Kantonsspital Baden AG
CH-5404 Baden
www.ksb.ch

Bericht

Strukturierte Datenextraktion am Beispiel Krebsdaten

Klassifizierung	nicht klassifiziert
Status	genehmigt zur Nutzung
Projektleiter	Ralph Schmid
Version	1.0
Datum	3. Juli 2025
Autor/Autoren	Ralph Schmid, Claudia Wecker (ETH-Nexus)
Empfänger / Konsument	Bundeskanzlei BK Digitale Transformation und IKT-Lenkung (DTI) Monbijoustrasse 91 CH - 3003 Bern

Änderungsverzeichnis

Version	Datum	Änderung	Autor
1.0	03.07.2025	Version 1.0	Ralph Schmid

Beschreibung

Das Kantonsspital Baden (KSB) muss Krebserkrankungen gemäss KRG an das Krebsregister Aargau melden. Bisher erfolgt die Datenextraktion manuell aus Patientenberichten. Um diesen Prozess zu optimieren und gleichzeitig den Datenschutz zu gewährleisten, entwickelt das KSB eine automatisierte Pipeline. Das Vorhaben prüft rechtliche und technische Herausforderungen, um eine zuverlässige und skalierbare Lösung zu entwickeln, die auch anderen Institutionen im Schweizer Datenökosystem zugutekommt. Dies fördert Standardisierung und sicheren Austausch sensibler Gesundheitsdaten. Die erste Implementierung konzentriert sich auf Brustkrebsdaten, einschliesslich TNM-Codes und G-Grading basierend auf Tumorboard-Berichten.

Kantonsspital Baden AG
CH-5404 Baden
www.ksb.ch

Management Summary

Das Kantonsspital Baden (KSB) hat eine automatisierte Datenextraktions-Pipeline entwickelt, um Krebsdaten gemäss dem Krebsregistrierungsgesetz (KRG) strukturiert aufzubereiten und an das Krebsregister Aargau zu übermitteln. Das Projekt ist die Herausforderung angegangen, strukturierte Daten aus unstrukturierten Quellen wie Textfeldern zu extrahieren.

Zu den Schlüsselkomponenten gehören die Datenmaskierung in einer Trusted Execution Environment (TEE), Flattening von verschachtelten Datenstrukturen, sowie die Extraktion mittels regulärer Ausdrücke (REGEX) und Large Language Models (LLMs). Die erste Implementierung konzentriert sich auf Brustkrebsdaten, insbesondere TNM-Codes und G-Grading basierend auf Tumorboard-Berichten.

Obwohl die Extraktionsqualität hoch ist, bleibt eine manuelle Überprüfung notwendig. Die am KSB entwickelten Lösungen bieten wertvolle Erkenntnisse für das Schweizer Datenökosystem mit dem Datenraum Gesundheit und zeigen effektive Ansätze zur Verarbeitung sensibler personenbezogener Daten in Cloud-Umgebungen sowie zur Extraktion von Informationen aus semi-strukturierten Textblöcken. Die entwickelten Module sind auf Kantonsspital Baden AG GitHub dokumentiert und publiziert.

Inhaltsverzeichnis

1. Abbildungsverzeichnis 4
2. Abkürzungsverzeichnis 4
3. Einleitung 5
3.1. Hintergrund 5
3.2. Zielsetzung 5
3.3. Mehrwert für das Datenökosystem Schweiz 5
4. Problemstellung 6
4.1. Ausgangslage Prozess Projektstart 6
4.2. Anforderungen an die Lösung 7
4.3. Herausforderungen 7
4.4. Umgang Strukturierte Daten Krebsregister Aargau - FHIR 8
4.5. Aufwand Datenextraktion KSB 8
5. Methodik 9
5.1. Lösungsansatz 9
5.2. Entwicklungsschritte 9
5.3. Technologische Basis 9
6. Implementierung 11
6.1. Datenschutz 11
6.2. Priorisierung Daten 11
6.3. Bewertung Technologien 11
6.4. Datenpipeline 13
6.5. Nutzung LLM Extraktion 13
6.6. Tests und Validierung 14
7. Ergebnisse 15
7.1. FHIR Struktur Krebsregister 15
7.2. Evaluation der Lösung 15
7.3. Grenzen der automatisierten Extraktion 15
7.4. Übertragbarkeit auf andere Spitäler im Rahmen Datenökosystem Schweiz 15
7.5. Kosten der Umsetzung 16
7.6. Open Source Code 16
7.7. Datenökosystem und Datenraum Gesundheit – Datenmodelle 16
8. Fazit und Ausblick 17

1. Abbildungsverzeichnis

Abbildung 1: Darstellung Prozess für die Aufbereitung von Daten vom Bericht bis zum Krebsregister 6
Abbildung 2: Maskierungspipeline vom Quellsystem zur Maskierung mit Speicherung in der Konkordanztabelle 11
Abbildung 3: JSON Flattening und Datenextraktion mittels LLM 13

2. Abkürzungsverzeichnis

Abkürzung	Bedeutung
BAG	Bundesamt für Gesundheit
DTI	Digitale Transformation und IKT-Lenkung
FHIR	Fast Healthcare Interoperability Resources
KISIM	Klinik Informationssystem (vom KSB)
KRG	Krebsregistrierungsgesetz
KSB	Kantonsspital Baden
LLM	Large Language Model(s)
ML	Maschinelles Lernen
mCODE	Minimal Common Oncology Data Elements
NLP	Natural Language Processing
OCR	Optical Character Recognition
PII	Personally Identifiable Information (personenbezogene Daten)
REGEX	Reguläre Ausdrücke
TEE	Trusted Execution Environment
TNM-Code	Klassifikation des Tumorstadiums (Tumor, Nodus, Metastase)

3. Einleitung

3.1. Hintergrund

Das Krebsregistrierungsgesetz (KRG; SR 818.33) legt fest, dass Krebserkrankungen in einer umfassenden und einheitlichen Form erfasst werden müssen. Das Kantonsspital Baden (KSB) hat die Verpflichtung, die entsprechenden Daten an das Krebsregister Aargau zu übermitteln.
Die Quelldaten liegen grösstenteils als unstrukturierte Informationen in Textfeldern oder PDF-Dateien vor. Diese Daten können nicht direkt für das Krebsregister verwendet werden und erfordern derzeit manuelle Aufbereitung.
Um diesen Prozess zu optimieren, plant das Kantonsspital Baden den Einsatz einer eigens entwickelten Pipeline. Diese soll die automatische Extraktion von Daten aus den Quellen ermöglichen und automatisch eine Aufbereitung durchführen, um den Anforderungen des Krebsregisters zu genügen.
Im Zuge dieses Vorhabens wird geprüft, welche rechtlichen und technischen Herausforderungen zu bewältigen sind, um die spezifische Pipeline des KSBs anderen interessierten Parteien/Spitälern zugänglich zu machen. Von Anfang an wird in enger Abstimmung mit den zuständigen Behörden darauf geachtet, dass bei der Implementierung mögliche CH-Datenaustauschstandards berücksichtigt werden.

