsfreiheit. Leider konnte die Zusammenarbeit aufgrund von krankheitsbedingter Inaktivität des Gremiums nicht vertieft werden. Entsprechend konnten hier keine Ergebnisse generiert werden, die mittelbar in Form von Korrekturen und Eintragsentwürfen in die Normungsarbeit eingeflossen wären oder zur Publikation geeignet wären. Ebenso wurden inhaltliche Problemfelder nicht bei der Erstellung von Trainings- und Testdatensets in AP2 berücksichtigt. Pilotanwendungen laufen jedoch seit Projektende vor allem in IEC SC65E, WG2 und prospektiv IEC JPT3.

Zudem verblieb ein Call-for-Experts zur Etablierung eines Arbeitskreises für die eingehende Unterstützung und Begutachtung der Arbeit, initiiert von der DKE, ohne nennenswerte Rückläufer.

Ein weiteres Ergebnis aus AP1 war die von der TU koordinierte Korpusdefinition, die in Abstimmung mit DKE und Normungsexperten als sinnvoll erachtet wurde. In der ersten Hälfte des Projekts einigte man sich auf alle Begriffseinträge aus den "Terms and Definitions"-Abschnitten der aktuellen IEC-Normen in englischer Sprache, einschließlich des IEVs (IEC 60050). Die Korpusdefinition beinhaltet keine Entwurfsvorschläge und keine veralteten oder zurückgezogenen Normen. Die englische Sprache wurde gewählt, nachdem mit Normungsexperten abgestimmt wurde, dass die zu unterstützende Begriffsharmonisierung in der Praxis ausschließlich auf Englisch erfolgt und Übersetzungen von der englischen Fassung abgeleitet werden.

Des Weiteren wurden in der ersten Projekthälfte Anforderungen an die Darstellung von Begriffseinträgen abgestimmt. Hierbei konnte in großem Maße auf Vorarbeiten aus NISO-STS [15] und die Expertise bei der Normen-Verarbeitung seitens der DKE zurückgegriffen werden.

Eine eingehende Begutachtung der Ergebnisse aus Harbsafe 2 und dabei sowohl der Auswertungs-, als auch der Bedienkonzepte konnte durch Normungsexperten aus TC1 im September 2022 erfolgen. Die Ergebnisse der semantischen Modellierung, Klassifikation und Strukturierung aus **AP3** wurden als unabdingbar für die Zukunft der Harmonisierungsarbeit eingeschätzt und Potential für Weiterentwicklung identifiziert. Bedienkonzepte- und Präsentation aus **AP7** seien adäquat, übersichtlich und komfortabel. Die Ergebnisse aus AP4 zu den Qualitätsmetriken wurden als derzeit in der Normungspraxis nicht unmittelbar zielführend bewertet und in der Folge auch in der Darstellung (**AP7**) gering priorisiert. Die statistische Bewertung der Terminologien von Normen, Gremien und Komitees erscheint aufgrund der je unterschiedlichen inhaltlichen Voraussetzungen wenig vergleichbar. Langfristige Trends der Gesamtterminologie müssen in einem zukünftigen Produktivbetrieb eigens analysiert und aufbereitet werden.

Von Peter Bernard Ladkin (Causalis) wurde in Zusammenarbeit mit der TU Braunschweig die VDE SPEC „DKE Enhanced Glossary – Methods and Tools“ erarbeitet und von der DKE zur Veröffentlichung vorbereitet.

Arbeitspaket 1: Ergebnisse

1. Die im Projekt erarbeiteten exemplarischen Harmonisierungstasks und Analyseergebnisse wurden in diversen Gremien verbreitet.
2. Validierung durch Experten der in **AP3** erarbeiteten der Analyse- und Annotationsverfahren, Prüfkriterien und Strukturierungsmethoden. Validierung der Gebrauchstauglichkeit der in **AP7** erarbeiteten Bedienkonzepte.
3. Übersichten, zahlenmäßige Auswertungen und beispielhafte Analyseergebnisse wurden der Öffentlichkeitsarbeit zur Publikation zugeführt (**AP8**).
4. Entwurf der VDE SPEC „DKE Enhanced Glossary – Methods and Tools“.

Arbeitspaket 2: Zielstellung

AP-Nr. 2	Auswahl und Annotation von Terminologieelementen als Trainings- und Testdatensets
Zeitraum	Projektmonat 3 bis Projektmonat 12
AP-Leiter	ITL
	Ziele des Arbeitspaketes
	<p>Das Arbeitspaket dient der manuellen Goldstanderstellung für Training und Evaluierung des ML-Prädiktionsmodells. Dabei wird Spezifität und Sensitivität der entwickelten Verfahren gemessen und Leistungsfähigkeit dokumentiert. Zunächst werden für die diversen Analyseergebnisse des komplexen Modells eigens Evaluationsziele und Methoden definiert und priorisiert. Auf dieser Grundlage werden aus dem DKE/IEC-Terminologie-Datenbestand (ISO-STS, TBX, IEC Glossary) Datensets extrahiert und manuell annotiert. Niedrig-priorisierte können intern annotiert werden und dienen dem Training prototypischer Klassifikatoren. Hoch-priorisierte Prädiktionsmodelle, wie die Homonymerkennung, deren Ergebnisse zur Harmonisierung von Normungsexperten zu bearbeiten sind, unterliegen hohen Anforderungen an Spezifität und werden daher im Rahmen einer Erhebung unter Experten annotiert. Dabei wird ein repräsentatives Datensets identifiziert, Ratingexperten akquiriert und auf einer webbasierten Umfrageplattform für jedes Item annotiert, ob es diesem Harmonisierungstask unterliegt oder nicht. Die Umfrageplattform steht aus dem Vorgängerprojekt „Harbsafe“ zur Verfügung und wird geringfügig adaptiert. Bei der Entwicklung der Datensets kann ebenfalls auf Ergebnisse aus Harbsafe aufgesetzt werden, allerdings erscheint es unbedingt erforderlich diese auf diverse Domänen und Subdomänen auszuweiten, um Validität der Ergebnisse in anderen Fachbereichen zu sichern.</p> <p>Weder Umfrageplattform noch Datensets sind Bestandteil der Systemarchitektur der zu entwickelnden ML-Datenkomponente. Die Datensets werden lediglich zur Entwicklung, Training und Evaluation des integrierten Prädiktionsmodells eingesetzt.</p> <p>ITL extrahiert und bereinigt die Datensets mit teilweise vorhandenen Tools (Python/Anaconda) und Vorerfahrung aus dem Vorgängerprojekt „Harbsafe“ aus DKE/IEC mittels XML-Parsing, Webcrawlern oder Datenbankzugriff. Die Datensets werden in eine projektinterne Datenbank beim ITL abgelegt, zum Teil in die adaptierte Umfrageplattform des ITL importiert oder manuell annotiert und für Training und Evaluation (AP4) bereitgestellt.</p> <p>DKE stellt Terminologiedaten bereit, sofern nicht bereits im Web veröffentlicht, verifiziert kritische Annotationsergebnisse, hilft bei der Akquise von Ratingexperten und verbreitet entsprechende Anfragen.</p> <p>INOSOFT verfolgt die Entwicklungen und berät hinsichtlich verfügbarer Datenquellen.</p>

Output

1. Diverse annotierte Datensets für Training und Evaluation der ML-Prädiktionsmodelle (**→AP3**), bei Eignung ebenfalls zur Veröffentlichung im Kontext von Terminologiewissenschaft und Information Retrieval vorgesehen.

Arbeitspaket 2: Ablauf

Die DKE hat einen umfassenden Terminologiekorpus in strukturierter Form an die TU Braunschweig geliefert. Hieraus wurden Datensets gemäß projektierter Trainingsaufgaben in AP3 generiert.

Einschränkungen / Hinweise:

Auf die Erhebung von Expertenbeiträgen zu den Test- und Trainingssets wurde verzichtet, da der Call To Experts und die starke Integration mit Arbeiten von Normungsgremien nicht im Projekt darstellbar waren. Stattdessen wurde koordiniert von der Projektleitung die Ratingkriterien mit Experten abgestimmt und sämtliche Ratingtasks von Studierenden im Master „Kultur der technisch-wissenschaftlichen Welt“ durchgeführt.

In enger iterativer Verknüpfung mit AP3 wurden Datensets erhoben, bewertet und in AP3 zur Weiterentwicklung der Prädiktionsmodelle eingesetzt, sobald eine valide Anzahl an Datenexemplaren erreicht werden konnte. Nach Bedarf in AP3 wurden bedarfsgemäß neue Beispiele nachgeliefert, um weitere Anwendungsfälle und Hinwestypen zu generieren. Begonnen wurde mit den kritischen STS-Datensätzen für die semantischen Spektren. Hier wurden ca. 7.000 Ähnlichkeitspaare erzeugt und in AP3 für die hochprioritäre STS-Modell-Optimierung eingesetzt. Um das Ziel einer automatischen Strukturierung aller Homonyme abzubilden, wurden über alle Definitionen zu häufigen Begriffsbenennungen sämtliche Ähnlichkeitspaare gebildet und bewertet. Zur finalen Validierung wurde schließlich ein repräsentatives Sample an Ähnlichkeitspaaren bewertet, um den Schwellenwert bei der Homonymprädiktion zu bestätigen.

Für HintAn, das Verfahren zur automatischen Detektion von Anforderungsverstößen bei Definitionen, wurden zunächst allgemeine Typen von Verstößen aus den Anforderungen abgeleitet und auf Implementierbarkeit geprüft (siehe Tabelle 5: Hinwestypen zu detektierbaren Anforderungsverstößen.). Hierzu wurde vor allem auf ISO/IEC DIR 1:2019, Annex SK im Abgleich mit diversen Terminologienormen zurückgegriffen. Insgesamt konnten 16 unterschiedliche Hinwestypen identifiziert werden. Die Konzeption wurde von der TU Braunschweig durchgeführt und von der DKE geprüft.

Tabelle 5: Hinweistypen zu detektierbaren Anforderungsverstößen.

