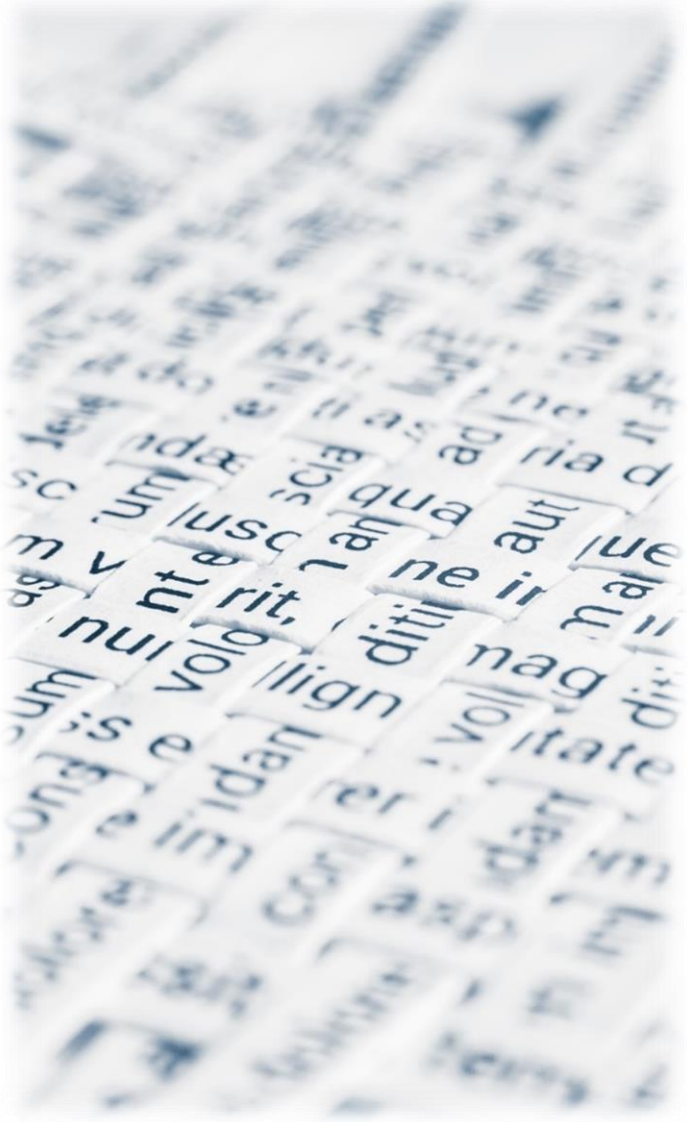


Migration und Konsolidierung

Dipl.-Ing. José Manuel de la Rosa Govantes

Vortrag am 26.04.2017 zur GSE Tagung Hamburg

Agenda

Über uns
Die semantics Kommunikationsmanagement GmbH

Ausgangspunkt und Ziel
Vorbemerkungen und Motivation

Ohne Fleiß kein Preis
Import und maschinelle Analyse

Automatisierung der Redaktionsarbeit
Maschinelle Konsolidierung

Ergebnis
Erwartungen und Ungeplantes

Status Quo und Ausblick
Anwendungen und Entwicklungen

Über uns

Die semantics Kommunikationsmanagement GmbH

semantics Kommunikationsmanagement GmbH

- Gründung 2000 als Spin-off der RWTH Aachen (Prof. Dr. Christian Stetter)
- Interdisziplinäres Team aus z.Zt. 38 Informatikern, Sprachwissenschaftlern und Psychologen
- Verbindung von Forschung und Praxis zur Textverständlichkeit und Texttechnologie
  - > Sprachschulungen, -richtlinien, -analysen, Workshops etc.
  - > Technische Lösungen für große Textbestände (Analyse, Management, Pflege)
  - > Empirische Studien
  - > Forschung in Linguistischer Informatik
- Ferner: Digitalisierungs- und Recherchesysteme für Bibliotheken, Archive und Museen (Marktführer D-A-CH)

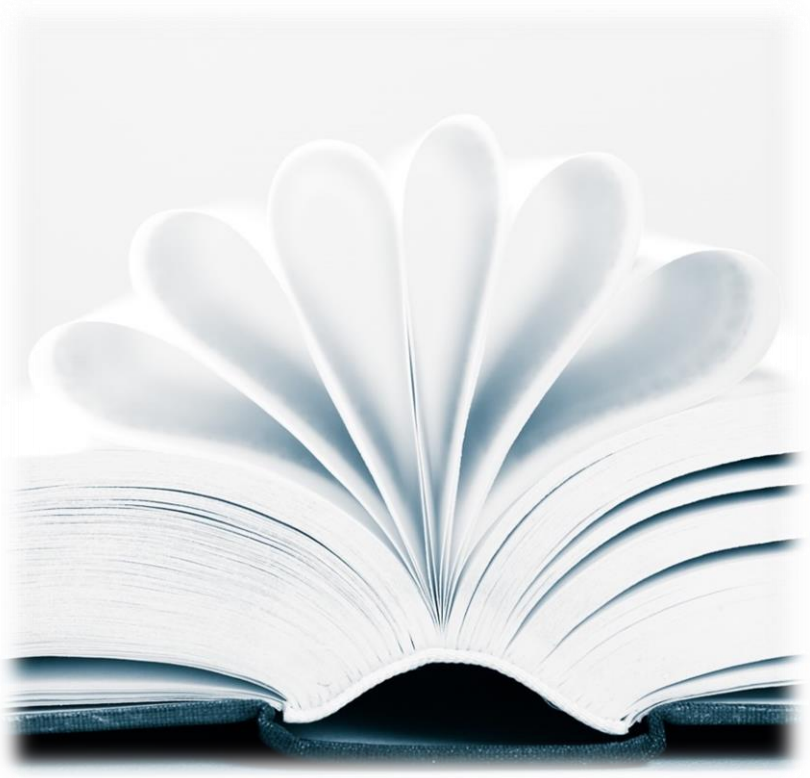
Ausgangspunkt und Ziel

Vorbemerkungen und Motivation

Kontext – Zeitalter der Digitalisierung

- Zeitalter der *Digitalisierung*, BigData, künstlichen Intelligenz
  auch Chatbots, Fake News, Toxic Comments, …
- Ansprüche der Verbraucher *steigen*
  Bzgl. Kommunikation: Inhalt, Reaktionszeit, Kanal
- Kundenansprache optimieren: Bindeglied *Sprache*
  Die richtigen Infos, zur richtigen Zeit, im richtigen Umfang und richtigen Kanal – und dabei den richtigen Ton treffen
- Kommunikation muss *entworfen* und *geplant* werden
  Brief, Email, Portale, Apps, Kurznachrichten, Social Media…
  sind unterschiedlich zu bedienen; Aufwand steigt
- Potential *Gestaltung der Sprache*:
  ⇒ *Nachfrage- und Beschwerdeaufkommen reduzieren*
  ⇒ *Empfehlungen, Kundenbindung und -zufriedenheit erhöhen*