3.2. Zielsetzung

Die Zielsetzung dieses Projekts ist es, die technischen Möglichkeiten zur Datenextraktion aus den Klinikinformationssystemen des Kantonsspitals Baden (KSB) aufzuzeigen und diese in strukturierter Form für die Übermittlung an das Krebsregister bereitzustellen. Im Vordergrund steht dabei die technische Implementierung einer robusten Pipeline, die auch den Umgang mit sensiblen Gesundheitsdaten in einer Cloud-Umgebung ermöglicht. Dieser Bericht beleuchtet die technischen Herausforderungen und die entwickelten Lösungsansätze zur effizienten und präzisen Datenextraktion. Datenschutzanforderungen und das medizinische Berufsgeheimnis bilden hierbei die unverhandelbaren Rahmenbedingungen für die technische Konzeption und Umsetzung.

3.3. Mehrwert für das Datenökosystem Schweiz

Die Herausforderung der strukturierten Datenübermittlung betrifft nicht nur das KSB, sondern alle Spitäler, sowie die kantonalen Krebsregister selbst. Diese müssen unstrukturierte Daten effizient aufbereiten, um sie einheitlich zu erfassen und weiterzuverarbeiten.

Das KSB entwickelt eine automatisierte Pipeline zur Datenextraktion und Strukturierung, die den hohen Anforderungen an Datenschutz und Sicherheit entspricht. Die gewonnenen Erkenntnisse und erprobten Lösungen sollen innerhalb des Datenökosystems mit anderen Spitälern und Institutionen geteilt werden. Dies fördert die Standardisierung und den sicheren Austausch sensibler Gesundheitsdaten, was langfristig die Digitalisierung des Gesundheitswesens in der Schweiz voranbringt.

4. Problemstellung

4.1. Ausgangslage Prozess Projektstart

Im Folgenden wird der Prozess zu Beginn und des Projekts dargestellt.

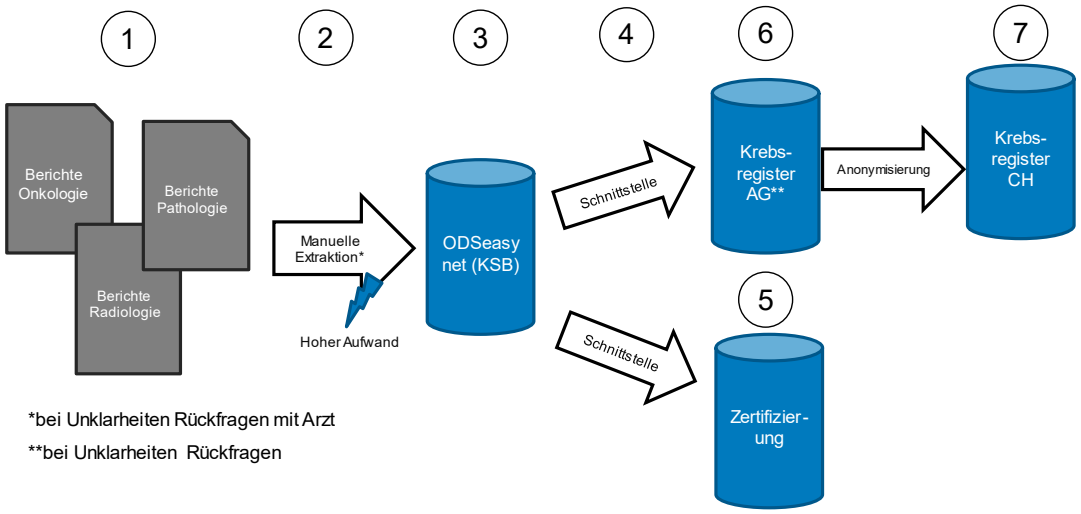

Abbildung 1: Darstellung Prozess für die Aufbereitung von Daten vom Bericht bis zum Krebsregister

1. Erfassung der Medizinischen Daten in den Quellsystemen. Insbesondere in KISIM (Klinik Informationssystem vom KSB)
2. Extraktion der Daten manuell durch spezialisierte Tumordokumentarinnen. Dabei verarbeitet jede Tumordokumentarin mehrere Tumorarten, auf welche sie spezialisiert ist.
3. Die Daten werden in ODSeasyNet eingetragen. ODSeasyNet unterstützt die Dokumentarinnen im Ausfüllen der richtigen Daten. Das System wird für die Übermittlung von zertifizierungsrelevanten Tumordaten verwendet. Es unterstützt auch die Übermittlung der Daten an das Krebsregister.
4. Daten werden als XML-File an das Krebsregister gesendet. Zusätzlich werden diverse Berichte in unstrukturierter Form mitgeschickt, um dem kantonalen Krebsregister mehr Kontext zu den einzelnen Krebsfällen zu geben.
5. Für das KSB ist die Zertifizierung der onkologischen Therapien sehr wichtig. Daher wird in die Aufbereitung der Daten für die Zertifizierungen viel Wert gelegt. Ein grosser Teil der Daten sind mit denen für das Krebsregister identisch.
6. Das Krebsregister dokumentiert die Fälle in den eigenen Systemen. Siehe hier auch Kapitel 4.4 Umgang Strukturierte Daten Krebsregister Aargau
7. Das Krebsregister sendet die Daten anonymisiert an das Krebsregister Schweiz. Der genaue Prozess wird nicht weiter beleuchtet.

Dieser Prozess hat sich im Laufe des Projekts verändert. Es werden weniger Daten in strukturierter Form an das Krebsregister geliefert als noch zu Beginn des Projekts. Die Hintergründe dazu finden sich im Kapitel 4.4 Umgang Strukturierte Daten Krebsregister Aargau.