Title	Description
Definition Introduces Synonym or Hyperonym	The definition contains only a nominal phrase. Please, specify characteristics.
Contains Sentences	The definition contains full sentences. Please, consider using notes.
Definition Contains Items Which Should be in Notes	The definition contains unspecific, exemplary or usage-related informations. Please, consider using notes.
Definition is Comparatively Long	The length of the definition is part of the top 1% percentile of the longest definitions. Check if this is appropriate.
Definition Contains Superfluous Phrase	The Definition contains superfluous phrases like "in this specification", "according to this standard".
Definition Addresses Use but not Concept	The definition seems to define the usage of the term, rather than the concept. Consider writing an intensional definition starting with a hyperonym.
Definition indicates a Field of Use without Angle Brackets	The definition seems to start with a field of use. Consider adding angular brackets.
Constructional Definition	The definition focusses on constructional aspects. Consider writing a functional definition.
Term cannot be Replaced by the Definition (Violates Substitutability Principle)	The definition seems not be formulated in a way that allows its replacement for the term.
Definition Starts with a Determiner	The definition shall not start with a determiner.
Too many Definitions	There seem to be multiple definitions. Consider choosing one that is valid throughout the document.
Definition Contains a Requirement	The definition seems to contain a requirement. Avoid using "shall" and "must".
No Definition	A verbal definition seems to be missing or may be inappropriately stated only as a formula.
Cyclical Definition	The term shall not be repeated in the definition.
Definition Terminates with Full Stop	To allow substitutability, the definition should not terminate with a full stop.
Definition Starts with Upper-Case Letter	To allow substitutability, the definition should not start with an upper-case letter.

Im Anschluss an die Hinweistypen-Identifikation wurde ein Korpus zufällig ausgewählt und nach Zutreffen eines Hinweistyps bewertet und erweitert, bis für jeden Hinweistypen mehr als 20 Beispiele gefunden werden konnten. Dabei wurden insgesamt 3.300 Beispiele manuell bewertet. Für die Bewertung war die TU Braunschweig zuständig.

Einschränkungen/Hinweise

Die Erhebung von Normungsanforderungen an die Formulierung von Definitionen wurde maßgeblich anhand ISO/IEC Directives, Part 1:2019, Annex SK durchgeführt, die verbindliche Direktiven an die Normungserstellung vorgibt. Der Annex SK wurde jedoch in 2022 durch eine Revision der Direktiven abgelöst. Die neuen Direktiven

finden sich in ISO/IEC DIR1:2022, Annex SL, Appendix 3. Diese sind augenscheinlich kompatibel, konnten aber nicht im Detail geprüft werden.

Parallel wurden ein Datenset für das SemAn-„Pretty-Print“-Format erstellt. Hier wurde im Rahmen von Workshops zwischen Vertretern der TU und Causalism explorativ ein konsistentes Format iterativ entworfen und erprobt, das möglichst viele Eigenschaften aus der sortalen Logik aus dem spezifischen Set grammatischer Phänomene in empirischen Definitionen herleitet und textuell darstellt. Aufgrund des hohen Aufwands bei der logischen Erfassung ist das manuelle Datenset eher klein geblieben. Die hohe erreichte Genauigkeit der resultierenden Implementierung lässt sich daher nicht quantitativ generalisieren. Die Pretty-Print-Darstellung verfügt über einen heuristischen Wert, bei der Optimierung der Lesbar- und Konzeptualisierbarkeit beim Nutzer und muss sich daher im Gebrauch beweisen.

Die Datensets konnten im Ablauf rechtzeitig fertiggestellt werden, wurden jedoch bis Projektabschluss punktuell iterativ ergänzt und korrigiert. Die Veröffentlichung im Kontext von Terminologiewissenschaft und Information Retrieval wird eruiert.

Arbeitspaket 2: Ergebnisse

Zu den Datensets gehören:

- 1) ein **vertikales STS-Datenset** (für Evaluation und Training von Semantic Spectrums) mit allen Definitionen aus dem Normungskorpus zu 39 Begriffsbenennungen mit mehr einer Definition. Die Definitionen wurden pro Begriffsbenennung nach Phrasenäquivalenzen gruppiert und Gruppenähnlichkeiten auf einer fünfstufigen Likert-Skala nach begrifflicher Ähnlichkeit bewertet. Hieraus resultieren ca. 9.000 Ähnlichkeitswerte von Definition zu Definition.
- 2) ein **horizontales STS-Datenset** mit zufälligen Definitionsvergleichen. Aus der Menge aller Definitionsvergleiche von Definitionen mit derselben Begriffsbenennung wurden zufällig ca. 950 Exemplare ausgewählt und auf einer definierten Likert-Skala nach begrifflicher Ähnlichkeit bewertet.
- 3) ein **HintAn-Datenset** mit 3.300 Begriffseinträgen, die nach Anforderungsverstößen an die Formulierung von Definitionen geprüft wurden.
- 4) ein kleineres **SemAn-Datenset** mit knapp 100 Definitionen, das zur Erarbeitung des SemAn-Pretty-Print-Formats diente. Hierbei wurden manuell Pretty-Print-Präsentationen erstellt, um diese in AP3 programmatisch zu reproduzieren.

Arbeitspaket 3: Zielstellung

AP-Nr. 3	Modellierung und Klassifikation der Terminologiedaten mit Machine-Learning-Verfahren
Zeitraum	Projektmonat 4 bis Projektmonat 20
AP-Leiter	ITL
	Ziele des Arbeitspaketes
	<p>Es werden datenwissenschaftliche Modelle entwickelt, die Harmonisierungsfälle und kontextuelle semantische Zusammenhänge in großen Terminologiekorpora prädizierbar machen. Dazu werden die semantischen Inhalte von terminologischen Einträgen mittels Embeddingverfahren, sowie weiteren, aus Metadaten abgeleiteten, Eigenschaften einer Repräsentation durch quantitative Features zugeführt. Ein ML-Modellierungsprozess in vier Phasen (Feature construction and weighting, Feature selection and projection, Classifier training, Evaluation) wird für jeden Harmonisierungstask iterativ durchlaufen, wobei Inputs, Parameter und Verfahren systematisch variiert und mit den Testdatensets (AP2) abgeglichen werden. Ein neuer Ansatz von Harbsafe II besteht zudem in der Automatisierung der semantischen Analyse (SemAn) als vorbereitende Strukturierungsmaßnahme zur manuellen konzeptuellen Analyse (ConcAn): die Dekomposition von Definitionen in atomare semantische Elemente und Matching von bedeutungsgleichen Varianten zu abstrakten Bedeutungsfeatures.</p> <p>ITL entwickelt Prädiktionsmodelle mittels öffentlich zugänglicher semantischer Features, aus einem Transferlernen-Ansatz mit neuronalen Netzen (sog. <i>word embeddings</i>). Die Features werden adäquat für terminologische Einträge verarbeitet und dienen als Input unterschiedlicher Klassifikatoren und Clustering-Verfahren, die vom ITL angepasst und optimiert werden. Zur Validierung der Verfahren werden Pilotanwendungen in einzelnen Fachbereichen abgegrenzt und Ergebnisse begutachtet. Die flexiblen, prototypischen Datenmodelle demonstrieren die Funktionalität der zu entwickelnden Softwareassistenz vollständig, ohne zunächst andere softwaretechnischen Eigenschaften (wie Performanz, Modularität und Robustheit) zu erfüllen, deren Implementation in AP5 erfolgt.</p> <p>DKE nimmt laufend an Präsentationen des Modells teil und bewertet die Ergebnisse hinsichtlich Qualität und Integrierbarkeit in technische und institutionelle Prozesse und auf Eignung für die Gremienpräsentation.</p>

Output

1. Diverse evaluierte ML-Prädiktionsmodelle aus Datenanalysen zur Implementation in einen Server-gebundenen Datenservice (**→AP5**).
2. Geeignete Ergebnisse der Pilotanwendungen werden an die Gremienarbeit weitergeleitet und für Workshops vorgesehen (**→AP1**).

Arbeitspaket 3: Ablauf

Begonnen wurde mit der Modellierung der Bedeutungsspektren auf Basis von neuronalen Netzen mit einem vortrainierten Transformer-/Encoder-Modell. Die Bedeutungsspektren basieren auf einem STS-Verfahren, das die inhaltliche Ähnlichkeit zweier Texte (hier: Definitionen) quantifiziert. Für die Auswahl des optimalen STS-Modells wurde auf Basis des vertikalen Trainings- und Evaluierungssets (AP3) eine Pipeline aufgebaut, die einen direkten Vergleich der Performanz sinnvoller STS-Modelle implementiert. Hierbei wurden Baseline-Metriken, wie der Jaccard-Similarity Index, mit älteren neuronalen Embedding-Verfahren, wie word2vec und state-of-the-art Transformer-Architekturen, wie Sentence-BERT [2] verglichen. Letztere demonstrieren signifikante Verbesserung in der Korrelation mit den manuellen Bewertungen gegenüber älteren Ansätzen und wurden in der Folge eingesetzt. Optimale Performanz konnte mit Sentence-BERT model [16] auf Basis von MiniLM [17] erreicht werden. Im Anschluss wurde versucht, die Performance des ausgewählten Modells, das auf allgemeinsprachlichen und nicht auf die elektrotechnische Domäne spezialisierte, neuronale Modell durch Fine-Tuning weiter zu verbessern. Dies führte allerdings nicht zu den gewünschten Verbesserungen.

Nach Optimierung der STS-Performance, wurden die Clusterung der Bedeutungsspektren anhand der Sentence-BERT-Embeddings zu allen Definitionen einer Begriffsbenennung durchgeführt. Die hierarchisch-agglomerative Clusterung resultierte in einer sinnvollen Strukturierung und Reihenfolge der Definitionen (siehe Abbildung 2). Für die übersichtliche Darstellung in der Web-Applikation wurde nun ein Schwellenwert für die Cluster-Distanz gesucht, der als sinnvolles Kennzeichen für distinkte oder nur partiell übereinstimmende Begriffsbedeutungen hinwies. Dieser Parameter wurde mittels des horizontalen Dateensets (AP2) optimiert.