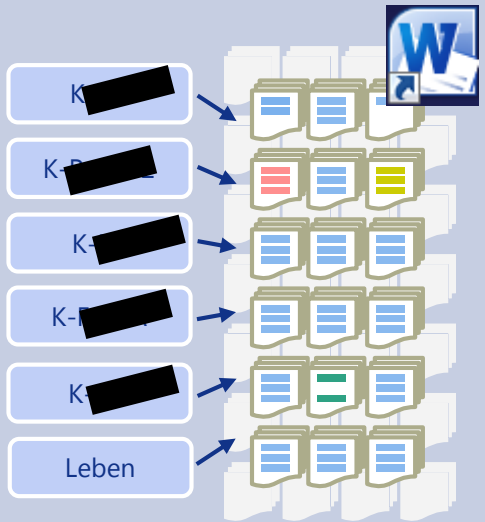
Kontext – Sprache gestalten



- Heute kein (sprach-)wissenschaftlicher Vortrag
- Unsere Technik adressiert das *systemische Problem*: Spannungsfeld Textpflege
  - > Sehr große Textbestände, lange Historie
  - > Große Organisationen, viele beteiligte Personen
  - > Verschiedene Kompetenzen, Rollen und Interessen
  - > Gemeinsame Verantwortung für den Textbestand
- Anwendung dieser Technik im vorgestellten Projekt
- 2 abgeschlossene Großprojekte (D, CH); 8 laufende Projekte
- Migration ohne Konsolidierung möglich und umgekehrt

Ausgangssituation – Redundante AVB

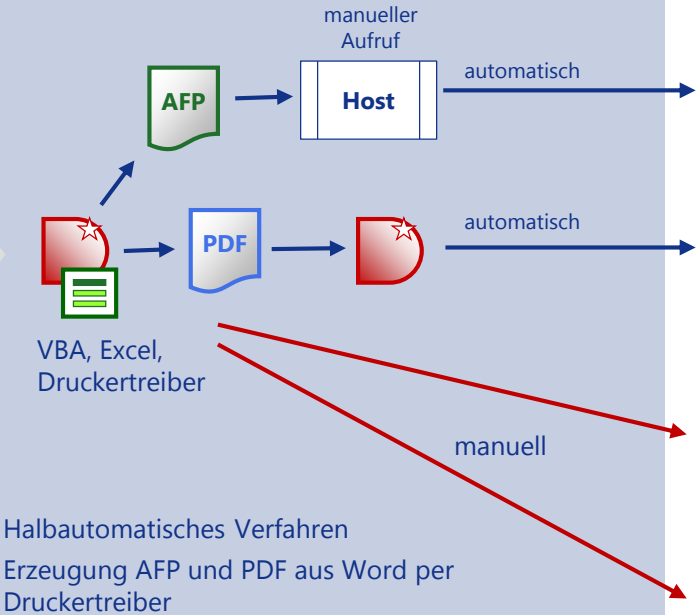
Pflege (Word)



- Pflege mit Word, Koordination rein organisatorisch (zahlreiche Referate)
- Vollausprägung aller Dokumente
- Hohe Redundanzen, erheblicher Aufwand in Mehrfachpflege
- 5 Generationen, 4 Layouts
- Ca. 2.500 Dokumente

manuell

Bereitstellung



- Halbautomatisches Verfahren
- Erzeugung AFP und PDF aus Word per Druckertreiber
- Übergabe AFP an Host und Verteilung PDF per VBA-Makro
- Erzeugung Overlays per Host-Befehl
- Steuerung per Excel und Host-Tabellen
- Punktuell manuelle Bereitstellungen notwendig

Nutzung

OVL
- Policierung, Batch (Overlays)

PDF
- Auskunft, Sachbearbeiter
- Vertrieb
- Interne Dokumentation (versionierte Dokumente, Belegexemplare)

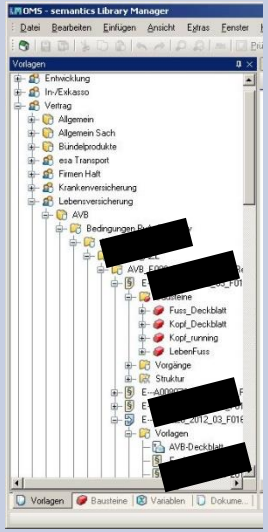
PDF
- Web-Server

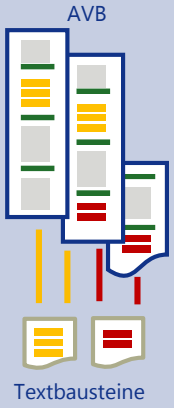
DOC
- Online-, Bündel-Produkte sehr geringe Mengen

Etablierte Verfahren und Technik sollen unverändert beibehalten werden

Ziel: Konsolidierte Redundanzen

Pflege (LM)