4.2. Anforderungen an die Lösung

Die geplante Lösung zur automatisierten Datenextraktion und -weitergabe an das Krebsregister muss spezifische Anforderungen erfüllen, um die Effizienz und Genauigkeit der Tumordokumentation erheblich zu steigern:

- Reduktion der Arbeitslast in der Dokumentation
- Automatisierung und Effizienzsteigerung im gesamten Prozess
- Die Qualität der extrahierten Daten muss analog IST oder besser sein
- Kompatibilität mit bestehenden Systemen – KISIM, ODSeasyNet, Datenmanagementplattform KSB
- Konformität mit Datenschutzrichtlinien und Berufsgeheimnis
- (Strukturierte und standardisierte Extraktion – FHIR – siehe Kapitel 4.4)
- Zuverlässigkeit und Skalierbarkeit

4.3. Herausforderungen

Die Umsetzung einer Lösung zur automatisierten Datenextraktion und strukturierten Weitergabe an das Krebsregister steht vor mehreren zentralen Herausforderungen:

Unstrukturierte Natur der Daten
Die klinischen Daten liegen oft in unstrukturierter Form vor, beispielsweise als Freitext in Arztberichten oder in gescannten PDF-Dokumenten. Diese Formate erschweren die maschinelle Verarbeitung und erfordern aufwendige Verfahren zur Datenextraktion und -strukturierung.

Fehlende oder ungenaue Daten
Fehlende/Doppelte oder von der Ärzteschaft ungenau erfasste Daten stellen eine grosse Herausforderung für die strukturierte Erfassung von Krebserkrankungen dar, da sie die Qualität und Vollständigkeit der Registereinträge beeinträchtigen.

Hohe Variabilität in der medizinischen Terminologie
Medizinische Fachbegriffe, Abkürzungen und Synonyme variieren stark, selbst innerhalb des Kantonsspital Baden. Dies führt zu Inkonsistenzen, die eine standardisierte Datenverarbeitung und -klassifizierung erschweren.

Komplexität und Interpretationsbedarf bei Tumordaten
Die korrekte Klassifizierung von Tumoren setzt nicht nur detaillierte medizinische Informationen voraus, sondern auch ein tiefes Verständnis der zugrunde liegenden klinischen Kontexte. Automatisierte Systeme müssen daher in der Lage sein, komplexe Zusammenhänge zu erkennen und zu interpretieren, um Fehler zu vermeiden. Ein Beispiel ist die untenstehende Abbildung von mehrfachen Tumoren.

Abbildung von mehrfachen Tumoren
Ein weiterer Aspekt betrifft die Abbildung von Fällen mit mehreren Tumoren. Die medizinische Behandlung erfolgt in der Regel fallbezogen, wobei ein Fall mehrere Tumoren umfassen kann, etwa bei multifokalen oder multizentrischen Tumorerkrankungen. Diese werden häufig gemeinsam in einem einzigen medizinischen Bericht dokumentiert. Für die Weiterleitung der Daten an das kantonale oder nationale Krebsregister müssen jedoch die multizentrischen Tumore einzeln gemeldet und eindeutig identifizierbar zugeordnet werden (eindeutiger TNM-Code). Die Herausforderung besteht darin, aus einem zusammengefassten Bericht die relevanten Informationen pro Tumor korrekt zu extrahieren, strukturiert aufzubereiten und jeder Tumorerkrankung separat zuzuweisen. Eine präzise Trennung und Zuordnung der Daten ist dabei essenziell, um unvollständige oder fehlerhafte Meldungen sowie potenzielle Mehrfachmeldungen zu vermeiden.

Anforderungen an Datenschutz und Datensicherheit
Besonders schützenswerte Gesundheitsdaten unterliegen strengen rechtlichen und ethischen Anforderungen. Die sichere Verarbeitung und Speicherung, insbesondere im Cloudumfeld, stellt hohe Ansprüche an Verschlüsselung, Zugriffskontrolle und Protokollierung, um sowohl Datenschutz als auch Datensicherheit zu gewährleisten.

Diese Herausforderungen verdeutlichen die Komplexität des Vorhabens und unterstreichen die Notwendigkeit einer sorgfältig entwickelten, technisch ausgereiften und rechtlich abgesicherten Lösung.

4.4. Umgang Strukturierte Daten Krebsregister Aargau - FHIR

Im Rahmen der obligatorischen Krebsregistermeldung wurde das Austauschformat Krebsregistergesetz (KRG) festgelegt. Dabei können die Informationen als unstrukturierte Berichte oder als FHIR-Datei gemäss eigenem FHIR Profil CH.FHIR.IG.CH-CRL\Home - FHIR v4.0.1 an das Krebsregister zugestellt werden.

Das Krebsregister Aargau hat jedoch heute keine Möglichkeit die FHIR Datei zu verarbeiten oder zu lesen. Auch ist uns bis heute keine Gesundheitseinrichtung bekannt, die in der Lage ist die Daten in strukturierter Form im FHIR Profil zu liefern.

Das führt dazu, dass alle Melder die Krebsregistermeldung mittels unstrukturierten Berichtes erledigen und das Krebsregister selber die notwendigen Daten aus diesen extrahiert. Für die richtige Extraktion und Interpretierung der Daten durch das Krebsregister ist eine grosse Anzahl an Daten notwendig, welche auch Informationen enthalten kann, welche nicht krebsrelevant sind. Hier hat sich das Krebsregister verpflichtet, die nicht benötigten Informationen zu vernichten. Für diese Extraktion wurde beim Krebsregister entsprechende Organisationen aufgebaut.

Update vom 24.01.2025
Gemäss der Pressemitteilung vom BAG «Landesweite Erfassung Krebsfälle: Bund prüft neue Lösung» (Quelle: Landesweite Erfassung Krebsfälle: Bund prüft neue Lösung) wird das nationale Projekt zur Beschaffung einer Landesweiten Lösung für die Erfassung der Krebsfälle eingestellt.
Damit ist derzeit nicht absehbar, ob und wann die kantonalen Register in der Lage sein werden, strukturierte FHIR-Daten zu verarbeiten. Daher hat das KSB entschieden, den Datenexport in FHIR-Struktur vorerst nicht weiterzuführen. Die Daten werden jedoch weiterhin in der internen KSB-Datenplattform in einer FHIR-ähnlichen Struktur gespeichert, sodass ein Export jederzeit möglich bleibt.