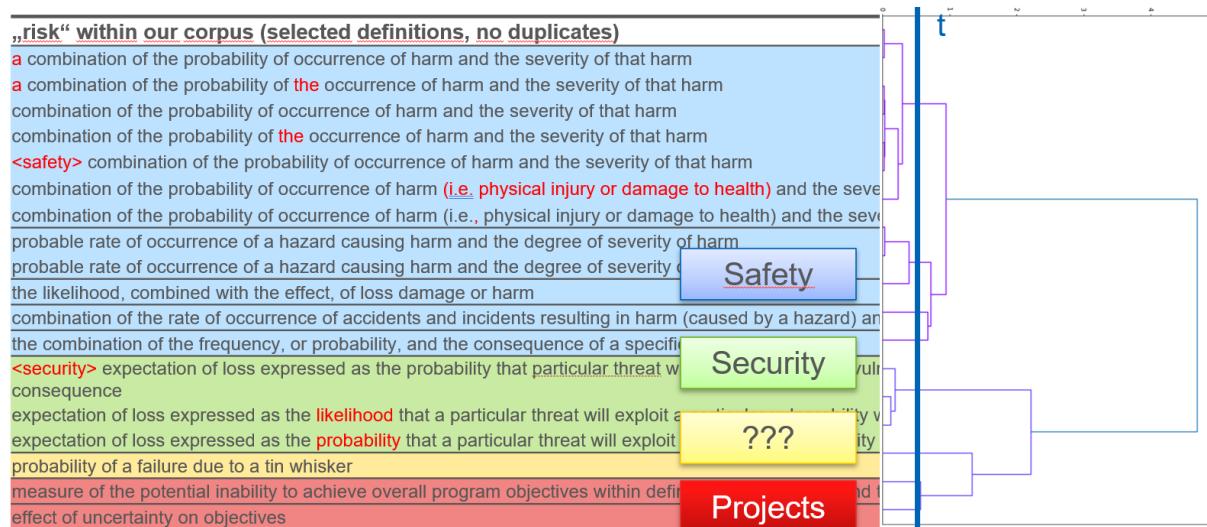


Abbildung 2 Ein geclustertes Bedeutungsspektrum mit beispielhaft vergebenen Cluster-Labels für einen sinnvollen Schwellenwert t .

Im Anschluss stand die Entwicklung und Evaluierung der harmonisierungsrelevanten Annotationen im Vordergrund. Für die HintAn-Konformitätsprüfung wurde regelbasiertes Matching mittels spaCy [1] angewandt. Hierzu wurden Indikationen und Kontraindikationen für Anforderungsverstöße aus den manuell bewerteten Konformitätsset identifiziert, in

grammatische, lexikalische und Zeichen-basierte Regeln implementiert und experimentell auf das Validierungsset angewandt. Dabei konnte Sensitivität und Spezifität der Detektoren iterativ verbessert werden. Beide Verfahren wurden an der TU Braunschweig entwickelt und im Projektkonsortium abgestimmt. Die Anzahl der prädizierten Hinweise ist überraschend hoch ausgefallen: Insgesamt wurden über 19.000 Hinweise für insgesamt 65.000 Einträge generiert. Die Genauigkeit der Vorhersage liegt (je nach Hinweistyp) bei 70-99%.

Parallel wurde das im Zusammenspiel mit AP2 gemeinsam von TU Braunschweig und Causalism entwickelte SemAn-Pretty-Print-Format von Causalism umgesetzt. Hierbei kamen reglbasierte Ansätze mittels spaCy [1] zum Einsatz. Gemessen am überschaubaren Evaluationsset konnte eine fehlerfreie programmatische Reproduktion von über 90% erreicht werden. Darunter fielen diverse geringfügige Fehlinterpretationen, die den subjektiven Übersichtswert der logisch-hierarchischen Darstellung nicht völlig negieren. Auf eine objektivierende Evaluation des subjektiven Mehrwerts beim bewussten Nachvollziehen der Definitionen wurde in diesem Projekt jedoch verzichtet.

Arbeitspaket 3: Ergebnisse

Die Ergebnisse bestehen aus prototypischen Implementierungen der erarbeiteten Tools und Verfahren, einschließlich Evaluation für: Semantic Spectrums (Sammlung und Strukturierung aller Definitionsvarianten durch semantische Sortierung und Gruppierung aller Einträge zu einer Benennung), HintAn (konkrete Harmonisierungshinweise bezüglich der Konformität der Definitionen gemäß IEC Direktiven und Terminologienormen), SemAn (eine Analyse der Definitionen in Bedeutungskomponenten und ihre hierarchische Darstellung).

Arbeitspaket 4: Zielstellung

AP-Nr. 4	Globale Konsistenzmetriken entwickeln
Zeitraum	Projektmonat 15 bis Projektmonat 24
AP-Leiter	ITL
	<p>Ziele des Arbeitspaketes</p> <p>Um die Qualität der Terminologie steuerbar zu machen, müssen Qualitätsmetriken erarbeitet werden, die den Harmonisierungsstand und -fortschritt und die entsprechenden Qualitätsziele der Messbarkeit zuführen. Hierbei ist zu zeigen und zu dokumentieren, welche messbaren Eigenschaften des Terminologiebestandes sich durch automatisch generierte Maßnahmenvorschläge im Sinne einer Harmonisierung beeinflussen lassen. Eingehen könnte neben einer Dublettenrate zum Beispiel ein Maß für die semantische Divergenz von Definitionen derselben Benennung (Mehrfachdefinitionen), Maße zur Bennennungskonsistenz und diverse weitere.</p> <p>ITL entwickelt statistische Verfahren zur Erhebung zentraler Kennzahlen für die Beschreibung des globalen Harmonisierungszustandes der DKE-Terminologie. Die Entwicklung erfolgt auf Grundlage der dem ITL vorliegenden Terminologiedaten und dem in AP3 erstellten Prädiktionsmodell. Das Verfahren wird dokumentiert, in Python/Anaconda implementiert und für die Integration in das finale Einsatzszenario im DKE Serversystem vorbereitet.</p> <p>DKE bewertet die Ergebnisse auf Eignung zur Überwachung der Harmonisierung der DKE-Normungsterminologie und zur Veröffentlichung in geeigneten Publikationen.</p>

Output

3. Ein adäquates, zur Veröffentlichung geeignetes statistisches Verfahren zur Erhebung von Harmonisierungs-relevanten Qualitätsmetriken für große Terminologiekorpora (**→AP8**).
4. Eine Implementation des Verfahrens zur Berechnung der Metriken für die DKE-Terminologiedaten zur Integration im zu entwickelnden ML-Softwaresystem (**→AP5**).

Arbeitspaket 4: Ablauf

Die konzeptionellen Arbeiten in AP 4 begannen mit ersten konsolidierten Arbeitsergebnissen aus AP3 in der zweiten Projekthälfte. Methodisch wurde versucht, aus Zählung- und Aggregation der präzisierten harmonisierungs-relevanten Annotationen aussagekräftige Qualitätsmetriken zu entwickeln. Dabei konnten Maße definiert werden, die vergleichende Aussagen über den Harmonisierungszustand in unterschiedlichen Normungsbereichen erlauben, jedoch keine vollständige Abbildung der komplexen semantischen Zusammenhänge und Inkonsistenzen darstellen. Bei der Evaluation durch Vertreter des für die Terminologie-Koordination zuständigen IEC TC 1, äußerten die Experten derzeit kein Interesse an diesem Steuerungstool. Der entscheidende Mehrwert läge in der Unterstützung der Arbeit, nicht in der Bewertung bei teilweise drastisch unterschiedlichen Ausgangsvoraussetzungen und möglicher Zweifel an der Aussagekraft in der

Normungscommunity. Die Publikation der Maße in vergleichender Form über das Enhanced Glossary wurde daher nicht empfohlen und ist nicht erfolgt.

Arbeitspaket 4: Ergebnisse

Erarbeitet wurden folgende Qualitätsmetriken (Tabelle 6: Qualitätsmetriken.).

Tabelle 6: Qualitätsmetriken.

A) Global / pro TC / pro Normungsdokument
- Hinweisrate bei den Einträgen (total/pro Hinweistyp)
Global / pro TC / pro Benennung
- Anzahl und Anteil der Homonyme an den Benennungen (= Benennungen mit mehreren Bedeutungen)
- Anzahl und Anteil der wiederverwendeten Definitionen bei Term-gleichen Einträgen
Global, experimentell:
- mittlere STS-Ähnlichkeit der Definitionen von Term-gleicher Einträgen

Die Metriken wurden Datenbank-seitig weitgehend umgesetzt und können administrativ eingesehen werden, sind aber derzeit nicht zur Veröffentlichung vorgesehen.

Arbeitspaket 5: Zielstellung

AP-Nr. 5	Entwicklung der Datenservices
Zeitraum	Projektmonat 9 bis Projektmonat 17
AP-Leiter	ITL
	Ziele des Arbeitspaketes
	<p>Die datenwissenschaftlichen Modelle und Funktionen aus AP3 werden im Rahmen einer ML-Pipeline ausentwickelt und integriert. Das resultierende Softwaresystem („Datenkomponente“) erfüllt Anforderungen an Performanz, Robustheit, Fehlertoleranz, lose Kopplung und Nachrichten-basierte, asynchrone Kommunikation zu anderen Diensten. Ziel ist eine modulare Architektur aus wiederverwendbaren Services mit definierten Datenschnittstellen als auslieferbare Softwarekomponente (z.B. als Docker) zur Integration ins Backendsystem (AP6). Der zentrale Dienst kapselt das ML-Prädiktionsmodell und seine diversen Funktionen und legt die prädizierten Relationen und Harmonisierungstasks (AP3) in einer Datenbank zum Zugriff für AP7 ab. Daneben wird ein Embedding-Service, der zu Eingabetexten eine semantische Vektorrepräsentation liefert, für die einheitliche Wiederverwendung von semantischen Textvergleichsdaten im DKE-Serversystem angeboten. Zudem ist ein TermEntry-Service zu entwickeln, der aus dem Altbestand aus Normdokumenten Begriffseinträge extrahiert und strukturiert. (Siehe Anhang 3.)</p> <p>ITL entwickelt auf Grundlage der datenwissenschaftlichen Klassifikatoren eine modulare Softwarekomponente, welche die in AP3 demonstrierten klassifikatorischen, relational strukturierenden und gruppierenden Funktionen über eine performante und robuste ML-Pipeline implementiert. Ferner stimmt das ITL Anforderungen ab, die die Softwarekomponente an die Serverumgebung stellt. Eine Datenbank wird entworfen, abgestimmt und dokumentiert, die die Analyseergebnisse aufnimmt und die Datengrundlage für das Web-System der Ausgabekomponente bildet.</p> <p>DKE stellt Anforderungen an Datenschnittstellen und Softwareintegration der zu entwickelnden Serverkomponente und unterstützt die Entwicklungsarbeit durch Support und Dokumentationen der Serverumgebung.</p> <p>INOSOFT begleitet die Entwicklungen und berät hinsichtlich Eignung zur Softwareintegration.</p>

Output

5. Modulare Softwarekomponente (Services) (**→AP6**)
6. Datenbank mit vorläufigen Prädiktionsergebnissen zur Entwicklung der Webkomponente (**→AP7**)
7. Dokumentation des finalen Datenschemas der Ergebnisdatenbank (**→AP7**).
8. Anforderungen der Softwarekomponente an die Serverumgebung (**→AP6**)

Arbeitspaket 5: Ablauf

Der Verlauf von AP5 gestaltete sich planmäßig und fand unter der Zuständigkeit der Technischen Universität Braunschweig statt. Die im Rahmen der NLP-Modellentwicklung aus AP3 hervorgegangenen Softwarekomponenten wurden weiterentwickelt, um ein angemessenes Reifegradniveau der prädiktiven Funktionen zu erreichen. Anschließend wurden sie technisch aufbereitet, um sie zur Integration in die Server-Umgebung (AP6) vorzubereiten. Alle Softwarekomponenten aus AP5 wurden an der Technischen Universität Braunschweig entwickelt, mit Ausnahme der Funktion zur SemAn-Generierung. Letztere wurde von Causalismus entwickelt und anschließend von der TU Braunschweig in die Service-Architektur eingebunden.