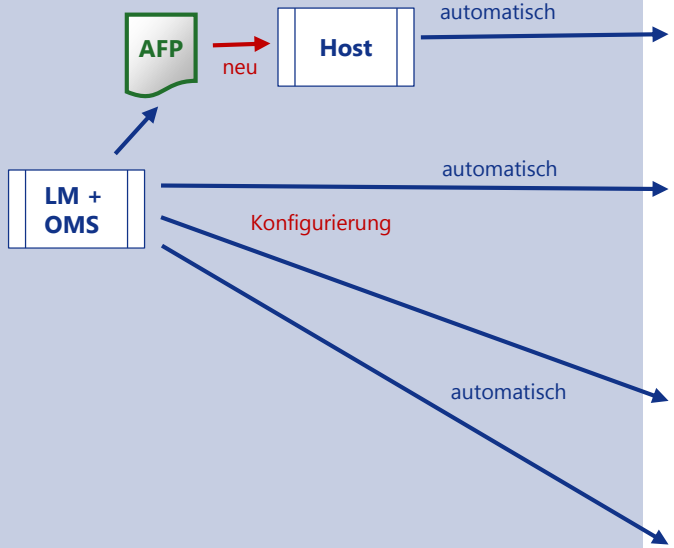




automatisch

- Minimierung Pflegeaufwand: Auslagerung Redundanzen in Textbausteine
- Organisation und Ablage unter Nutzung Bibliothek, Zugriffssteuerung, Rollenkonzept etc.
- Dokumentation, Revisionsführung, Redaktionsprozesse etc. nativ in LM
- Technisch geführte Koordination

Bereitstellung



- Automatische Übergabe- und Freigabeprozesse inkl. Historie
- Nutzung Standards (OMS)
- Keine externen Komponenten (VBA, Druckertreiber), Dokumentations- und Steuerungsinstrumente (Excel)

Nutzung
(Unverändert)

OVL
- Policierung, Batch (Overlays)

PDF
- Auskunft, Sachbearbeiter
- Vertrieb
- Interne Dokumentation (versionierte Dokumente, Belegexemplare)

PDF
- Web-Server

XML, PDF
- Online-, Bündel-Produkte sehr geringe Mengen

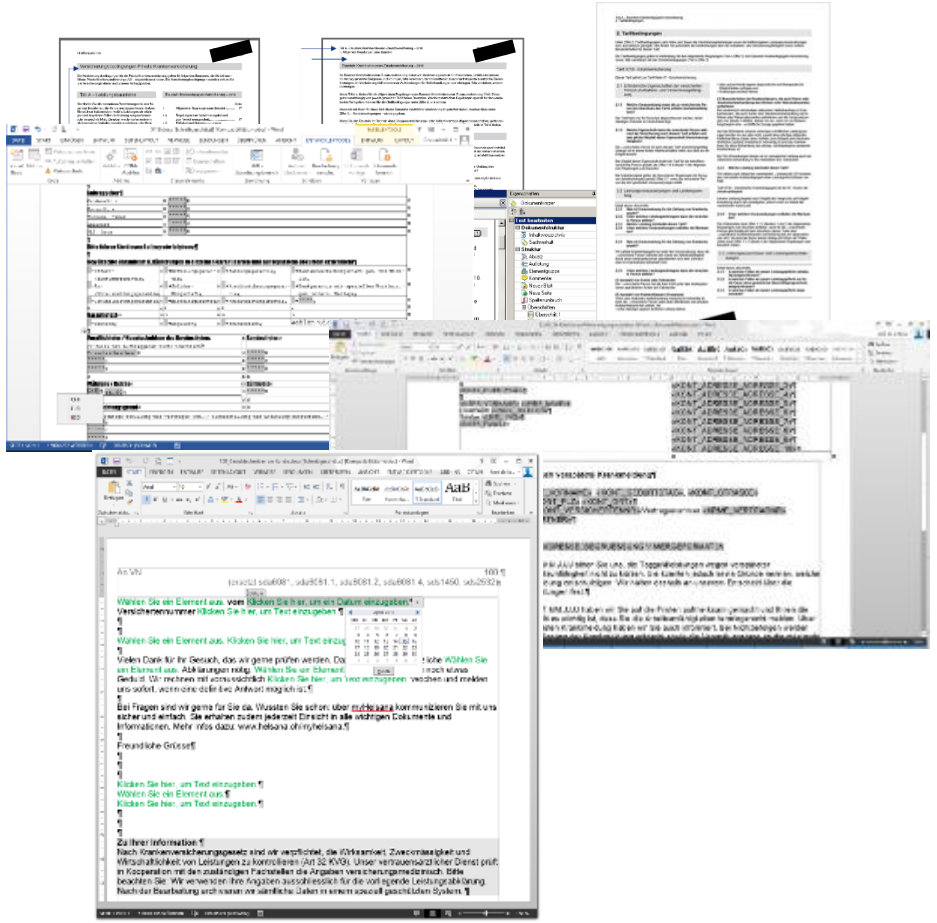
- *Etablierte Technik unverändert*
- *Manuelle Prozesse automatisiert*