4.5. Aufwand Datenextraktion KSB

Die Datenextraktion und -aufbereitung für die strukturierte Lieferung an das Krebsregister erfordert einen erheblichen Aufwand, ohne dass dabei ein direkter Mehrwert für das KSB entsteht. Im Gegensatz dazu bietet die Datenaufbereitung im Rahmen der Zertifizierung einen klaren Nutzen. Aus diesem Grund wurde die Entscheidung getroffen, die Datenextraktion für das Krebsregister nur noch für die Tumorarten zu machen, welche auch Zertifizierungsrelevant sind. Für alle anderen Fälle wird die Extraktion der Daten dem Krebsregister überlassen.
Für die Extraktionspipeline stellt dies jedoch keine Änderung dar.

5. Methodik

5.1. Lösungsansatz

Zur Optimierung der Datenextraktion für das Krebsregister sollen unterschiedliche Lösungsansätze untersucht werden. Der Schwerpunkt liegt dabei auf der automatisierten Detektion und Extraktion relevanter Informationen aus unstrukturierten Textdaten.
Dafür werden folgende methodische Ansätze betrachtet:

- Natural Language Processing (NLP): Einsatz von Sprachverarbeitungstechnologien zur Identifikation und Extraktion medizinisch relevanter Entitäten wie Diagnosen, Tumorstadien und Therapien.
- Maschinelles Lernen (ML): Training von Modellen zur Klassifizierung und Strukturierung der extrahierten Daten.
- Regelbasierte Methoden: Nutzung vordefinierter Regeln zur Erkennung und Extraktion spezifischer medizinischer Informationen, z.B. REGEX.
- Optical Character Recognition (OCR): Extraktion von Text aus gescannten Dokumenten oder PDF-Dateien zur weiteren Verarbeitung.
- Vergleich ganzer Krankenakten vs. Textchunks: Untersuchung, ob eine Verarbeitung ganzer Krankenakten oder die gezielte Extraktion relevanter Textabschnitte (Textchunks) effektiver ist.

Durch die Kombination dieser Ansätze soll eine robuste Pipeline entwickelt werden, die die Datenqualität und Verarbeitungsgeschwindigkeit verbessert und den manuellen Aufwand reduziert.

5.2. Entwicklungsschritte

Die Entwicklung der Pipeline erfolgt iterativ in mehreren Schritten:

- Analyse der Datenquellen und Datenstruktur: Untersuchung der bestehenden Datenformate (Textfelder, PDFs) und relevanter medizinischer Terminologien.
- Auswahl der Technologien und Werkzeuge: Evaluierung geeigneter NLP-Frameworks (z. B. spaCy, Transformers), OCR-Technologien (z. B. Tesseract, Google Vision API) und Datenverarbeitungsmethoden.
- Erstellung eines Datenschemas: Definition eines strukturierten Datenmodells für die Speicherung und Weiterverarbeitung.
- Validierung Datenextraktion: Die extrahierten Daten werden in einem ersten Schritt manuell validiert, bevor die automatisierte Pipeline gebaut wird.
- Implementierung und Integration: Entwicklung der Pipeline unter Berücksichtigung bestehender IT-Infrastrukturen wie KISIM und ODSeasyNet.
- Testing und Validierung: Überprüfung der Datenqualität durch Vergleich mit manuellen Extraktionsergebnissen und kontinuierliche Optimierung der Algorithmen.
- Einsatz und Skalierung: Integration der Lösung in den produktiven Betrieb und Skalierung zur Abdeckung weiterer Tumorarten.

5.3. Technologische Basis

Die Implementierung der Pipeline stützt sich auf moderne Technologien zur Texterkennung, Verarbeitung und Speicherung. Dabei wird für die Maskierung der PII-Daten vergleichbare Komponenten wie für die Extraktion der Daten gesetzt:

- NLP-Modelle: Nutzung von Deep-Learning-Ansätzen zur Erkennung und Extraktion relevanter Entitäten (z. B. Named Entity Recognition für medizinische Begriffe).
- LLM: Nutzung von kleinen Sprachmodellen, um spezifische Aufgaben zu erfüllen.
- Regelbasierte Systeme: Ergänzung der ML-Modelle durch festgelegte Regeln zur Erhöhung der Präzision.
- Datenbanken und Schnittstellen: Speicherung der strukturierten Informationen in einer relationalen oder dokumentenbasierten Datenbank sowie Bereitstellung von Exportmöglichkeiten im FHIR-Format.

Kantonsspital Baden AG
CH-5404 Baden
www.ksb.ch

- Cloud- und On-Premise-Architekturen: Sicherstellung der datenschutzkonformen Verarbeitung und Speicherung sensibler Gesundheitsdaten.
- OCR-Technologien: Extraktion von Text aus gescannten Dokumenten und PDFs, um unstrukturierte Daten maschinenlesbar zu machen.

Durch die Kombination dieser Technologien kann eine effiziente und datenschutzkonforme Lösung für die automatische Verarbeitung medizinischer Daten geschaffen werden.

6. Implementierung

6.1. Datenschutz

Der Schutz sensibler Gesundheitsdaten hat höchste Priorität. Die Implementierung der Datenverarbeitung erfolgt unter strikter Einhaltung gesetzlicher Vorgaben und Datenschutzrichtlinien. Dabei werden Verschlüsselungstechniken, Zugriffskontrollen und Protokollierungsmechanismen eingesetzt. Um die Sicherheit der Patientendaten zu gewährleisten, wird auf die folgende Maskierungspipeline gesetzt:

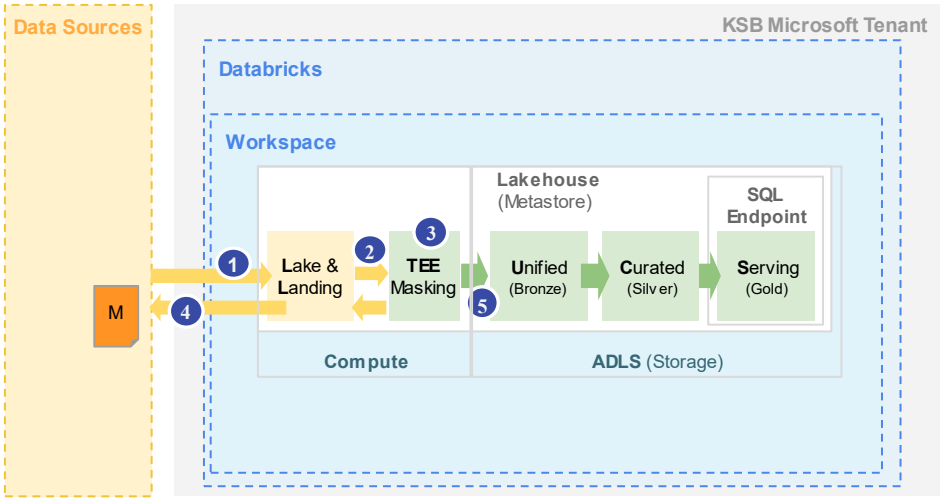

Abbildung 2: Maskierungspipeline vom Quellsystem zur Maskierung mit Speicherung in der Konkordanztabelle