Basierend auf den Erfahrungen aus dem Modelleinsatz und dem Ressourcenverbrauch bei der Nutzung von NLP konnten Anforderungen für die Verarbeitung in der Server-Umgebung (AP6) abgeleitet und weitergegeben werden. Dabei wurde, wie vorgesehen, ein modularer Flask-Web-Service innerhalb einer Docker-Containerumgebung implementiert und bereitgestellt. Der Komponentenentwurf der einzelnen Softwarepakete im Projekt ermöglichte eine entwicklungsfreundliche Umgebung mit geringen Abhängigkeiten und klar definierten Schnittstellen. Die Arbeiten konnten innerhalb des zeitkritischen Rahmens abgeschlossen werden, und sämtliche Detailfragen, die während der Projektarbeit auftauchten, wurden bis zum Projektende softwaretechnisch behandelt.

Um die Entwicklung der Frontend-Web-Anwendung (AP7) gezielt zu unterstützen und auszurichten, lag der Fokus frühzeitig auf dem Zieldatenbank-Schema und den Integritätsbedingungen. Ein Datenbankentwurf wurde abgestimmt, dokumentiert und in PostgreSQL implementiert. Dank der Erfahrungen im Terminologie-Management erwies sich der Datenbankentwurf als erfolgreich und blieb im Verlauf des Projekts weitestgehend stabil.

Arbeitspaket 5: Ergebnisse

Das von der TU Braunschweig entwickelte System dient der Analyse und Anreicherung von terminologischen Daten mit den Harmonisierungs-relevanten Annotationen, die im Projekt Harbsafe 2, AP3 erarbeitet wurden: **Semantic Spectrums** (Sammlung und Strukturierung aller Definitionsvarianten durch semantische Sortierung und Gruppierung aller Einträge zu einer Benennung), **HintAn** (konkrete Harmonisierungshinweise bezüglich der Konformität der Definitionen gemäß IEC Direktiven und Terminologienormen) und **SemAn** (eine Analyse der Definitionen in Bedeutungskomponenten und ihre hierarchische Darstellung).

Die Implementierung besteht in einem flexibel einsetzbaren Web-Service, im Folgenden „TermAnalyzer“ genannt. Dieser wurde mit Python 3.0 implementiert und umfasst folgende statische und programmatische Komponenten:

1. OpenAPI 3.0 REST-Schnittstelle
 - Selbstdokumentierendes interaktives Web-Interface
 - Python Flask Service-Implementation
2. Datenquellenkonfiguration
3. SQL-Zieldatenspezifikation
4. Datenpipeline
 - Konfigurierbare Datenselektion

- Homonymidentifikation
- Semantic-Spectrum-Analyse
- HintAn-Prädiktion
- SemAn-Pretty-Print
- Serialisierter Output

5. Datenbank-Synchronisation

Der Service nimmt terminologische Daten im spezifizierten Format entgegen, annotiert diese und spielt die annotierten Daten zurück an den Service-Nutzer. Alternativ kann eine Synchronisationsroutine ausgelöst werden, die terminologische Daten aus verschiedenen XML-basierten Datenquellen abruft, annotiert und in einer PostgreSQL-Datenbank ablegt. Die einsetzbaren Datenquellen administrativ konfigurierbar und können per Datei- oder Webzugriff eingesetzt werden.

1. OpenAPI 3.0 REST-Schnittstelle

Bei der Definition der Datenschnittstelle wurde auf den OpenAPI 3.0-Standard zurückgegriffen. Dieser dient dazu, um Anwendungsprogrammierschnittstellen oder APIs zu beschreiben. Die OpenAPI integriert Spezifikationen und Benutzerdokumentation der Serviceendpunkte und Outputdefinition. Auf der OpenAPI-Definition basieren einerseits die Implementationen der Service-seitigen Endpunkte und ein dokumentiertes, interaktives Web-Interface für flexiblen, transparenten Einsatz des Services. Die Service-Endpunkte wurden mit Python 3.6 implementiert und ermöglichen Zugriff auf alle Funktionen der Datenpipeline. Zu analysierende Begriffe können im per OpenAPI 3.0 definierten JSON-Format zur Analyse übergeben werden und mit den statischen Norm-Inhalten, definiert in der Datenquellendefinition abgeglichen werden.

Tabelle 7: Service-Endpunkte des TermAnalyzers

Endpoint	Function	Description
Corpus Comparison and Input Analysation		
POST/analyze	performs all annotations while comparing the definitions to the corpus	Checks all submitted TermEntries for definition quality, compares them with the current IEV (IEC 60050) and the "Clause 3"-Glossary-Corpus and returns them annotated with hints, relational hints and SemAn-Pretty-Print. If you want to restrict the comparisons within a technical committee use the "tc"-property, otherwise leave it empty.
Excel In / Excel Out		
POST /annotateExcelGlossary/	analyses and annotates Excel glossary	Checks the submitted Excel glossary for definition quality, compares them with the current IEV (IEC 60050) and the "Clause 3"-Glossary-Corpus and returns them annotated with hints, relational hints and SemAn-Pretty-Print. Within the Excel sheet certain columns are expected, these are "terms" (comma-separated), "definition" and "tc". Those column names correspond to the values in the first row. If you want to restrict the comparisons within a technical committee per entry use the "tc" property, otherwise leave it empty.

Endpoint	Function	Description
Input Analysat		
POST/annotateHints	annotates all submitted Entries with hints regarding definition quality	Checks all submitted TermEntries for definition quality, adds Hints to each TermEntry and returns them back
POST /annotateSemAn	annotates all submitted Entries with a SemAn-Pretty-Print	Returns all submitted Entries with a generated SemAn-Pretty-Print for each definition
POST /clusterSemanticSpectrums	returns all submitted Entries grouped and sorted by their designation and definition similarity	Uses all preferred Terms from the submitted TermEntries to produce all SemanticSpectrums for this body of terminology. A Semantic Spectrum orders and clusters all the TermEntry for each Term in an intuitively meaningful way.
Corpus Query		
GET /getSemanticSpectrumForTerm/{term}	finds all definitions clustered by meaning	queries all configured sources (IEV/IEC 60050 and the "Clause 3"-Glossary-Corpus) for homonyms of the input term and returns all entries for that term clustered and sorted by meaning.
Database Synchronization		
GET/synchronize	performs an update of the result database	synchronizes all configured sources with the result database (IEV/IEC 60050 and the "Clause 3"-Glossary-Corpus) for updating the frontend application. (The current results database will be updated, previous results will be archived.)

2. Datenquellenkonfiguration

Die Datenquellenkonfiguration ermöglicht die Integration verschiedener XML-Terminologie-Datenquellen in die Synchronisationsroutine (GET/synchronize). Sie dient ferner der Durchführung von Abfragen über dem Datenbestand (GET/getSemanticSpectrumForTerm/{term}). Datenquellen werden per URI konfiguriert. Die terminologischen Einträge, sowie darunter liegende Datenfelder werden per XML-Selector (XPath) konfiguriert. Die Konfiguration erfolgt administrativ mittels JSON. Ferner kann für die Persistierung der Ergebnisdaten eine PostgreSQL-Datenbank konfiguriert werden.

3. SQL-Zieldatenspezifikation

Die SQL-Zieldatenspezifikation umfasst ein Datenschema, das sämtliche Inputdaten annotiert mit den Ergebnissen aus HintAn, SemAn und Semantic Spectrums erfassen und ablegen kann. Die Zieldatenspezifikation ist in PostgreSQL-Format verfasst und erlaubt die

vollständige Erzeugung einer kompatiblen Datenbank mit den erforderlichen Integritätsbedingungen. Relevante Datenfelder, wie Benennung und Definition, werden in Normalform angelegt. Quellformate bleiben erhalten. Teil der Zieldatenspezifikation in PostgreSQL sind gespeicherte Prozeduren zur Population der Datenbank. Diese ermöglichen die Datenbank-seitige objektrelationale Abbildung und entkoppeln das Datenmodell des TermAnalyzers von der physischen Persistierung im Datenbankschema.

4. Datenpipeline

Die Datenpipeline wurde modular mit Python 3.6 implementiert und generiert je nach Service-Endpunkt harmonisierungsrelevante Annotation für den Service-Output oder zur Synchronisation mit der Datenbank.

- **Konfigurierbare Datenselektion**
Infofern für die Funktion des Service-Endpunkts erforderlich, werden gemäß Datenquellenkonfiguration Terminologie-Ressourcen geladen, geparsst und Harmonisierungs-relevante Datenelemente (Benennungen, Definitionen) extrahiert.
- **Homonymidentifikation**
Terminologische Einträge mit identischer Benennung werden gruppiert.
- **Semantic-Spectrum-Analyse**
Sämtliche Definitionen jeder Gruppe werden mittels Sentence-BERT, einem Transformer-basierten vortrainierten neuronalen Modell für textuelle semantische Ähnlichkeit, eingebettet und die resultierenden semantischen Vektoren in einem hierarchisch-agglomerativen Clustering strukturiert. Wird ein bestimmter empirisch erarbeiteter Schwellenwert im Distanzmaß zwischen zwei Clustern überschritten, werden diese als Begriffseinträge mit unterschiedlicher Bedeutung aufgefasst und umgekehrt sämtliche Begriffseinträge in hinreichend ähnlichen Clustern als begrifflich identisch. Letztere werden als semantische Gruppe zusammengefasst. Durch Sortierung in der Reihenfolge des kürzesten Wegs durch den semantischen Raum werden die semantischen Gruppen in eine sinnvolle Reihenfolge gebracht und optimieren die Übersicht über alle Definitionen zu einer Benennung.
- **HintAn-Prädiktion**
Die HintAn-Prädiktion generiert Hinweise beim Verstoß gegen Normungsanforderungen in der Formulierung von Definitionen. Verwendet wird das verbreitete Python-Framework spaCy [1] in einer parallelisierten Pipeline mit den Komponenten: Satzsegmentation, Tokenisierung, Part-of-Speech-Tagging und Dependency-Parsing. Um die Pipeline auf Abweichung bei der Definition von Begriffseinträgen in der Normung gegenüber allgemeinsprachlich vortrainierten Modellen anzupassen, wurde die Satzsegmentation Regel-basiert angepasst. Ferner wurde ein regelbasiertes Entity-Recognition verwendet, um Normungsbegriffe in Definitionen wieder zu finden (etwa um zyklische Definitionen zu erkennen). Die mittels spaCy grammatisch angereicherten Definitionen werden anschließend durch ein modular erweiterbare Set an Detektoren analysiert, die jeweils einen Typ üblicher Anforderungsverstöße mit gewisser Genauigkeit erkennen und einen entsprechenden Hinweis generieren. Die Detektoren erheben keyword-basiert und aufgrund von

grammatischen Eigenschaften, Indikationen und Kontraindikationen für den Anforderungsverstoß. Prädizierbar sind so Verstöße gegen die folgenden Anforderungen an Definitionen. Die Anforderungen sind abgeleitet ISO/IEC DIR 1:2019, Annex SK im Abgleich mit diversen Terminologienormen:

