Datenbank Archivierung am Beispiel der 2014 durch Loganto ersetzten Einwohnerdatenbank der VRSG (VRSG EK)

Christian Eugster / 1. Juni 2017

Inhaltsverzeichnis

1 Einleitung

Im Mai 2015 wurde eine Arbeitsgruppe, bestehend aus den Stadtarchiven von Bern, Luzern, Zürich und St. Gallen, zusammengestellt mit dem Zweck, in einem gemeinsamen KOST-Projekt die Archivierung von Daten aus dem Einwohnerregister zu untersuchen. Das Projekt wurde in zwei Phasen unterteilt, wobei an der ersten Phase alle beteiligten Stadtarchive aktiv teilnahmen, während die zweite Phase durch das Stadtarchiv St. Gallen durchgeführt wurde.

In der ersten Phase sollten der Istzustand in den Einwohnerämtern evaluiert und die Daten aus archivischer Sicht bewertet, ein Datenmodell für die archivische Aufbewahrung erstellt und der Ablieferungsprozess inklusive Schnittstellen aus dem Quellsystem beschrieben und das ganze in Form eines Konzepts zusammengestellt werden.

Die zweite Phase befasste sich mit der konkreten Umsetzung in Form eines „proofs of concept", in deren Verlauf das Konzept aus Phase eins auf seine Machbarkeit hin überprüft und die konkreten Daten archiviert werden sollten.

Die Arbeitsgruppe zur Archivierung der Einwohnerdaten beschloss, dass Einwohnerdaten unter folgenden Umständen archiviert werden sollten:

1. Ein neues Datenbanksystem löst das bestehende ab.
2. Es erfolgen rechtliche Änderungen, die mit Änderungen am Datenbanksystem verbunden sind.
3. Am Datenbanksystem werden Änderungen vorgenommen, die auf den Datenbestand Einfluss haben.

2014, also noch vor bevor die Arbeitsgruppe diesen Beschluss fassen konnte, war in St.Gallen die Inbetriebnahme von Loganto erfolgt. Das Stadtarchiv St.Gallen entschied, rückwirkend den Beschluss der Arbeitsgruppe umzusetzen und die Vorgängerdatenbank zu archivieren. Die Umsetzung erfolgte mit dem Ziel:

1. allgemeine Erfahrungen mit der Archivierung von Datenbanken zu sammeln;
2. die fragliche Datenbank so zu archivieren, dass zu einem späteren Zeitpunkt, wenn die Datenbank selber nicht mehr in Betrieb ist, Abfragen an ihr durchgeführt werden können.

Im Herbst 2015 wurde mit der Archivierung begonnen. Am 2. Februar 2017 fand eine Besprechung zwischen Christian Eugster und Martin Kaiser von der KOST statt, deren Zusammenfassung Martin Kaiser uns per Email zukommen liess[1]. Dieses Dokument verweist jeweils auf die dort aufgelisteten Punkte.

2 Phase 1: Die konzeptionelle Arbeit: Evaluation, Datenbewertung, Ablieferungsprozess

Das Vorhaben konnte als Projekt der Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST)[2] durchgeführt werden. Die Projektarbeit startete mit der Erstellung eines Projektantrages zuhanden des Steuerungsausschusses der KOST, der am 29. Januar 2015 eingereicht und laut Email

[1] Siehe Anhang 6.2. Das von Martin Kaiser erstellte Kurzprotokoll dieser Sitzung enthält eine gute Zusammenfassung der Problematik.

[2] Informationen über die KOST/CECO sind unter http://kost-ceco.ch/cms/ abrufbar.

vom 9. Februar genehmigt wurde[3]. Während dieser ersten Phase wurden drei Sitzungen durchgeführt, an denen jeweils alle beteiligten Stadtarchive vertreten waren[4].

2.1 1. Sitzung vom 16. Januar 2015

An der ersten Sitzung wurden das Vorgehen diskutiert und die Informationen für den Projektantrag zusammengestellt. Eine IST-Analyse in den Verwaltungen der beteiligten Archive mittels eines Fragebogens[5] sollte Auskunft geben über die in den Einwohnerregisterämtern eingesetzte Software, über die rechtlichen Grundlagen sowohl bezüglich der Archivierung der Einwohnerregisterdaten als auch des Datenschutzes. Ausserdem sollten spezifische Eigenheiten in Bezug auf diese Daten ermittelt und Kontakte mit den betroffenen Behörden/Stellen (Einwohneramt, Informatikamt, etc.) initialisiert werden. Aufgrund der zusammengestellten Unterlagen sollte in der folgenden Sitzung über das weitere Vorgehen diskutiert werden.