1. Data loading from on-premise to Databricks cluster (in-memory/disk)
2. Data loading from Databricks cluster to Confidential Computing – Trusted Execution Environment (TEE)
3. Data masking in TEE
4. Long-term storage of concordance table M on-premise
5. Masked data encrypted on lake & landing and all subsequent data layers

Die Maskierung erfolgt mittels einer SHA-256 Hash Funktion mit Salt und ist nicht reversibel.
Die Datensicherheit und Datenspeicherung in Azure mittels double Key-Encryption richtet sich nach Best Practices und ist hier nicht weiter beschrieben.

6.2. Priorisierung Daten

Es werden vorrangig Daten extrahiert, die für die Tumordokumentation und Zertifizierung von onkologischen Therapien relevant sind, in der Entwicklungsphase der Pipeline sind dies insbesondere Brustkrebsdaten.

6.3. Bewertung Technologien

Im Zuge der Evaluation der einzusetzenden Technologien wurden die folgenden Erkenntnisse gewonnen.

NLP-Modelle
Der Einsatz grosser, genereller Sprachmodelle (z.B. OpenAI Chat 4o mit 1,8 Billionen Parametern) zur Durchsuchung und Strukturierung ganzer Krankenakten nach dem Ansatz „Prompt and Pray“ erscheint zunächst verlockend und zeigt auf den ersten Blick gute bis sehr gute Ergebnisse. Diese Modelle können umfangreiche Textmengen schnell verarbeiten und relevante Informationen extrahieren. Allerdings erweist sich die Sicherstellung konsistenter Abfragen und die Fehlersuche als schwierig. Zudem sind die Kosten solcher Modelle bei Skalierung sehr hoch. Stand 04.03.2025 liegen die Kosten auf Azure bei $2,50 pro 1 Million Tokens für den Input und $10,00 pro 1 Million Tokens für den Output. Auch

sind die Rechenzentrumskapazitäten in Europa begrenzt, was die Beschaffung ausreichender, bezahlbarer Ressourcen erschwert. Aus diesen Gründen werden diese Modelle in der Implementierung nicht eingesetzt.

LLM
Der Einsatz kleinerer Sprachmodelle (z.B. Llama 70b mit 70 Milliarden Parametern) zur Erfüllung spezifischer Aufgaben hat sich als gute Alternative zu grossen Modellen erwiesen. Diese Modelle sind deutlich günstiger und das Open-Source-Modell Llama 70b ist leichter verfügbar. Das Preis-Leistungs-Verhältnis ist deutlich besser, da sie spezifische Aufgaben effizient und kostengünstig erledigen können. Daher setzen wir im Projekt auf dieses Modell. Stand 04.03.2025 liegen die Kosten auf Azure/Databricks bei $0,66 pro 1 Million Tokens für den Input und $1,97 pro 1 Million Tokens für den Output.

Regelbasierte Systeme
Regelbasierte Abfragen wie REGEX werden für einfach zu extrahierende Daten verwendet, deren Form konstant bleibt. Diese Methoden sind leistungsstark, kostengünstig und bieten konsistente Ergebnisse. Sie eignen sich besonders gut für strukturierte Daten, bei denen die Formatierung gleichbleibend ist. Für Abfragen mit variierenden Ergebnissen werden hingegen LLMs eingesetzt, da diese flexibler auf unterschiedliche Datenformate reagieren können.

Datenbanken und Schnittstellen
Die Daten werden auf der Datenplattform Databricks gespeichert. Die Struktur entspricht der Medaillon-Architektur, wobei die Daten in mehreren Schritten strukturiert werden. Der Gold Layer ist an die FHIR-Struktur angelehnt, was eine flexible Kombination von Datenbanken und Tabellen sowie die Bereitstellung der Daten für verschiedene Anwendungsfälle ermöglicht. In diesem Fall dient die Datenbereitstellung dem Krebsregister, wodurch eine effiziente und flexible Nutzung der Daten gewährleistet wird.

Cloud- und On-Premise-Architekturen
Durch die Maskierung aller PII-Informationen im Trusted Execution Environment (TEE) in Azure wird die Nutzung von Cloudtechnologien bei nicht-europäischen Anbietern ermöglicht, während die Daten weiterhin in der Schweiz gespeichert bleiben. Die Cloud bietet den Vorteil, neue Technologien zur Datenstrukturierung einzusetzen, ohne grosse lokale Infrastruktur bereitstellen zu müssen. Diese wäre in der aktuellen Situation mit begrenzten Ressourcen nur schwer zu gewährleisten. Daher wird die gesamte Pipeline in der Cloud betrieben, was eine skalierbare und flexible Lösung darstellt.

OCR-Technologien
OCR ist zunächst nicht notwendig, da der Fokus auf eigenen Berichten liegt, die als Textblöcke in den Datenbanken verfügbar sind. Diese Berichte sind bereits in einem semi-strukturierten Textformat vorhanden, wodurch der Einsatz von OCR-Technologien überflüssig wird. OCR wäre aber denkbar für den Einsatz von externen Berichten.

6.4. Datenpipeline

Die gesamte Datenpipeline ist in Abbildung 2 dargestellt. Im Nachfolgenden wird auf die eigentliche Datenaufbereitung eingegangen.