Ohne Fleiß kein Preis

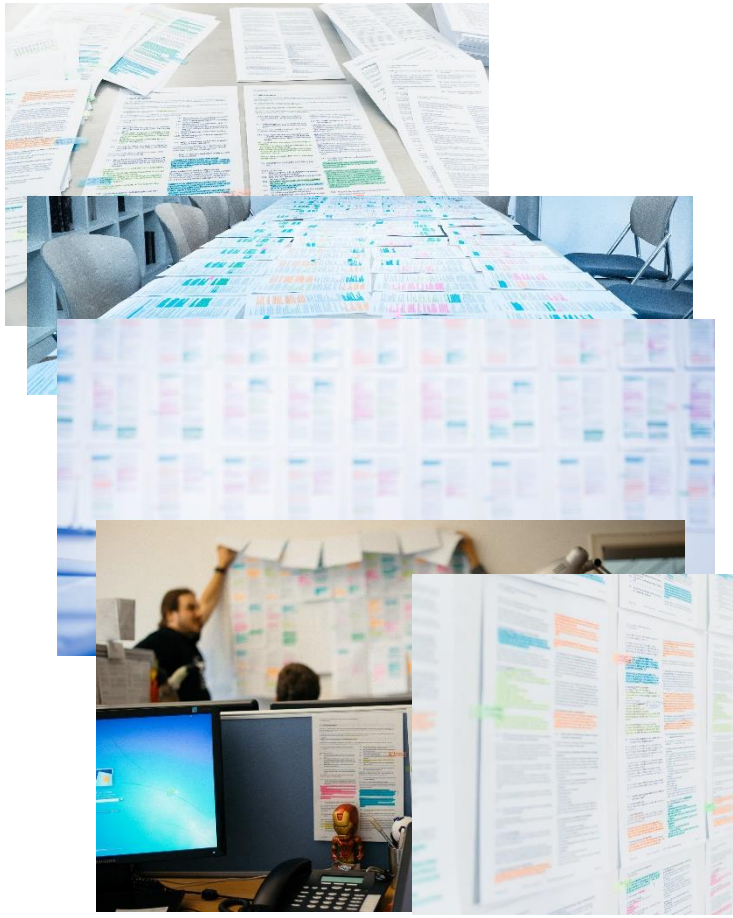
Import und maschinelle Analyse

Import – Standard Word, daher „ganz einfach“…

- Import Word-Format (andere ebenfalls möglich)
  hier: Jahrzehnte gewachsene Variationen…
- Material moderat anspruchsvoll
  ein-/zweispaltiges Layout, Formatierungen, Inhaltsverzeichnisse, …
- Metadaten aus Dateinamen, Textbereichen
  auch zur Zuweisung Layout, Dokumenttypen
- Herausforderungen durch „manuelles Layouting“
  lokale Formatierung, manuelle Umbrüche und Silbentrennung, Tabellen durch Tabs, manuelle Zähler / Inhaltsverzeichnisse, …
  (Lösung z.B. durch Ersetzungstabelle und Baustein-Templates)
- Vielfalt durch Variablen und Eingabeelemente…
  Textmarken, Formularfelder, Feldfunktionen, Inhaltssteuerelemente, gelbe Farbe. Oder: Sprachliche Anweisungen.
  (Einrichtung des Imports je Textbestand)



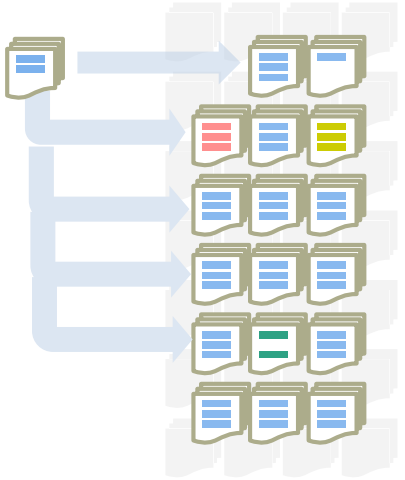
Analyse – Erfahrungen



- Erfahrung mit „kleinen“ Projekten: z.B. Redundanzanalyse schweizerische KV, 120 Dokumente, überschaubar (…?)
- Nutzung Vergleichstool; Ergebnisse konservieren schwierig
- Nach kurzer Zeit entsteht Raumbedarf – wie übergreifende Redundanzen visualisieren?
- Überblick gelang durch Nutzung aller Dimensionen…
- Interesse an weiteren Projekten sehr gering (bei der Belegschaft)

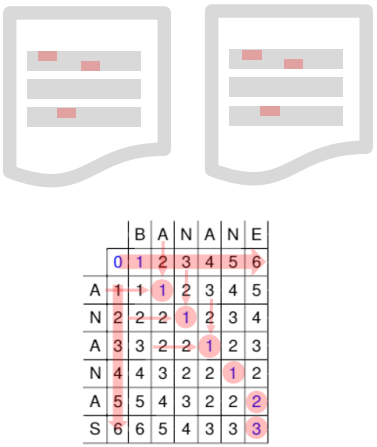
Analyse – Herleitung Zahlen

- Vergleich *d* Dokumente führt zu $D = \frac{d*(d-1)}{2}$ Operationen
- Vergleich jedes Absatzes zweier Dokumente $A = a_1 * a_2$
- Insgesamt $G = D * A$ Textpassagen zu vergleichen
- Im Beispiel
  - > 120 Dokumente, je 15-20 Absätze
  - > $D = 7.080$, $\bar{A} \approx 289 \Rightarrow G \approx 2.046.120$
  - > Optimierung durch intuitive Bewertung potentieller Ähnlichkeit
- Im vorgestellten Projekt
  - > 2.500 Dokumente, je ca. 200 Absätze
  - > $D = 3.123.750$, $\bar{A} \approx 40.000 \Rightarrow \boldsymbol{G \approx 124.950.000.000}$
  - > Maschinelle Verarbeitung und Optimierung notwendig *(sonst bei 1.000/s ca. 4 Jahre)*

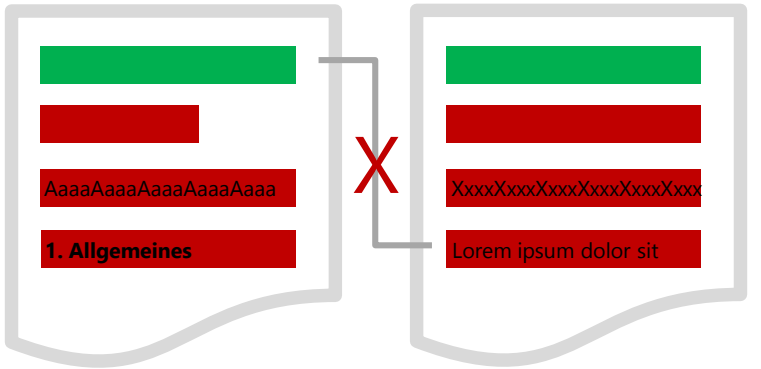


Analyse – Algorithmen

- Suche nach *identischen* und *ähnlichen* Textpassagen; kein Tokenizing o.ä. sinnvoll *(punktuelle Abweichungen: Tarifbezeichner, Erstattungsquote, …)*
- Auswahl: Levenshtein-Distanz, Entfernung zwischen Wörtern; auf Absätze anwenden *(35% Toleranz)*
- Bewährte Algorithmen vorhanden, aber höherer Ordnung *(gewisse Laufzeit, aufgrund Masse signifikant)*
- Optimierung bzw. Vorausschluss u.a. durch
  - > Identität *(Speichervergleich in „Nullzeit“ – keine weitere Prüfung notwendig)*
  - > Länge *(Zeichenanzahl Absatz)*
  - > Fingerprint *(Checksumme Absatz)*
  - > Absatzformat *(Überschriften vs. Text vs. Aufzählungen vs. …)*
  - > Entfernung *(Position im Dokument)*