2.2 2. Sitzung vom 3. Juli 2015

Während anlässlich der zweiten Sitzung die IST-Analyse durchwegs positiv bewertet wurde, was die Zusammenarbeit mit den Ämtern betrifft, wurde die Tatsache als problematisch erachtet, dass die Bundes- wie auch die kantonalen Gesetze keine Löschung der Daten vorsehen, sondern lediglich eine Deaktivierung, welche unter besonderen Umständen (z.B. bei Wiederauftauchen einer als tot eingetragenen Person) rückgängig gemacht werden kann. Da es so gesehen im Einwohnerregister nie zu Dossierabschlüssen kommt und aktive Daten nicht in Archive übernommen werden sollten, wurde folgender Lösungsweg erarbeitet:

Zwischen den Dienststellen und den Archiven wird mittels einer Ablieferungsvereinbarung festgelegt, dass für die Dienststellen eine Anbietepflicht eintritt, sobald die nachhaltige Verfügbarkeit der Daten nicht mehr gewährleistet ist. Für einen solchen Fall wurden drei mögliche Szenarien vorgeschlagen:

- bei Änderungen von Gesetzen, Verordnungen, Reglementen, welche die Registerdaten betreffen;
- bei Migration der Daten eines bestehenden in ein neues System;
- bei Änderungen innerhalb des bestehenden Systems.

Da durch die Ämter keine ganzen Datensätze gelöscht werden (können), muss der Fokus vor allem auf den Metadaten liegen; dass ein Feld (z.B. Beruf, Notizen) gelöscht wird, wird als wahrscheinlicher angesehen.

Was das erwünschte Datenmodell der archivierten Daten betrifft, so wurde festgelegt, dass alle im Registerharmonisierungsgesetz genannten Daten obligatorisch übernommen werden müssen. Einzelne Metadaten auszuscheiden, macht im Hinblick auf den damit verbundenen Arbeitsaufwand wenig Sinn.

---

[3] Der Projektantrag kann zum Zeitpunkt der Niederschrift dieses Dokuments auf der Seite http://kost-ceco.ch/cms/index.php?id=310,562,0,0,1,0 eingesehen werden. Möglicherweise ändert die URL nach Abschluss des Projekts.

[4] Die erste Setzung fand am 16. Januar, die zweite am 3. Juli und die dritte am 18. September 2015 statt.

[5] Der Fragebogen ist in fünf Bereiche aufgeteilt (Archiv, Einwohneramt, Hersteller, Alle, Datensatz), und umfasst Fragen zu Archivierungsvorgaben, Bewertung, Verzeichnungspraxis, verwendete Systeme, Datenmodell, Schnittstellen und Migration. Die ausgefüllten Fragebögen sind bei den einzelnen beteiligten Stadtarchiven einsehbar.

Die Zusammenstellung möglicher vorhandener Schnittstellen wurde an dieser Sitzung zurückgestellt.

2.3 3. Sitzung vom 18. September 2015

An der dritten Sitzung wurde der zwischenzeitlich ausgearbeitete Entwurf einer Vereinbarung zwischen dem Einwohneramt und dem Archiv bezüglich Datenhaltung der Einwohnerdaten diskutiert und angenommen, mit dem Zusatz, dass Punkt 2 der Vereinbarung in den jeweiligen Städten an die vorhandenen rechtlichen Vorgaben angepasst wird[6]. Mit dieser Entscheidung kann die Phase 1 des Projekts als abgeschlossen gelten.

3 Phase 2: Proof Of Concept: Die Registerdaten des Einwohneramtes St. Gallen

Mit der Phase 2 war, wie bereits erwähnt, nur das Stadtarchiv St. Gallen beschäftigt. Dies vor allem deshalb, weil bei den anderen Archiven im gegebenen Zeitrahmen keine der in der Vereinbarung genannten Bedingungen eingetreten war.