- The definition shall not start with a determiner.
- The definition shall be formulated in a way that allows replacement for the term.
- The definition shall not contain additional information that may be better suited for the notes.
- The definition shall not start with a field of use without brackets.
- The definition shall not contain superfluous phrases.
- The definition shall not focus on constructional aspects.
- The entry shall contain a verbal definition.
- The term shall not be repeated in the definition.
- The entry shall not contain multiple definitions.
- The definition shall not define the usage of the term, rather than the concept.
- The definition shall not contain a requirement.
- The definition shall not introduce a synonym.

- **SemAn-Pretty-Print**

Die Ergebnisse der spaCy-Pipeline [1] werden außerdem verwendet, um mittels Dependency-Parsing die logische Struktur einer Definition in semantische Komponenten zu zerlegen und ihre Abhängigkeiten darzustellen. Das Ergebnis ist ein rein Text-basiertes hierarchisches Format mit einfacher logischer Syntax. [18]

- **Serialisierter Output**

Ergebnisse der Datenpipeline werden serialisiert. Das Transferformat ist JSON.

5. Datenbank-Synchronisation

Die verarbeiteten Ergebnisse, i.e. JSON-serialisierten Quelldaten zusammen mit prädizierten harmonisierungsrelevanten Annotationen, werden bei der Synchronisation (GET/synchronize) in einer PostgreSQL-Datenbankpersistiert. Um ältere Datenstände zu erhalten und das Rollback zu ermöglichen, wird die bestehende Datenbank datiert und archiviert. Das Schema für die neuen Ergebnisse wird gemäß der Zieldatenspezifikation neu angelegt und die serialisierten Ergebnisse werden an gespeicherte Prozeduren übermittelt, die die objektrelationale Abbildung durchführen. Die resultierende Datenbank enthält sämtliche Ergebnisse und ist als Datenbank-Image oder nativ geeignet für den Einsatz im Enhanced-Glossary oder weiteren zukünftigen Anwendungen.

Arbeitspaket 6: Zielstellung

AP-Nr. 6	Technische Integration der Komponenten (Leitung: DKE)
Zeitraum	Projektmonat 13 bis Projektmonat 24
AP-Leiter	DKE
	Ziele des Arbeitspaketes
	<p>Die Ergebnisse aus den vorherigen Arbeitspaketen (AP3, AP5) werden in Form von Algorithmen, Funktionen und Subsystemen in ein Backend-System integriert. Das Backend-System stellt das Gesamtsystem ohne Ausgabekomponente (AP7) dar und ist Bestandteil der IT-Umgebung der DKE. Das Backend-System muss den sicheren Datenzugriff auf den Normen- und Terminologiebestand der DKE gewährleisten, eine Umgebung für die Ausführung von ML-Algorithmen zur Verfügung stellen, sowie einen überwachten und kontrollierten Zugriff nach außen realisieren, welcher z.B. von der zu entwickelnden Webapplikation in AP7 verwendet wird. Für die Konzeptionierung und Umsetzung, insbesondere für die Integration der ML-Komponenten, wird die DKE unter Berücksichtigung möglicher Synergieeffekte (z.B. zu dem Projekt „DiTraNo“) auf entsprechend qualifizierte Unterauftragnehmer zurückgreifen. Das resultierende System wird nachhaltig in die DKE-Systemlandschaft integriert.</p> <p>DKE stellt die Serverumgebung für die Backend-Softwarekomponente unter Berücksichtigung der Anforderungen aus AP5 bereit. Die DKE führt Integration und Deployment des Backendsystems durch und definiert die Datenbankschnittstelle zum Frontend-System (AP7). Die Integration des Frontend-Systems wird abgestimmt. Die DKE erstellt einen Plan zur anschließenden Nutzung und Pflege des Softwaresystems.</p> <p>ITL begleitet die Softwareintegration und Deployment, nimmt resultierende Anforderungen für die Integration der ML-Komponente auf und stimmt Anforderungen ab.</p> <p>INOSOFT begleitet, berät und unterstützt Softwareintegration und Deployment, nimmt resultierende Anforderungen für die Integration der Ausgabekomponente auf und stimmt Anforderungen ab.</p>

Output

1. Einsatzfähiges ML-System zur Unterstützung der Terminologieharmonisierung. Die ML-Backend-Komponente ist in die DKE-Serverumgebung integriert, Echtzeit-fähig mit aktuellem Datenzugriff und zugreifbarem Datenbankoutput. Die Integration der Frontend-Terminologierecherche und -Harmonisierungsplattform ist vollzogen und bereit für den Produktivbetrieb.
2. Ein Plan für den Produktivbetrieb wurde erstellt.

Arbeitspaket 6: Ablauf

Unter dem Namen DKE Content as a Service (DKE CaaS) stellt die DKE einen Dienst bereit, mittels dem ein XML-basierter Normenbestand (sowie ggf. unterstützendes Datenmaterial, wie z.B. Inhalte des Internationalen Elektrotechnischen Wörterbuchs, kurz IEV) bedarfsgerecht (im Sinne der Anwendung) an einen Client ausgespielt werden kann.

Die DKE nutzt hierzu das Open Source XML-Datenbankmanagementsystem BaseX [19], das zur Speicherung und Abfrage von Inhalten größtenteils auf gängige W3C-konforme Web-Technologien, wie z.B. XML, XPath und XQuery, baut und mit dessen Hilfe RESTful Webservices zur bedarfsgerechten Ausspielung von Inhalten einfach entwickelt werden können.

Das Harbsafe2-Projekt nutzt das DKE CaaS, um auf die Terminologie-Daten aus den Normen zuzugreifen. Hierzu wird über das DKE CaaS ein Webservice bereitgestellt, der über hierarchische REST-URIs sowohl Metadaten zur gesamten Norm als auch einzelne Terminologie-Einträge einer Norm in verschiedenen Formaten (NISO STS, TBX und XHTML) ausliefern kann.

Die ML-Komponente des Harbsafe2-Projekts setzt genau dort an und fragt über entsprechende REST-Aufrufe die aktuell gültige Terminologie-Daten des XML-basierten Normenbestands im XHTML-Format an, welches durch das DKE CaaS entsprechend geliefert wird und anschließend weiterverarbeitet werden kann.

Die Verwendung des Docker-Containers für das Softwaresystem bat zahlreiche Vorteile für die Nutzung und Pflege der Anwendung. Durch die Plattformunabhängigkeit können verschiedene Betriebssysteme und Hardwarekonfigurationen problemlos unterstützt werden. Die hohe Skalierbarkeit ermöglicht es, das System bei steigender Last oder zunehmender Nutzeranzahl effizient anzupassen, um eine kontinuierlich hohe Leistungsfähigkeit zu gewährleisten. Darüber hinaus tragen die verbesserten Sicherheitsstandards dazu bei, das System vor potenziellen Bedrohungen zu schützen und die Vertrauenswürdigkeit der Anwendung für die Nutzer zu erhöhen. Die Flexibilität bei der Nutzung als Service erlaubt es, das System einfach in bestehende Infrastrukturen zu integrieren oder bei Bedarf zu erweitern, was die langfristige Pflege und Weiterentwicklung des Softwaresystems erleichtert.

Arbeitspaket 6: Ergebnisse

Für eine fortlaufende weitere Einsatzfähigkeit wurden die Services als Docker-Container bereitgestellt, mit der ein hohes Maß an Interoperabilität, Skalierbarkeit sowie Sicherheit gewährleisten zu können, sowie zukunftsorientiert die Nutzung in Produktivsystem zu ermöglichen. Das System wird durch Portainer (ein Container Management System mit einer grafischen Benutzerschnittstelle) mittels einer Benutzerschnittstelle im Webbrowser gesteuert und verwaltet. Durch Portainer ist die Anbindung weiterer Container sichergestellt und eine Integration weiterer Funktionen sowie Services ist möglich.

VDE Services und VDE Operations administriert und verwaltet zukünftig diese Systemarchitektur und nutzt die im Projekt definierten Schnittstellen für eine anschließende Nutzung.