		B	A	N	A	N	E
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3	3	2	2	1	2	3
N	4	4	3	2	2	1	2
A	5	5	4	3	2	2	2
S	6	6	5	4	3	3	3

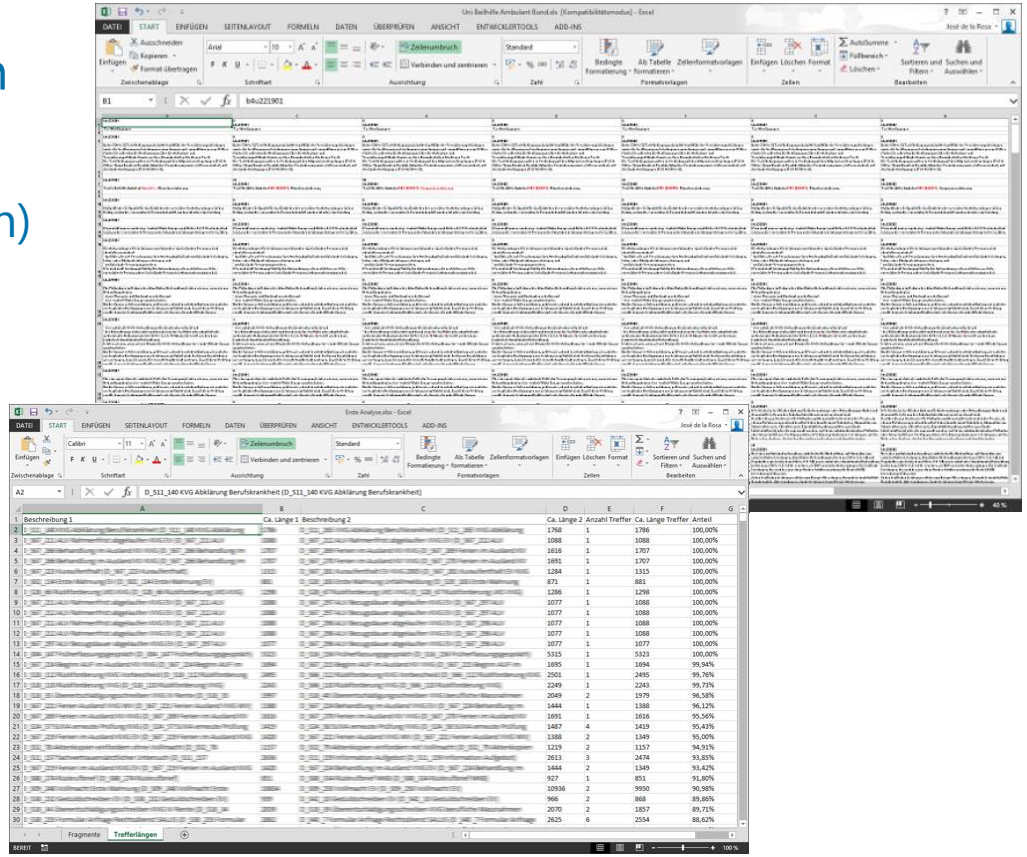


Analyse – fachliche Optimierungen zur Wiederverwendbarkeit

- Splitten großer Fragmente
  z.B. anhand Überschriften, künftige Wiederverwendung
- Splitten von „Overlapps“
  höhere Wiederverwendung
- Zusammenhalten kurze „Ausreißer“
  durch Variablen aufzufangen
- Zusammenhalten fachlicher Einheiten
  z.B. Einleitende Sätze und Aufzählungen