2014 führte die Verwaltungsrechenzentrum St. Gallen AG (VRSG) eine neue Datenbankgeneration für die Bewirtschaftung der Einwohnerdaten ein. Da Loganto, wie das neue Produkt heisst, noch nicht über alle benötigten Schnittstellen verfügt, bleibt die bislang verwendete Datenbank mit der Bezeichnung VRSG EK vorläufig in Betrieb. Die Datenbewirtschaftung erfolgt ausschliesslich in Loganto. Datenänderungen werden in der alten Applikation nachgeführt, und aus dieser werden die Exportdaten für die unterschiedlichen Anwendungen innerhalb der Stadtverwaltung generiert.

3.1 Datenauswahl

Es standen am Anfang drei Optionen zur Auswahl, wie wir an die Daten kommen könnten:

1. Wir archivieren die Daten, welche die Informatikdienste St. Gallen (IDS) nachts erhalten und in ihr System einspielen. Dabei handelt es sich um eine Teilmenge der Daten aus dem Ursprungssystem, die von verschiedenen Amtsstellen benötigt werden. Diese Daten befinden sich in einer relationalen Datenbank (Microsoft SQL-Server).
2. Wir archivieren Loganto selber. Diese Datenbank liegt ebenfalls in einer relationalen Datenbank. Da sie eben erst eingeführt worden ist, verfügt sie noch über wenige Schnittstellen.
3. Wir archivieren das Vorgängersystem VRSG EK. Dieses System liegt nicht als relationale Datenbank vor. Es handelt sich um ein teilrelationales System, das zur Produktivitätssteigerung die Daten nicht in der ersten Normalform[7] ablegt (Adabas).

Wir entschieden uns aus folgenden Gründen für die dritte Variante:

[6] Der Wortlaut des ersten Entwurfs kann im Anhang (Seite 12: Kapitel 6.1) nachgelesen werden.
[7] Daten werden in Datenbanken in Tabellen abgelegt. Die Normalform macht Aussagen über die Datenorganisation. Die 1. Normalform liegt vor, wenn sämtliche Attribute in einer Tabelle atomar vorliegen (zur Normalform siehe https://de.wikipedia.org/wiki/Normalisierung_(Datenbank) (Stand 30.5.2017).

1. Bei VRSG EK handelt es sich um die originalen Daten des Ursprungssystems. Wir können somit eher davon ausgehen, dass keine Konvertierungsfehler vorhanden sind.
2. Diese Daten wurden für die Datenmigration in das neue System exportiert. Der exportierte Datenbestand steht unmittelbar zur Verfügung und kann ohne weiteren Aufwand für die Verarbeitung und Archivierung übernommen werden.
3. Der Datenbestand umfasst die Daten, wie sie zu einem präzise festgelegten Zeitpunkt bestanden haben (nämlich zum Zeitpunkt des Exports). Die Gefahr von Inkonsistenzen ist daher sehr gering.
4. Die Daten liegen in einem gut zu verarbeitenden Format vor. Es handelt sich um 12 Dateien im XML-Format.

Es existiert keine Dokumentation; die VRSG konnten (oder wollten) auch keine oder nur geringe Unterstützung bei der Archivierung bieten.

3.2 Datenanalyse

3.2.1 Fehlende Dokumentation

Da von der VRSG keine Dokumentation für die Daten erhältlich ist, musste eine Analyse der Daten vorgenommen werden. Diese ist Voraussetzung, damit die Daten zu einem späteren Zeitpunkt sinnvoll verwendet und interpretiert werden können. Da die Dateien sehr umfangreich sind (die grösste derselben, RESIDENT.XML, umfasst fast 4 GB), musste ein geeignetes Verfahren gefunden werden, das eine umfassende Analyse ermöglicht. Mit der Erstellung von Schema-Dateien (siehe Seite 8, Erstellen von XML Schema-Dateien als Unterstützung bei der Datenanalyse) schaffen wir uns ausserdem eine Möglichkeit, uns einen Überblick über die Struktur der Inhalte zu verschaffen, ohne die zum Teil riesigen XML-Dateien öffnen zu müssen.