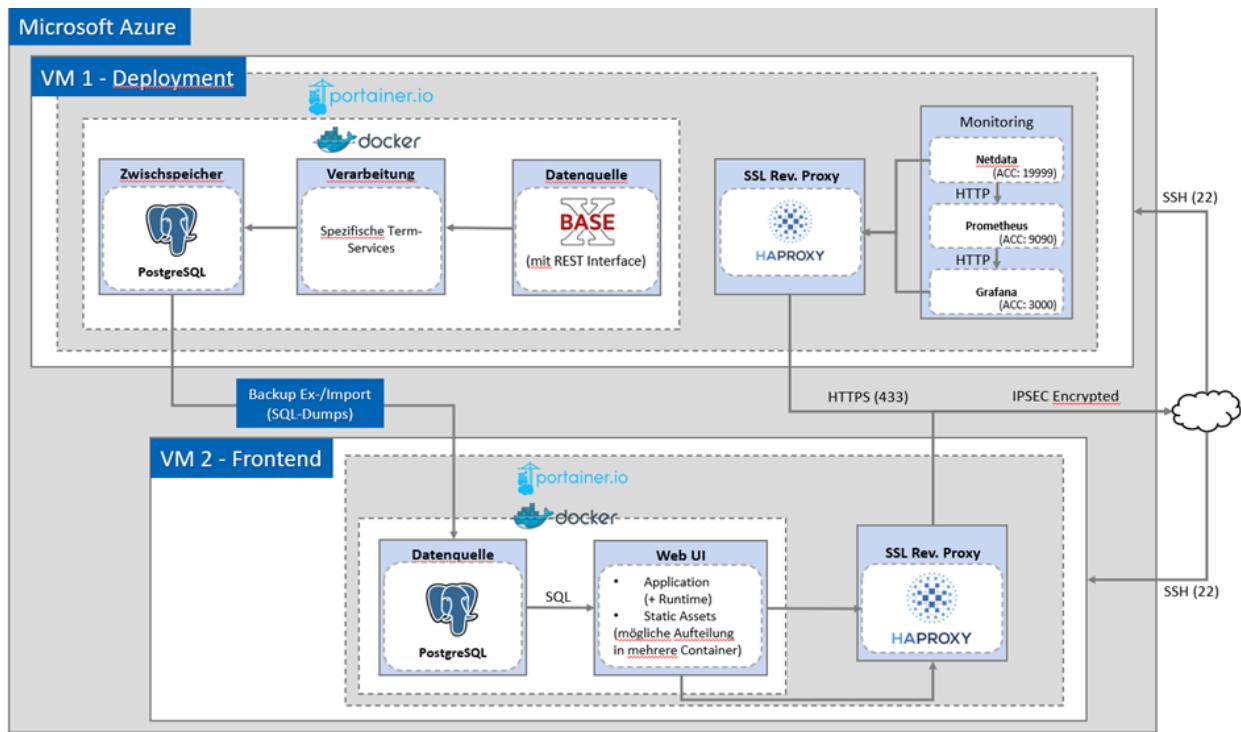


Abbildung 3 Gesamtsystem

Durch die Container Nutzung ist es möglich einen Parallel-Entwicklungsprozess einzuführen. Bei einem Update eines Service können somit Test- sowie Produktivbetrieb zeitgleich mit der jetzigen Systemarchitektur betrieben werden [19].

Arbeitspaket 7: Zielstellung

AP-Nr. 7	Design und Implementierung der Ausgabekomponente und Proposalgenerierung
Zeitraum	Projektmonat 13 bis Projektmonat 25
AP-Leiter	INOSOFT
Ziele des Arbeitspaketes	
<p>Die Frontend- bzw. Ausgabekomponente für Normexperten und interessierte Normanwender bildet eine separate Webapplikation, die Prädiktionsergebnisse (AP3) und die erhobenen Metriken (AP4) übersichtlich in einem Dashboard präsentiert. Harmonisierungstasks werden aufgelistet und Begriffseinträge mit den prädierten Strukturrelationen angezeigt. Moderne Such- und Filterfunktionen werden implementiert. Das Modul ermöglicht die Erstellung von Proposalformularen in den herkömmlichen Textformaten, die manuell angepasst und über die gängigen, ggf. erweiterten tabellarischen Templates (Commenting Table) in den Normungsprozess eingehen können.</p> <p>INOSOFT entwickelt nach modernen Standards eine klassische Web-Applikation zur Darstellung der Terminologie mit allen ML-Ergebnissen (Harmonisierungsfälle, Relationen, strukturierte Suchergebnisse).</p> <p>DKE begleitet die Entwicklungsarbeiten, begutachtet und validiert die Gebrauchstauglichkeit des Softwareprodukts und stimmt Anforderungen zur Softwareintegration ab.</p> <p>ITL begleitet die Entwicklungsarbeiten, stimmt Datenmodell ab und stellt Anforderung zu Funktion und Darstellung der ML-Ergebnisse. Ferner wird der Unterauftragnehmer zur Begutachtung und Beratung hinsichtlich der Gebrauchstauglichkeit aus Anwenderperspektive eingesetzt.</p>	

Output

1. Einsatzfähiges und gebrauchstaugliche Web-Applikation zur Darstellung der Ergebnisse, einschließlich moderner Such- und Filteroptionen zur Integration mit dem ML-Backendsystem und der DKE-Service-Umgebung. (→ AP 6)
2. Veröffentlichung im Netz für einen definierten Nutzerkreis gemäß Nachnutzungsplan.

Arbeitspaket 7: Ablauf

Die Web-Applikation wurde von INOSOFT AG nach Anforderungen (AP 1) und Datenmodellen (AP 5) der TU Braunschweig geplant, entworfen und implementiert, und von der DKE validiert und zur Integration in die Serverumgebung aufgenommen (AP 6). Die Veröffentlichung erfolgte plangemäß als Prototyp unter <http://eglossary.dke.de>.

Tabelle 8: Entwicklungsablauf der Web-Applikation stellt den Ablauf der maßgeblichen Entwicklungsarbeiten dar.

Tabelle 8: Entwicklungsablauf der Web-Applikation

1. Ausarbeiten einer Systemarchitektur
a) Entwurf/Ausarbeitung eines Maskendesign bezogen auf Basis der Anforderungen
b) Wählen einer mehrschichtigen, skalierungsfähigen, Plattform-unabhängigen und Cloud-fähigen Architektur
2. Erstellen eines Rahmens für die Systemarchitektur
a) Rahmen besteht aus Frontend, Backend und Plattform-spezifischer Konfiguration
3. Implementierung
a) Umsetzen des Backends als REST-Interface
b) Umsetzen des Frontends als clientseitige Webanwendung
4. Integrationsphase
a) Integration der bereitgestellten PostgreSQL-DB
b) Bereitstellung der Lösung auf Docker-Infrastruktur
5. Testphase
a) Durchführen von Entwickler-Tests
b) Durchführen von Integrationstests
6. Rollout-Phase
a) Bereitstellung der Lösung in der DKE-Umgebung

Arbeitspaket 7: Ergebnisse

Das Ergebnis ist ein moderner, performanter und voll einsatzfähiger Prototyp einer Web-Applikation zur Darstellung der Ergebnisse, einschließlich moderner Such- und Filteroptionen (siehe Abbildung 4). Die Seite umfasst ca. 65.000 englische Begriffseinträge aus dem DKE-Normenkorpus, auffindbar nach Komitee und Normen oder per Benennungssuche. Die Ergebnisse der Benennungssuche sind gruppiert nach teilweisen Übereinstimmungen und zu jeder Benennung werden alle Einträge als Semantic Spectrum, d. h. gruppiert nach präzisiert identischen Begriffen zu unterschiedlichen, phrasenäquivalenten Definitionen angezeigt. Auf der Detailseite zu jedem Eintrag werden SemAn-Pretty-Prints und ggf. Hinweise über Anforderungsverstöße an die Formulierung von Definitionen aufgeführt.

Einschränkungen / Hinweise:

Die Metriken zum Harmonisierungsgrad aus AP4 wurden nicht zur Präsentation vorgesehen. Da sie seitens TU Braunschweig und DKE nicht als zielführend Anwendungs-geeignet erachtet wurden.

Anstelle der Masken zur Erstellung von Proposalformularen (Commenting Table) zur Eingabe in den Normungsprozess wurde eine umfangreiche Funktion für Auswahl und Export von Begriffseinträgen entwickelt. Der Export erlaubt dem Anwender die flexiblere Weiterbearbeitung in herkömmlichen Textverarbeitungsprogrammen.

Die Web-Applikation ist aus Gründen der Modularität der Komponenten und der Datensicherheit Software- und Netzwerk-technisch vom TermAnalyzer (AP 5) entkoppelt. Zur Aktualisierung der Daten wurde ein unkomplizierter semi-automatischer Datenprozess entwickelt. Die gemeinsame Datenbankschema-Definition erlaubt den transformationsfreien manuellen Transfer der Ergebnisse zur Darstellung der Ergebnisse im Web per SQL.

The screenshot shows the homepage of the DKE Enhanced Glossary. At the top, there is a search bar with placeholder text "Term or standard reference number" and a magnifying glass icon. Below the search bar is a navigation menu with a house icon and other icons. On the left, there is a sidebar titled "SEARCH RESULT FILTER" with the subtext "no filter set". The main content area contains a brief introduction about the glossary, mentioning it is a compilation of electrotechnical terminology in English extracted from IEC standard publications. It highlights three ways definitions are checked: automatically against formal requirements, cross-checked for semantic similarity using machine learning, and parsed and pretty-printed using SemAn. Below this is a note about the project being funded by the German Federal Ministry of Education and Research. At the bottom of the main content area is a section titled "BROWSE STANDARD DOCUMENTS BY TECHNICAL COMMITTEE" with a table:

ISO	TC 2	TC 3	TC 4
TC 5	TC 7	TC 8	TC 9
TC 10	TC 11	TC 13	TC 14

Abbildung 4 Browser-Ansicht der entwickelten Web-Plattform <http://e-glossary.dke.de>.

Arbeitspaket 8: Zielstellung

AP-Nr. 8	Öffentlichkeitsarbeit, Präsentation und Schulung		
Zeitraum	Projektmonat 11 bis Projektmonat 23		
AP-Leiter	DKE		
Teilnehmer	ITL	DKE	INOSOFT
Ressourcen in Stunden	240	680	40
Ziele des Arbeitspaketes			
	<p>In Publikationen und Workshops werden die Ergebnisse einer interessierten Öffentlichkeit präsentiert. Darüber hinaus werden Anwender in der Handhabung geschult. Schulungen erheben Nutzerrückmeldungen, die in die Finalisierung des Softwareprodukts eingehen.</p> <p>DKE informiert die Normungsexperten über Entwicklungen über die DKE-Webpräsenz und eigene Publikationsorgane, richtet Schulungen und Workshops aus und koordiniert Konferenzauftreitte.</p> <p>ITL stellt Materialien und präsentiert Ergebnisse und Methoden, koordiniert mit der wissenschaftlichen Publikationstätigkeit.</p> <p>INOSOFT stellt Materialien zur Anwenderoberfläche bereit.</p> <p>Output</p> <ul style="list-style-type: none"> 3. 2-3 Artikel auf der DKE-Webpräsenz. 4. 2-3 Schulungen/Workshops für interessierte Mitarbeiter und Referenten. 5. Präsentationen auf Konferenzen. 		