Analyse – Darstellung Ergebnisse

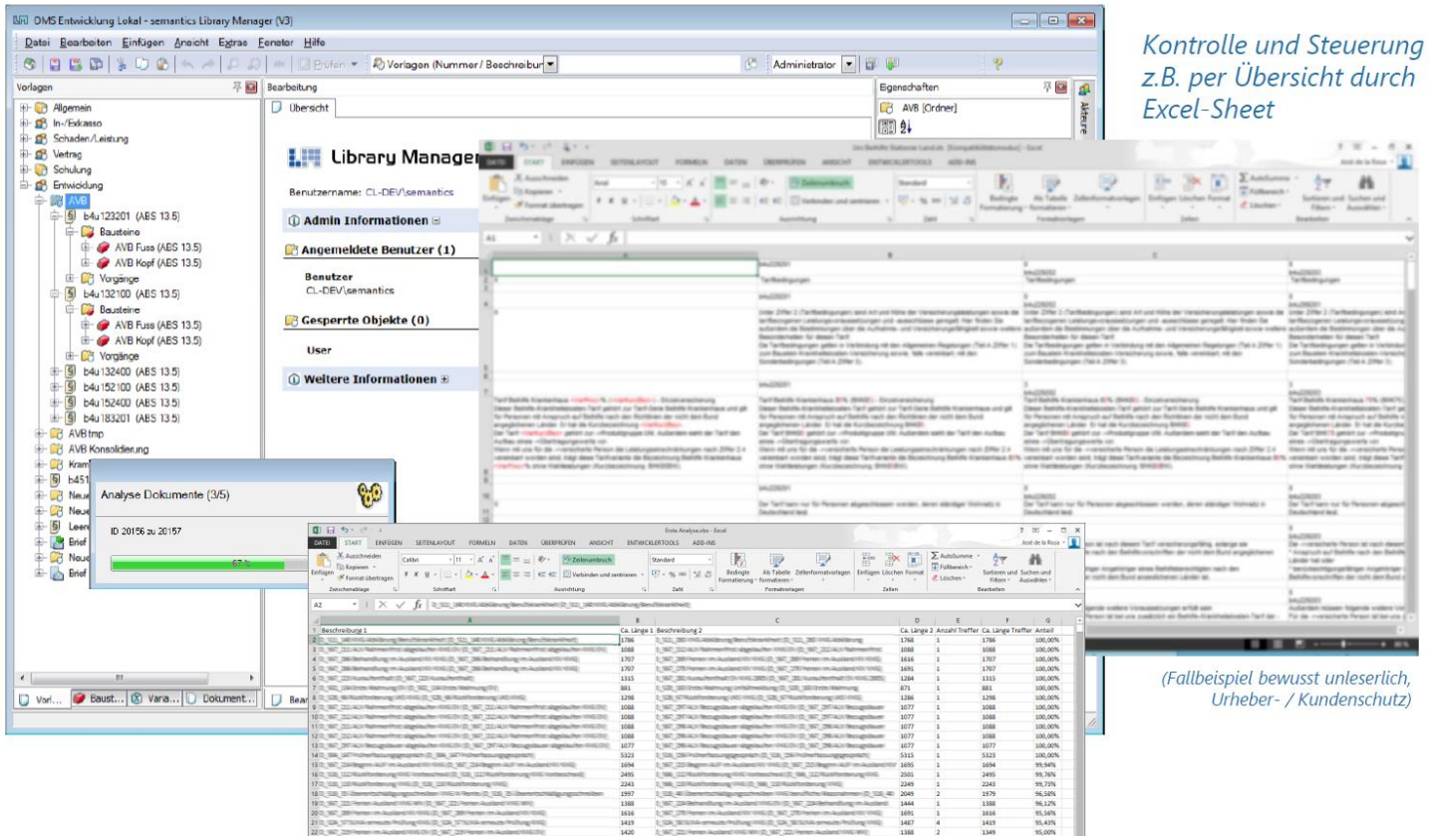
- Herausforderung: Tausende Ergebnisse visualisieren
- Frühe Entscheidung: Zweistufiges Verfahren
  - > Gruppenweise Konsolidierung (Tarif-, Produktgruppen)
  - > Nachträgliches 2. Konsolidieren Textbausteine
- Sorge ob fremder Formate (spitze Klammern…)
  - > Fachabteilung ohne technische Kenntnisse
  - > Konzept mit Excel (andere möglich)
- Visualisierung „Treffer“ als Zeilen im Sheet
  - > Nur Inhalt, keine Formatierung
  - > Angabe % Abweichung
- Zusätzlich Ranking Dokumentähnlichkeit möglich



Import und Analyse – Zusammenfassung

- Für Fachabteilung einfaches, effizientes Verfahren
- Import der Dokumente in Library Manager (LM)
- Analyse von Gruppen durch LM

- Bereit für weitere Verarbeitung: Erzeugung Textbausteine zur Konsolidierung



Kontrolle und Steuerung z.B. per Übersicht durch Excel-Sheet

(Fallbeispiel bewusst unleserlich, Urheber- / Kundenschutz)