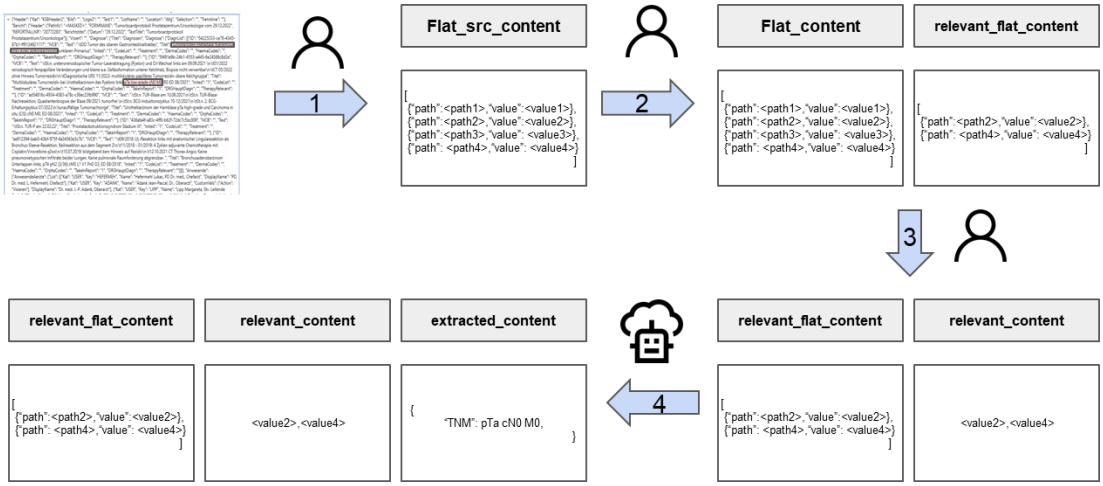

Abbildung 3: JSON Flattening und Datenextraktion mittels LLM

1. Glätten JSON aus der Rohdatenquelle

Die Daten werden aus KISIM importiert. Die KISIM-Berichte liegen als tief verschachtelte JSON-Dateien vor.
Mittels eines entwickelten Algorithmus wird diese komplexe Struktur vereinfacht (geglättet) und in einer neuen Tabelle als Key-Value-Paar gespeichert.

2. Manuelle Aufbereitung der Rohdatenquelle

Die "Flat_src_content"-Daten werden manuell kontrolliert und bei Bedarf bereinigt oder korrigiert.
Das Ergebnis dieser Aufbereitung wird in der Tabelle "Flat_content" abgebildet, welche eine verfeinerte Version der ursprünglichen, vereinfachten Daten darstellt.

3. Manuelle Auswahl relevante Daten

Aus der Tabelle "Flat_content" werden manuell die relevanten "Pfad"- und "Wert"-Paare ausgewählt.
Diese ausgewählten Daten werden in einer neuen Tabelle "relevant_flat_content" gespeichert.

4. Nutzung LLM

Durch die Formulierung präziser Prompts werden die gewünschten Informationen aus den "relevant_flat_content"-Daten extrahiert.
Die extrahierten Informationen werden in einer neuen Tabelle "relevant_content" oder "extracted_content" gespeichert, je nach Art der extrahierten Daten.

6.5. Nutzung LLM Extraktion

Fokus auf reine Datenextraktion

Der primäre Zweck des LLM-Einsatzes liegt in der reinen Extraktion von Daten.
Es erfolgt keine komplexe Logikimplementierung, wie beispielsweise die Zuordnung von Werten zu spezifischen Kodierungen oder das Durchführen von Datenmapping.
Die Prompts, die dem LLM übermittelt werden, sind auf maximale Präzision ausgelegt.
Es werden ausschliesslich die für die Extraktion notwendigen Textsegmente an das LLM gesendet, um die Effizienz zu steigern und unnötige Verarbeitung zu vermeiden.

Kostenaspekte und Datenvolumen
Obwohl die Kosten für die LLM-Nutzung im Allgemeinen relativ günstig sind, können sie bei der Verarbeitung grosser Datenmengen und einer hohen Anzahl von Datensätzen signifikant ansteigen.
Es ist daher entscheidend, die Datenmenge, die dem LLM zur Verarbeitung übergeben wird, auf das absolute Minimum zu beschränken, um die Kosten zu kontrollieren.

Datenqualität und Rahmenbedingungen
Das LLM bietet die Möglichkeit, Rahmenbedingungen für die erwarteten Ausgaben festzulegen.
Dies ermöglicht die Durchsetzung von Datenqualitätsstandards, indem beispielsweise das Format der erwarteten Ergebnisse spezifiziert wird.
Sollten die Ausgaben des LLM nicht den definierten Rahmenbedingungen entsprechen, wird die Anfrage automatisch wiederholt, um die Datenqualität sicherzustellen.

Asynchrones Aufgabenmanagement für hohe Performance
Um eine optimale Performance zu erzielen, wird ein asynchrones Aufgabenmanagement für die Kommunikation mit dem LLM implementiert.
Dieses asynchrone Vorgehen ermöglicht die parallele Verarbeitung mehrerer Anfragen, wodurch die Durchlaufzeit reduziert und die Effizienz gesteigert wird.

6.6. Tests und Validierung

In der ersten Projektphase kann aus Projektsicht noch nicht vollständig sichergestellt werden, dass die extrahierten Daten korrekt sind. Die Datensätze erscheinen zwar korrekt, bestehen jedoch nicht immer eine fachliche Überprüfung. Dies macht eine automatische Übermittlung an ODSeasyNet risikoreich, da dort nur noch eine oberflächliche Validierung durch Mitarbeitende erfolgt.

Aufgrund dieser Qualitätsbedenken wurde entschieden, in dieser Phase keine Daten direkt an ODSeasyNet zu übermitteln. Stattdessen wird ein übersichtliches Dashboard bereitgestellt, das alle relevanten Informationen einschliesslich der Originalquellen präsentiert. Dieser Ansatz verbessert die Prozesseffizienz insofern, als dass nicht mehr zahlreiche einzelne Berichte geöffnet werden müssen, um Informationen zu lokalisieren und zu überprüfen.
Die Validierung dieser Daten erfolgt durch die Tumordokumentarinnen selbst. Schrittweise werden weitere Datenelemente hinzugefügt und validiert. Sobald erste Daten in validierter Form vorliegen, wird die automatisierte Pipeline zur Datenübermittlung an ODSeasyNet implementiert.

7. Ergebnisse

7.1. FHIR Struktur Krebsregister

Zu Beginn des Projekts wurde die FHIR-Spezifikation des Krebsregistergesetzes eingehend analysiert. Dabei stellte sich heraus, dass die vorhandenen Vorgaben zwar eine grundsätzliche Orientierung bieten, jedoch in vielen Bereichen nicht ausreichend präzise definiert sind. Zahlreiche Details sind nicht eindeutig festgelegt, was zu einem erheblichen Interpretationsspielraum führt.