Arbeitspaket 8: Ablauf

Für eine möglichst zielgerichtete Präsentation wurde im Verlaufe des Projekts eine Liste mit allen interessierten Gruppen und Gremien erstellt, bei denen das Projekt und die Ergebnisse vorgestellt werden sollten. Im Fokus stand dabei besonders die Interessierten und thematisch in Bereich der Terminologie verzahnten Gremien und Gruppen. Das Softwaretool konnte so direkt bei möglichen Anwendern vorgestellt werden. Bei der Auflistung der potenziellen interessierten Gruppen und Gremien wurde zudem zuvor und fortlaufend eine Priorisierung durchgeführt. Durch die zahlreichen Präsentationen sollte außerdem Feedback gesammelt werden, welches in der Entwicklung der Software berücksichtigt wurde. Des Weiteren wurden zwei Beiträge für die DIN-Mitteilung verfasst (Ausgaben: 11/2021 und 10/2022). Am 20.09.2022 fand der Abschlussworkshop des Projekts statt, welcher als Hybridmeeting online und vor Ort in Frankfurt am Main zur Teilnahme eingeladen hat. Eingeladen wurden alle interessierten Gruppen aus der Liste. Auch die VDE SPEC wurde zum Ende des Projekts fertiggestellt und wird zeitnah nach Projektabschluss veröffentlicht. Auf der DKE-Seite wurde eine Projektseite erstellt, auf der das Projekt und die Zielsetzung

erläutert werden. Auf dieser werden zudem die Vorträge aus dem Abschlussworkshop zur Verfügung gestellt. Als wichtiges Medium zur Präsentation der Projektergebnisse, wurde die Roadmap Industrie 4.0 identifiziert, da diese auch von vielen Experten, welche in der Normung tätig sind, gelesen wird. Die wichtigsten Ergebnisse wurde in der „Roadmap Industrie 4.0“ in Form von Handlungsempfehlungen eingebracht.

Arbeitspaket 8: Ergebnisse

Die Vorstellung des Projekts und dessen Ergebnisse haben in den Gremien und interessierten Gruppen zu positiven Rückmeldungen geführt. Teilweise durch Folgetermine oder Empfehlungen zu anderen Personen/Gruppen, für die das Thema relevant ist. Hierdurch wurde eine Aufklärungsarbeit geleistet, welche in der Normung sehr wichtig ist, da am Ende des Tages die Experten, welche in den Gremien sitzen und Normen erarbeiten auch für die Qualität der Terminologie mitverantwortlich sind. Da die Problematik um die Begriffsdefinitionen nicht allen Experten bekannt ist, ist die Aufklärungsarbeit in diesem Gebiet eines der Ergebnisse aus dem Aufgabenpaket. Verzeichnet werden konnte außerdem eine erfolgreiche Präsentation in neun verschiedenen Gremien, TC's und Gruppen, welche in der angehängten Übersicht aufgeschlüsselt sind. Vermerkt wurde das Datum des Treffens, eine Information, um welches Gremium/Gruppe es sich handelt und welches Ergebnis aus der Vorstellung hervorging. Zusammenfassend lässt sich sagen, dass es eine sehr positive Rückmeldung zu den Bestrebungen des Projekts gab und so auch Kontakt zum IEC und dem TC1 sowie zum IEC SMB SG13 TF3 entstanden ist, was für die Verwertung der Projektergebnisse ein wichtiger Meilenstein ist. Zudem wurde mit dem AK STD_1941.0.1 welcher für die Erarbeitung der Normungsroadmap Industrie 4.0 verantwortlich ist, der Bedarf für mehrere Handlungsempfehlungen in Bezug auf Terminologie erkannt. In vier Handlungsempfehlungen konnten somit die wichtigen Erkenntnisse aus der Projektarbeit verankert werden. Die Handlungsempfehlungen werden in der nächsten Ausgabe der Roadmap Industrie 4.0 integriert werden. Der Abschlussworkshop hat zu spannenden Diskussionen zum Projektende beigetragen. Herrn Patrick Zimmermann (IEC TC 1), welcher unter anderem auch die Pflege des IEV mit betreut, hat als Referent beim Workshop die Problematiken im IEV dargestellt und in einer anregenden Diskussion wurden weitergehende Möglichkeiten für die Verwertung der Projektergebnisse besprochen.

Tabelle 9: Projektpräsentationen und Workshops

Datum	Komitee Nummer	Komitee Name
01.02.2021	K111	Terminologie
01.02.2021	UK 931.2	Begriffe der industriellen Automatisierung
10.09.2021	IEC SMB SG12 TF1	Standardization Management Board/Digital Transformation and Systems Approach
08.06.2022	DIN-TERM	Terminologiedatenbank DIN-TERM

Datum	Komitee Nummer	Komitee Name
21.06.2022	AK STD_1941.0.1	Normungsroadmap Industrie 4.0
23.06.2022	TBINK AK IT-Security	TBINK-Arbeitskreis IT-Security und Security by Design
07.07.2022	K931	Systemaspekte der Automatisierung
14.07.2022	IEC TC65/SC 65E/WG 2	Devices and integration in enterprise systems/Product properties & classification
20.07.2022	AK STD_1941.0.1	“IDiS informiert”
22.09.2022	AK STD_1941.0.1	Normungsroadmap Industrie 4.0
20.09.2022	n.A.	Abschlussworkshop Harbsafe 2: “Neuronale Sprachassistenz für die Normungscommunity”
10.10.2022	AK 931.0.1	Merkmale
03.11.2022	IEC TC65/SC 65E/WG 2	Devices and integration in enterprise systems/Product properties & classification
Geplant	IEC SMB SG13 TF3	Standardization Management Board/Working with consortia

Schließlich wurde von der Inosoft AG unter Beteiligung der DKE ein Imagefilm zum Thema „Harbsafe 2“ und „E-Glossary“ produziert, welcher auf der Projektseite bei der DKE und auf der Website von INOSOFT AG veröffentlicht wurde und für weitere Veröffentlichungen vorgesehen ist.

II.2 Wichtigste Positionen des zahlenmäßigen Nachweises

Der quantitative Nachweis umfasst hauptsächlich Personalkosten für die qualifizierten Projektmitarbeiter (siehe Anlage 3 und 4). Reisekosten wurden aufgrund der pandemischen Lage weitgehend umgewidmet bzw. nicht in Anspruch genommen.

II.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Aufgrund des hohen Zeitaufwands für die Harmonisierung der Terminologie in der Normung besteht ein dringender Bedarf für Werkzeuge und Tool welche in diesem Bereich unterstützen können. Sowohl für die Normanwender als auch Ersteller spielt die Nutzung von Harmonisierten Normen eine große Rolle. Dies spart Ressourcen und beugt Missverständnissen vor. Besondere Relevanz gewinnt die Thematik im Kontext der Digitalisierung, da sich neue Geschäftsmodelle für innovative Marktsegmente wie

Elektromobilität, Smart Home, künstliche Intelligenz und Industrie 4.0 nur schwer ohne einheitliche Begriffe etablieren lassen.

Bereits jetzt zeichnet sich in diesen Themenfeldern eine Vielfalt an Mehrfachdefinition ab. Um dies zu verhindern und zu bereinigen ist eine umfangreiche Pflege von verbindlichen Quellen wie dem IEV-Wörterbuch unumgänglich. Auch für die digitale Transformation der Normung ist ein ausschlaggebender Faktor, um Normen maschinenlesbar und interpretierbar zu machen, stellt eine Harmonisierte Terminologie das grundlegende Fundament dar.

Das in Harbsafe 2 weiterentwickelte Softwaretool setzt genau an dieser Stelle an und ermöglicht derzeitig sehr zeitaufwendige und manuelle Prozesse im Rahmen der Harmonisierungsbestrebungen in der Normung deutlich zu verkürzen. Auch das breite Interesse bei den vorgestellten Gremien und Gruppen spiegelt den Nutzen und den Bedarf an einem solchen Softwaretool wider.

Im Zentrum des wissenschaftlichen und technischen Erfolges des Förderprojekts Harbsafe 2, steht die entwickelte Software, welche über das e-glossary.dke.de zugänglich ist. Hier wurde zum ersten Mal eine nutzbare Softwarelösung zur Voll- bzw. Teilautomatisierung der Harmonisierungsarbeiten in den Terminologie Datenbanken der IEC entwickelt, welche einen greifbaren Mehrwert liefert.

Die Weiterentwicklung und einen nutzbringenden Einsatz der im Projekt entwickelten Software bei der DKE, mit Projektpartnern ggf. auch auf internationaler Ebene wird noch besprochen und weiterverfolgt.

Eine der zukünftigen Anschlussmöglichkeiten der Projektergebnisse, liegt in der Anwendung von SMART Standards. Erst durch die Möglichkeit den Normenbestand der IEC in einem maschinenlesbaren Format (XML) zu generieren hat es ermöglicht im Rahmen des Förderprojekts den schieren Umfang der Terminologiebestände digital auszuwerten und zu analysieren. Für die weiteren Schritte im Rahmen der Digitalisierung der Normung und auch für andere Normungsorganisationen sollte die Ergebnisse und das Softwaretool aus Harbsafe 2 berücksichtigt werden.

Weitere Anschlussmöglichkeiten bilden sich auch in einer Weiterentwicklung der Software ab, abgestimmt auf die Anwendungsfälle für die Pflege und Bearbeitung der Terminologiebestände des IEV-Wörterbuchs. Die Adaption dieser Verfahren auf andere Normungsinstitutionen und Anwendungsfälle stellt eine weitere Fortführungsmöglichkeit dar.

II.4 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Keine Fortschritte bei anderen Stellen bekannt.

II.5 Erfolgte oder geplante Veröffentlichungen der Ergebnisse

Zur Verbreitung der Projektergebnisse und Erläuterung des Umgangs mit den harmonisierungsrelevanten Annotationen in der Normungspraxis wurden eine VDE SPEC

entwickelt und in den Veröffentlichungsprozess gegeben. Das Format VDE SPEC wurde alternativ zur VDE-Anwendungsregel gewählt, welches zunächst geplant war. Das Projektteam hat sich unter Absprache zwischen DKE und ITL für die Erstellung einer VDE SPEC entschieden. Hintergrund dieser Entscheidung war, dass eine VDE SPEC direkt vom Projektteam verfasst werden kann und somit eine transparentere und schnellere Verwertung der Ergebnisse darstellt.

Die Disseminationstätigkeiten werden ferner in den eingehenden Darstellungen von AP 8 detailliert erläutert.

Fachpublikationen:

- Kurzbeitrag in VDE Dialog 10/2021
- DIN-Mitteilungen 11/2021
- DIN-Mitteilungen 10/2022
- Normungsroadmap Industrie 4.0 (v5)
- VDE Spec 11/2022

Wissenschaftliche Publikationen:

- Peter Bernard Ladkin, Lou Xinxin, Dieter Schnäpp: The Terminological Analysis Method SemAn and its Implementation. Safety-Critical Systems eJournal vol.2 no.1 [SCSC-183], <https://scsc.uk/scsc-183>.