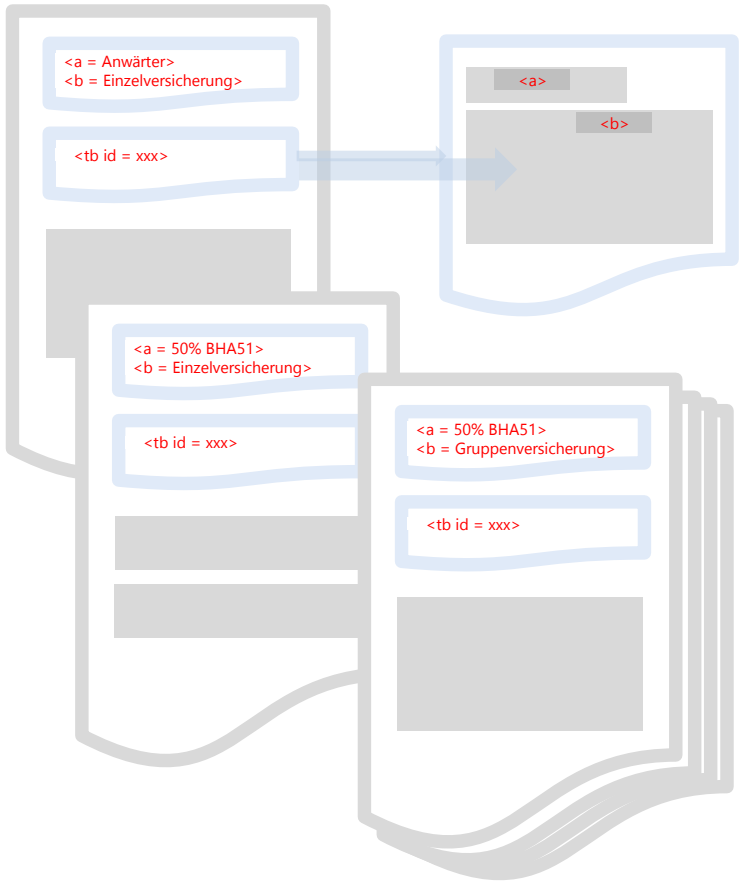
Automatisierung der Redaktionsarbeit

Maschinelle Konsolidierung

Konsolidierung – Vorgang

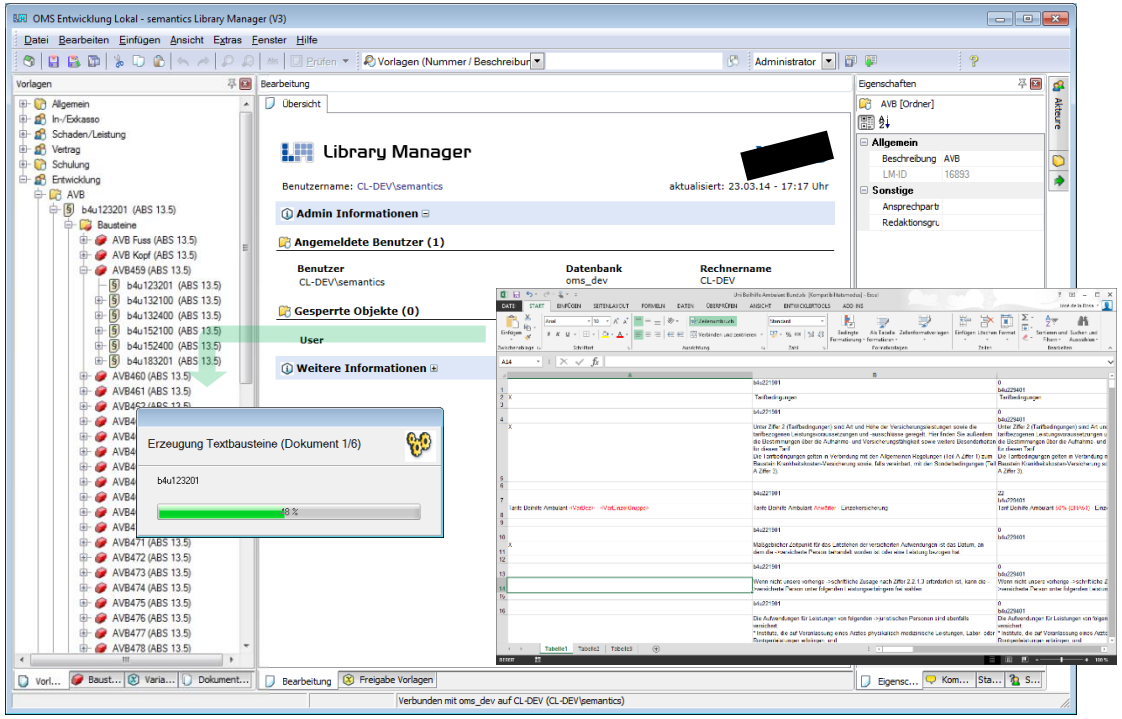
- Bausteine anlegen
  - > Inhalt aus Dokument kopieren
  - > Variablen einbetten
- In Dokument Texte durch Bausteinreferenzen ersetzen
- In Dokument Werte für Variablen eintragen
- Ersetzungen und Werte in allen Dokumenten

- Aufgrund Masse maschinelle Durchführung notwendig
- Fachliche Steuerung und Entscheidung gefordert
  dabei weiterhin ohne technische Kenntnisse



Konsolidierung – Technik

- Steuerung per Excel
- Kennzeichnung auszulagernder Fragmente
- Bezeichnung gewünschter Variablen
- Punktueller Ausschluss möglich
- Anschließend automatische Verarbeitung durch LM
  - > Fragmente aus Dokument bekannt
  - > Position Variablen aus Excel
  - > Variablenwerte aus Dokumenten gemäß Position in Excel



Ergebnis

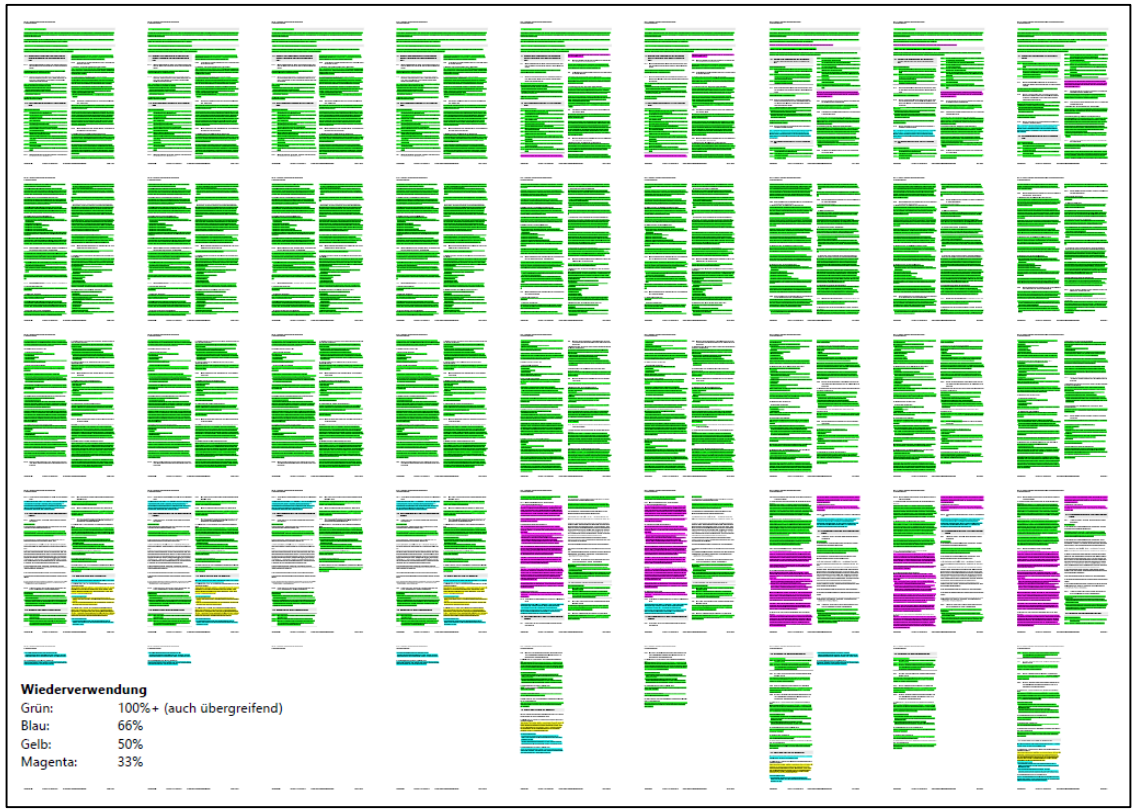
Erwartungen und Ungeplantes

Erwartungen übertroffen

- 99% Abdeckung
- 15.000 Textbausteine (TB) erzeugt
- 180.000 Referenzierungen TB
- 400fache Wiederverwendung relevante TB *(VVG-Reform, …*
  zudem zahlreiche TB in allen Dokumenten)
- 300 Variablen, bis 330 Ausprägungen
- Ersparnis vs. manuellem Projekt ca. 20 Personen*jahre*

Analyse 1 PT / 10 Dokumente	*≈ 250 PT*
Erzeugung je TB 10 Min.	*≈ 315 PT*
Einbettung je Verwendung 10 Min.	*≈ 3.750 PT*