Ein anschauliches Beispiel für die strukturierte und visuelle Darstellung von FHIR-Daten bietet das Minimal Common Oncology Data Elements (mCODE)-Projekt aus den USA ((Home - minimal Common Oncology Data Elements (mCODE) Implementation Guide v4.0.0). Dieses zeigt, wie eine standardisierte und interoperable Repräsentation onkologischer Daten gelingen kann.

7.2. Evaluation der Lösung

Die Evaluation der Datenextraktionslösung für Brustkrebsbefunde zeigt, dass die aktuelle Lösung bereits erfolgreich Daten aus Brustkrebsbefunden extrahiert, insbesondere den TNM-Code, das G-Grading sowie weitere Extraktionsdaten und Textinformationen aus Tumorboard-Berichten. Die Qualität der extrahierten Informationen ist dabei sehr hoch, allerdings sind die Daten noch nicht ausreichend strukturiert, um eine vollständig automatisierte Verarbeitungspipeline zu ermöglichen. Aus diesem Grund ist aktuell ein manueller Zwischenschritt erforderlich. Dennoch übertrifft der aktuelle Ansatz bereits alle bisherigen Methoden zur Datenextraktion deutlich. Wir sehen in diesem Ansatz ein erhebliches Potential, insbesondere hinsichtlich der Skalierbarkeit auf andere Datenfelder, und sind zuversichtlich, dass wir durch die Weiterentwicklung der Pipeline den manuellen Schritt in Zukunft ersetzen können.

7.3. Grenzen der automatisierten Extraktion

Die automatisierte Datenextraktion, insbesondere bei der Nutzung von Large Language Models (LLMs), stösst auf spezifische Grenzen. Eine wesentliche Herausforderung ist die Extraktion seltener Krebsarten oder selten vorkommender Informationen. Da LLMs auf Trainingsdaten basieren, haben sie Schwierigkeiten, Informationen zu erkennen und zu extrahieren, die im Trainingsdatensatz nur spärlich vertreten sind. Dies schränkt die Fähigkeit ein, seltene Fälle oder spezifische, wenig dokumentierte Informationen zuverlässig zu verarbeiten.

Zusätzlich stellt die bestehende Datenqualität eine bedeutende Hürde dar. Die verstreute Datenlage über verschiedene Berichte erschwert die Extraktion erheblich. Da relevante Daten nicht immer an denselben Stellen innerhalb der Berichte zu finden sind, entstehen Inkonsistenzen. Hier ist eine Nachbesserung der Datenerfassung erforderlich, um die Qualität und Konsistenz der Daten zu verbessern und somit die Effektivität der automatisierten Extraktion zu steigern.

Zusammenfassend lässt sich sagen, dass die automatisierte Extraktion, trotz ihrer Vorteile, durch die Verfügbarkeit und Qualität der Trainingsdaten sowie die Strukturierung der zu extrahierenden Informationen limitiert ist. Um diese Grenzen zu überwinden, sind sowohl Verbesserungen in der Datenerfassung als auch die Weiterentwicklung des Extraktionsprozesses inkl. LLMs notwendig.

7.4. Übertragbarkeit auf andere Spitäler im Rahmen Datenökosystem Schweiz

Die am Kantonsspital Baden (KSB) entwickelten Ansätze zur Datenextraktion, einschliesslich Maskierung, Flattening, REGEX und der LLM-Datenpipeline, bieten das Potenzial, auf andere Spitäler übertragen zu werden. Es ist jedoch wichtig zu betonen, dass eine 1:1-Übertragung aufgrund der unterschiedlichen Datenstrukturierungen von Spital zu Spital nicht ohne Anpassungen möglich ist. Die spezifischen Formate und Ablagestrukturen von medizinischen Berichten und Dokumentationen variieren erheblich, was eine individuelle Anpassung der Extraktionsprozesse erfordert.

Ein besonders vielversprechender Aspekt ist die Extraktion von Daten mittels Large Language Models (LLMs). Dieser Ansatz kann weitgehend 1:1 verwendet werden, sofern die Input-Daten in einer ähnlichen Weise vorliegen wie am KSB.

Die Fähigkeit von LLMs, Muster und Zusammenhänge in unstrukturierten Textdaten zu erkennen, ist unabhängig von spezifischen Datenformaten.

Um eine erfolgreiche Übertragung zu gewährleisten, sind folgende Schritte empfehlenswert:

- Analyse der Datenstruktur: Eine detaillierte Analyse der vorhandenen Datenstrukturen und -formate im Zielspital ist unerlässlich.
- Anpassung der Vorverarbeitung: Die Vorverarbeitungsschritte, wie Maskierung und Flattening, müssen an die spezifischen Gegebenheiten des Zielspitals angepasst werden.
- Validierung und Optimierung: Nach der Implementierung ist eine gründliche Validierung der Extraktionsergebnisse erforderlich, um die Genauigkeit und Zuverlässigkeit sicherzustellen.
- Durch diese Schritte kann die am KSB entwickelte Datenextraktionspipeline erfolgreich auf andere Spitäler übertragen werden, wodurch eine effizientere und präzisere Nutzung von medizinischen Daten ermöglicht wird.

7.5. Kosten der Umsetzung

Die Kosten für die Implementierung einer vergleichbaren Datenverarbeitungspipeline lassen sich nur bedingt aus den Erfahrungen am Kantonsspital Baden (KSB) ableiten. Dies liegt vor allem an den teils erheblichen Unterschieden in den bestehenden ICT-Landschaften der jeweiligen Institutionen, in die eine solche Lösung integriert werden muss.

Unter der Annahme, dass eine moderne Datenverarbeitungsplattform als Grundinfrastruktur bereits vorhanden ist, ergibt sich folgende grobe Kosteneinschätzung:

- Manuelle Datenaufbereitung: Der grösste Kostenblock entfällt auf die manuelle Extraktion und Strukturierung relevanter Textabschnitte aus den unstrukturierten klinischen Dokumenten. Je nach Umfang und Qualität der Ausgangsdaten ist hier mit Aufwänden zwischen CHF 50'000 und CHF 300'000 oder noch mehr zu rechnen.
- Infrastrukturkosten: Diese fallen im Vergleich moderat aus, insbesondere wenn bestehende Systeme genutzt oder geringfügig erweitert werden können.
- Kosten für den Einsatz grosser Sprachmodelle: Die Ausgaben für LLMs machen im Gesamtprojektbudget den kleinsten Anteil aus, insbesondere bei effizienten Nutzungsstrategien und gezieltem Feineinsatz.