Des Weiteren wurden die Arbeit und Ergebnisse in diversen Gremien präsentiert: (Siehe „Tabelle 9: Projektpräsentationen und Workshops“ in der eingehenden Erläuterung zu AP 8)

Für die Weiterentwicklung und Verwendung wurde das Softwaretool dem IEC TC 1 zur Verfügung gestellt. Im IEC TC 1 wird über verschiedene Arbeitsgruppen die Bearbeitung und Weiterentwicklung des IEV gesteuert, hier soll das entwickelte Softwaretool dabei unterstützen eine Teilautomatisierung der Prozesse zu erreichen. Ein Pilotbetrieb im Bereich der Automatisierung (z. B. DKE K 931.2) ist Experten-seitig gewünscht und wird organisatorisch geplant. Der Prototyp ist unter <http://e-glossary.dke.de> öffentlich passwortgeschützt verfügbar und kann auf Anfrage für Normungsexperten zugänglich gemacht werden. Zusätzlich wurde über das Webportal von VDE/DKE eine Projekt-Homepage veröffentlicht, auf der das Vorhaben vorgestellt wird, sowie auf die Tools zugegriffen werden kann.

Anlagen (einzelne von den Projektpartnern eingereicht.)

- Anlage 1 – Erfolgskontrollbericht (nicht öffentlich)
- Anlage 2 – Berichtsblatt und Document Control Sheet (kann veröffentlicht werden)
- Anlage 3 – Verwendungsnachweis (nicht öffentlich)
- Anlage 4 – Aufstellung nach Kostenarten (nicht öffentlich)

III Literaturverzeichnis

- [1] Honnibal, Montani, V. Landeghem und Boyd, „Zendo.org,“ 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.1212303>.
- [2] Reimers und Gurevych, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019.
- [3] Deutsches Institut für Normung e.v, „Deutsche Normungsstrategie, Berlin,“ 2017. [Online]. Available: <https://www.din.de/blob/234448/58f20dcc3cecf12cc6a91f956cc3c160/dns-2017-layout-data.pdf>. [Zugriff am 23 03 2023].
- [4] International Electrotechnical Commission, „Electropedia: The World's Online Electrotechnical Vocabulary,“ 2023. [Online]. Available: <https://electropedia.org/>. [Zugriff am 23 03 2023].
- [5] ISO/TC 37/SC 1 Principles and methods, *ISO 860: Terminology work — Harmonization of concepts and terms*, International Organization for Standardization, 2007.
- [6] Becker, Arndt und Schnäpp, „Schlussbericht Projekt: HARBSAFE (Verbundvorhaben),“ TIB Leibniz-Informationszentrum Technik und Naturwissenschaften Universitätsbibliothek, 2020.
- [7] T. Mikolov, „Efficient Estimation of Word Representations in Vector Space,“ 2013. [Online]. Available: <https://arxiv.org/abs/1301.3781>. [Zugriff am 24 03 2023].
- [8] H. Lee, „An efficient framework for learning sentence representations,“ 2018. [Online]. Available: <https://openreview.net/forum?id=rJvJXZb0W>. [Zugriff am 24 03 2023].
- [9] A. Conneau, D. Kiela, H. Schwenk, L. Barrault und A. Bordes, „Supervised Learning of Universal Sentence Representations from Natural Language Inference Data,“ 2017. [Online]. Available: <https://arxiv.org/abs/1705.02364>. [Zugriff am 24 03 2023].
- [10] S. Subramanian, A. Trischler, Y. Bengio und J. P. Christopher, „Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning,“ 2018. [Online]. Available: <https://arxiv.org/abs/1804.00079>. [Zugriff am 24 03 2023].

- [11] D. Cer, Y. Yang und S.-y. Kong, „Universal Sentence Encoder,“ 2018. [Online]. Available: <https://arxiv.org/abs/1803.11175>. [Zugriff am 24 03 2023].
- [12] S. Aora, Y. Liang und T. Ma, „A Simple but Tough-to-Beat Baseline for Sentence Embeddings,“ 2017. [Online]. Available: <https://openreview.net/pdf?id=SyK00v5xx>. [Zugriff am 24 03 2023].
- [13] J. Devlin, M.-W. Chang, K. Lee und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,“ 2019. [Online]. Available: <https://aclanthology.org/N19-1423/>.
- [14] H. Femmer, D. M. Fernández, S. Wagner und S. Eder, „Rapid quality assurance with Requirements Smells,“ 2016. [Online]. Available: <https://arxiv.org/abs/1611.08847>. [Zugriff am 24 03 2023].
- [15] Unbekannt, „STS4i – Standards Tag Suite XML for interoperability,“ 2023. [Online]. Available: <https://github.com/sts4i>. [Zugriff am 23 03 2023].
- [16] R. Nils und G. Iryna, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.
- [17] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang und M. Zhou, „Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers,“ 2020. [Online]. Available: <https://arxiv.org/abs/2002.10957>. [Zugriff am 24 03 2023].
- [18] B. Ladkin, L. Xinxin und D. Schnäpp, „The Terminological Analysis Method SemAn and its Implementation,“ 2023. [Online]. Available: <https://scsc.uk/journal/index.php/scsj/article/view/20>. [Zugriff am 24 03 2023].
- [19] BaseX GmbH, „Base X: The XML Framework,“ [Online]. Available: <https://basex.org/>. [Zugriff am 23 03 2023].

Anlage 1: Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart Schlussbericht	
3. Titel Verbundvorhaben: Softwareassistierte Harmonisierung - Teilvorhaben: Entwicklung und Erprobung einer Machine-Learning-basierten Softwareunterstützung für die Harmonisierung der DKE-Normungsterminologie, einschließlich formaler Methoden und Prozesse, im Rahmen von Pilotanwendungen in der Normungspraxis		
4. Autor(en) [Name(n), Vorname(n)] Behrend, Mark Salimi, Bahram Schnäpp, Dieter Lee, Sascha Man-Son	5. Abschlussdatum des Vorhabens 30.09.2022	6. Veröffentlichungsdatum 31.03.2023
	7. Form der Publikation Schlussbericht	
8. Durchführende Institution(en) (Name, Adresse) Technische Universität Braunschweig Institut für Verkehrssicherheit und Automatisierungstechnik (ITL) Hermann-Blenk-Straße 42, 38108 Braunschweig DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE Merianstraße 28, 63069 Offenbach am Main Inosoft GmbH Im Rudert, 35043 Marburg	9. Ber. Nr. Durchführende Institution	10. Förderkennzeichen ¹⁾ 03TN0018A-C
	11. Seitenzahl 44	
12. Fördernde Institution (Name, Adresse) Bundesministerium für Wirtschaft und Klimaschutz (BMWK) 53107 Bonn	13. Literaturangaben 19	14. Tabellen 8
	15. Abbildungen 3	
16. Zusätzliche Angaben		
17. Vorgelegt bei (Titel, Ort, Datum)		
18. Kurzfassung Bei der Erstellung und Pflege von Terminologie-Datenbanken, kommt es häufig zu Problemstellungen in dem Themenfeldern: Homonymie, Definitionsqualität und Definitionslesbarkeit. Das Projekt Harbsafe 2 hatte zum Ziel die Qualität des vollständigen und korrekten Terminologie-Bestands der IEC als Normungsorganisation zu verbessern. Dazu wurde ein Prototyp einer Webanwendung entwickelt, der es ermöglicht, anwenderfreundlich sämtliche Terminologie-Einträge zu recherchieren. Zusätzlich wurden drei innovative automatische Verfahren angewendet: HintAn, Bedeutungsspektren und SemAn. Diese annotieren die Definitionen und identifizieren Inkonsistenzen, geben einen Überblick über die Vielzahl definitorischer Varianten und schlüsseln die Bedeutungsmerkmale jeder Definition plausibel auf. Das Projekt gliedert sich in drei Arbeitsfelder: Erarbeitung, Modellierung und Evaluation der harmonisierungsrelevanten Annotationen, Entwicklung einer Daten-Pipeline und Komplexe Software-Entwicklung. Die Qualität der Normungsterminologie ist eine wichtige Voraussetzung für die Nutzbarkeit von Normen und Standards. Widersprüche und Inkonsistenzen stellen ihre Anwendbarkeit vor erhebliche Schwierigkeiten. Eine gezielte Terminologie-Harmonisierung ist daher notwendig.		
19. Schlagwörter HintAn, Bedeutungsspektren, SemAn, Semantic-Textual-Similarity,		
20. Verlag	21. Preis	

Anlage 2: Document Control Sheet

1. ISBN or ISSN	2. type of document (e.g. report, publication) Final report
3. title Joint project: Software-assisted harmonization - Subproject: Development and testing of a machine-learning-based software support for the harmonization of DKE standardization terminology, including formal methods and processes, in the context of pilot applications in standardization practice.	
4. author(s) (family name, first name(s)) Behrend, Mark Salimi, Bahram Schnäpp, Dieter Lee, Sascha Man-Son	5. end of project 30.09.2022
	6. publication date 31.03.2023
	7. form of publication Final report
8. performing organization(s) (name, address) Technische Universität Braunschweig Institut für Verkehrssicherheit und Automatisierungstechnik (ITL) Hermann-Blenk-Straße 42, 38108 Braunschweig	9. originator's report no.
DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE Merianstraße 28, 63069 Offenbach am Main	10. reference no. 03TN0018A-C
Inosoft GmbH Im Rudert, 35043 Marburg	11. no. of pages 44
12. sponsoring agency (name, address) Bundesministerium für Wirtschaft und Klimaschutz (BMWK) 53107 Bonn	13. no. of references ?
	14. no. of tables 19
	15. no. of figures 8
16. supplementary notes	
17. presented at (title, place, date)	
18. abstract When creating and maintaining terminology databases, problems often arise in the areas of: Homonymy, definition quality and definition readability. The Harbsafe 2 project aimed to improve the quality of the complete and correct terminology inventory of the IEC as a standards organization. For this purpose, a prototype of a web application was developed, which allows to search all terminology entries in a user-friendly way. In addition, three innovative automatic methods were applied: HintAn, Meaning Spectra, and SemAn. These annotate the definitions and identify inconsistencies, provide an overview of the multitude of definitional variants, and plausibly break down the meaning characteristics of each definition. The project is divided into three work areas: Elaboration, Modeling and Evaluation of Annotations Relevant to Harmonization, Development of a Data Pipeline and Complex Software Development. The quality of standardization terminology is an important prerequisite for the usability of norms and standards. Contradictions and inconsistencies pose considerable difficulties for their applicability. Targeted terminology <u>harmonization is therefore necessary</u> .	
19. keywords HintAn, Bedeutungsspektren, SemAn, Semantic-Textual-Similarity,	
20. publisher	21. price