⇒ *Projekt ohne Technik nicht realisierbar*



Ungeplante (positive) Nebeneffekte

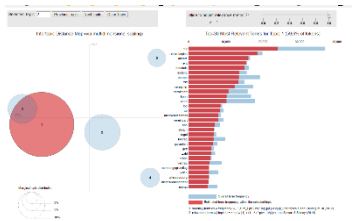
- „Selbstheilung“ großer Textkorpora: Massenvergleich spürt kleine Fehler auf
  - > Interpunktion, Orthographie, Doppellungen, Inkonsistenzen („oder“ statt „bzw.“), …
- Einsparungen und Minimierung Fehlerpotential in Linie durch Library Manager
  - > Einführung LM als strukturiertes Content- und Textsystem inkl. Darstellung Abhängigkeiten, Redaktionsprozessen, Dokumentation, Annotationen, Versions- und Historienführung, Variantenlogik
  - > Ablösung manueller Bereitstellungsverfahren durch fachliche und technische Freigabeprozesse
- Zentrale Bereitstellung AVB für heterogene Systeme
  - > Single-Source Bereitstellung für Altsystem, Kundenportal, Produktives OMS, …
- Belobigung durch Abteilung „Interne Revision“
  - > Turnusmäßige Untersuchung führte zu besonderer Auszeichnung der Arbeitsweise der verantwortlichen Abteilung

Status Quo und Ausblick

Anwendungen und Entwicklungen

Ausblick – Entwicklungen

- Neue Herausforderungen: Analyse von 24.000 Dokumenten
- Clustering der Dokumente z.B. mit „Latent Semantischer Analyse“ (LSA) oder „Latent Dirichlet Allocation“ (LDA), …



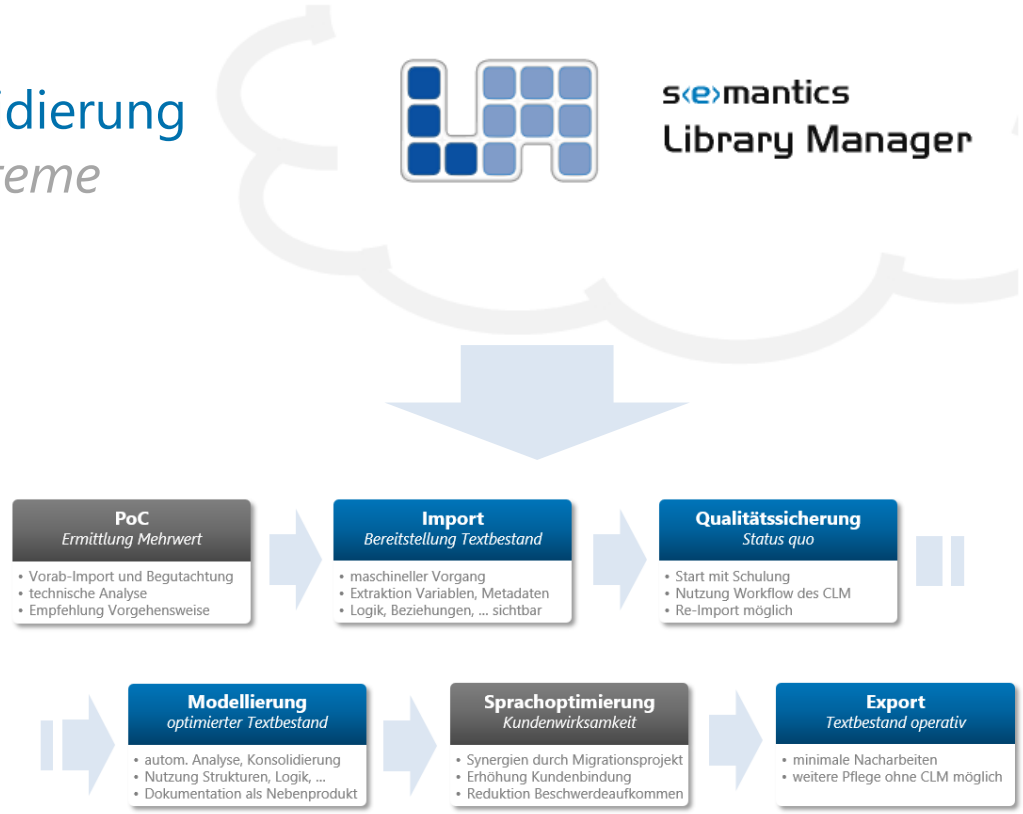
	doc1	doc2	doc3	doc4
angabe	2		1	
anschrift	1	1	1	1
arzt	3	1	2	1
bitte		1	1	
datum	1	1	1	1
fahrzeug				3
für	3	2	2	3
ja	2		1	
name	2	2	3	2
nein		1		1
ort	1	1	1	1
pensionsfond	2		2	
person	2	1		
seit		3		
tod			1	
überangebot				
unterschrift		1		
versorgungsb		1	1	
versorgungsv…ltnis	3	2		
vertragspartner	1	2		
wann	1	3	1	2
welche			3	2
werde			2	1
wurde	1	1	1	2
zahlt			2	2

- Kombination LSA und Thesaurus
  - > Semantische Ähnlichkeitssuche
  - > Operativer „Suchindex“ in Redaktionssystem
  - > Suche nach Sekundärbegriffen

- Reverse Engineering der Fachlogik
  - > Auswertung Zusammenhänge höherer Ordnung

Status Quo – Anwendungen

- Standard Dienstleistung: Migration und/oder Konsolidierung
  Migrations- und Konsolidierungstool für beliebige Systeme
- Bereitstellung „as a Service“: RemoteApps, Cloud
  Keine Installation, Infrastruktur etc. notwendig
- Temporärer Einsatz zur Konsolidierung möglich
  Dienstleistungen semantics inklusive



Fragen und Antworten

Besten Dank für Ihre Aufmerksamkeit!

Dipl.-Ing. José Manuel de la Rosa Govantes

semantics Kommunikationsmanagement GmbH
Viktoriaallee 45
52066 Aachen

+49 241 89 49 89 29
j.delarosa@semantics.de
www.semantics.de