Ähnliches gilt für die Betriebskosten: Je nachdem, welche Komponenten und Leistungen in die Berechnung einbezogen werden, können diese stark variieren – von wenigen Franken pro Monat bis hin zu mehreren tausend Franken.

7.6. Open Source Code

Die vom Kantonsspital Baden (KSB) entwickelten Werkzeuge werden als Open-Source-Software veröffentlicht und sind auf GitHub vom Kantonsspital Baden AG GitHub öffentlich zugänglich gemacht worden.

7.7. Datenökosystem und Datenraum Gesundheit – Datenmodelle

Die am Kantonsspital Baden (KSB) entwickelten Methoden zur Datenextraktion und -verarbeitung bieten wertvolle Erkenntnisse für das geplanten Schweizer Datenökosystem und den Gesundheitsdatenraum, wie auch für Digi Santé. Das KSB hat erfolgreiche Ansätze für die Verarbeitung sensibler Personendaten in Cloud-Umgebungen und die Extraktion von Informationen aus semi-strukturierten Textblöcken umgesetzt.
Die implementierten Technologien wie Maskierung, Flattening, REGEX und LLM-basierte Datenpipelines demonstrieren, wie eine effiziente und sichere Datenverarbeitung gewährleistet werden kann. Diese praktischen Erfahrungen sind besonders wertvoll für die Gestaltung von Datenmodellen und Prozessen im Rahmen von Digi Santé.
Die erfolgreiche Nutzung komplexer und unstrukturierter Datenquellen am KSB zeigt das Potenzial für den Datenraum Gesundheit, in dem verschiedene Datenquellen mit unterschiedlichen Formaten zusammengeführt werden sollen. Die gewonnenen Erkenntnisse können zur Definition von Best Practices und Entwicklung von Interoperabilitätsstandards beitragen.
Durch institutionsübergreifende Zusammenarbeit kann die Entwicklung von Digi Santé beschleunigt werden. Die KSB-Ansätze dienen als Ausgangspunkt, müssen jedoch an die spezifischen Anforderungen des Gesundheitsdatenraums angepasst werden, um höchste Datenschutzstandards zu gewährleisten.

8. Fazit und Ausblick

Fazit

Übersicht der Projektleistungen:
Das Kantonsspital Baden (KSB) hat ein bedeutendes Automatisierungsprojekt zur Extraktion von Krebsdaten implementiert, welches die gesetzlichen Anforderungen des Krebsregistrierungsgesetzes erfüllt und gleichzeitig innovative technische Lösungen einsetzt. Die entwickelte Pipeline transformiert unstrukturierte medizinische Daten in ein strukturiertes Format, das für Krebsregister und Zertifizierungszwecke geeignet ist. Die Pipeline konnte nicht bis zum Krebsregister Aargau fortgesetzt werden, da diese heute nicht in der Lage ist strukturierte Krebsdaten im FHIR-Format zu verarbeiten.

Technische Umsetzung:
Die Lösung verbindet mehrere fortschrittliche Technologien, darunter Datenmaskierung in einer Trusted Execution Environment, JSON-Flattening, reguläre Ausdrücke und Large Language Model (LLM)-basierte Extraktion. Die Datenextraktion mittels LLM erfolgt mit dem kostengünstigen Llama 70b-Modell, das eine ausgewogene Balance zwischen Leistung und Kosteneffizienz bietet. Der entwickelte Code ist als Open-Source-Software veröffentlicht. Die Details sind im Code dokumentiert und auf GitHub publiziert Kantonsspital Baden AG · GitHub.

Herausforderungen und Anpassungen:
Trotz der hohen Extraktionsgenauigkeit bei wichtigen Krebsdatenelementen wie TNM-Codes und G-Grading erfordert das System weiterhin manuelle Überprüfungsschritte aufgrund von Datenqualitätsproblemen und der Komplexität bei der Extraktion von Informationen zu seltenen Krebsarten. Das Projektteam musste sich zudem an sich ändernde Anforderungen anpassen, nachdem der Bund sein nationales Krebsregisterlösungsprojekt eingestellt hatte.

Bedeutung für den Schweizer Gesundheitsdatenraum des Datenökosystem Schweiz:
Die entwickelten Ansätze liefern wertvolle Erkenntnisse für das geplanten Schweizer Datenökosystem, insbesondere in Bezug auf:

- Sichere Verarbeitung sensibler personenbezogener Daten in Cloud-Umgebungen
- Effektive Extraktionstechniken für semi-strukturierte Textdaten
- Praktische Grundlagen für die Entwicklung von Best Practices und Interoperabilitätsstandards

Ausblick

Sicht KSB:
Das Kantonsspital Baden plant, die erfolgreiche Strukturierung der Tumordaten auf weitere Tumorarten auszuweiten. Die kontinuierliche Verbesserung und Optimierung der Extraktionsprozesse stehen dabei im Fokus. Sobald die Qualität der ausgegebenen Daten einen Reifegrad erreicht hat, der keine manuelle Überprüfung mehr erfordert, wird die gesamte Pipeline vollständig automatisiert. Dies wird nicht nur die Effizienz steigern, sondern auch personelle Ressourcen für andere wichtige Aufgaben freisetzen.

Sicht Datenökosystem - Gesundheitsdatenraum:
Für die Generierung eines signifikanten Mehrwerts und die erfolgreiche Automatisierung bestehender Prozesse ist die Implementierung schweizweiter Datenmodellierungsstandards von entscheidender Bedeutung. Diese Standards sollten nicht nur auf der Ebene des Datenaustauschformats etabliert werden, sondern idealerweise bereits bei der primären Datenerfassung Anwendung finden. Nur durch solche grundlegenden Standardisierungen können Datenpipelines effizient von einem Spital zum anderen übertragen werden.

In Abwesenheit entsprechender Grundstrukturen in den Quellsystemen bleibt die Datenstrukturierung für jedes Spital eine hochindividuelle Aufgabe, die erhebliche finanzielle Ressourcen erfordert. Dies führt voraussichtlich dazu, dass viele Implementierungen nicht realisiert werden können, da die meisten Spitäler nicht bereit oder in der Lage sein

Kantonsspital Baden AG
CH-5404 Baden
www.ksb.ch

werden, diese Kosten selbständig zu tragen. Eine koordinierte, nationale Herangehensweise zur Standardisierung würde daher nicht nur die technische Interoperabilität fördern, sondern auch die wirtschaftliche Machbarkeit solcher Datentransformationsprojekte wesentlich verbessern.