3.2.2 Verarbeitung der Dateien zwecks Datenanalyse

Es erwies sich in einer ersten oberflächlichen Analyse, dass die Struktur des Inhalts dieser Dateien mehrere Ebenen umfasst und in den verschiedenen Teilen unterschiedlich komplex ist. Eine tiefergehende Analyse war aufgrund der grossen Dateien ohne vorherige Verarbeitung derselben unmöglich. Aus diesem Grunde wurde ein Programm geschrieben, welches die Daten jeder Datei in eine (relationale) Datenbank schreibt, und zwar so, dass für jede sich wiederholende gleichbleibende Struktur eine Tabelle angelegt wurde, deren Name dem Elementnamen in der XML-Datei entspricht, und die zusammengehörenden Elemente für die Spaltenbenennung Verwendung fanden. Diese Umwandlung dauerte aufgrund der Datenmenge mehrere Stunden[8].

Aus dieser Umwandlung resultiert eine Datenbank, welche aus zahlreichen Tabellen besteht, die, durch zusätzliche Primär- und Sekundärschlüssel angereichert, die hierarchische Struktur der XML-Datei wiedergeben. Konkret weist die Datenbank folgende Eigenschaften auf:

[8] Der erste Versuch dauerte über eine Woche. Verschiedene Optimierungen am Code erlaubten eine Verkürzung auf ca. fünf Stunden. Die wichtigste Optimierung war die Aufteilung des Prozesses in zwei Läufe. Im ersten Lauf wurde der strukturelle Aufbau jeder XML-Datei ermittelt und darauf basierend das Datenbankmodell erstellt. Im zweiten Lauf wurden die zugehörigen Daten in CSV-Dateien eingelesen und diese am Schluss in die erstellte Datenbank importiert.

1. Elemente in der XML-Datei, welche Unterelemente enthalten, von denen mindestens eines keine weiteren Unterelemente hat, werden als Tabellen interpretiert.
2. Elemente in der XML-Datei, welche keine Unterelemente enthalten, werden als Tabellenspalten ihres Überelementes interpretiert.
3. Alle Tabellen, die aus 1 resultieren, erhalten einen zusätzlichen numerischen Primärschlüssel, der jeden Datensatz in der Tabelle eindeutig identifiziert.
4. Elemente in der XML-Datei, welche als Tabellen interpretiert werden und einem Überelement zugeordnet sind, welches ebenfalls als Tabelle interpretiert wird, erhalten einen zusätzlichen, numerischen Fremdschlüssel, der auf den Primärschlüssels des übergeordneten Datensatzes verweist. Dadurch kann die hierarchische Struktur der XML-Datei in der Datenbank abgebildet werden.

Um die Rückverfolgbarkeit der Datenbankdaten zu ihrer Herkunft gewährleisten zu können, wurde bei der Namensgebung der Tabellen- und Spaltennamen wie folgt vorgegangen:

1. Tabellennamen werden aus den Teilen <DATEINAME>_<NAMESPACE>_<ELEMENTNAME> zusammengesetzt. Ein Beispiel: Die Datei MIGADRES.XML enthält unter anderem sich wiederholende Elemente vom Typ <datamigzv-main:datamigzvheader>. Die Datenbanktabelle erhält dementsprechend den Namen MIGADRESS_DATAMGZV_MAIN_DATAMIGZVHEADER
2. Die Spaltennamen werden aus dem qualifizierten Elementnamen gebildet: <NAMESPACE>_<ELEMENTNAME>, z.B. wird <datamigZv-Main:housenumber> zu DATAMIGZV_MAIN_HOUSENUMBER.
3. In beiden Fällen werden Sonderzeichen als "_" umgesetzt, d.h. ein „:" oder ein „-„ wird zu einem „_".

Die zusätzlich eingefügten Spalten Primär- und Fremdschlüssel werden speziell gekennzeichnet:

1. Primärschlüsselspalten bestehen aus dem vereinfachten Tabellennamen mit dem Anhängsel _ID
2. Fremdschlüsselspalten bestehen aus dem referenzierten Tabellennamen mit dem Anhängsel „_FK".

Weil die XML-Dateistruktur aufgrund der Grösse der Dateien nicht immer einsehbar war, hat sich bei der Konvertierung der Daten eine Besonderheit ergeben: Wie sich nach der Konvertierung zeigte, gibt es Elemente gleichen Namens, die unterschiedlichen Überelementen angehören. Dies führte dazu, dass die entsprechenden Tabellen folgende Eigenschaften aufweisen:

1. Sie enthalten einen oder mehrere Fremdschlüssel (für jeden übergeordneten Elementnamen einen).
2. Jeder Datensatz in diesen Tabellen bezieht sich jeweils nur auf eines der vorkommenden Überelemente, das heisst nur jeweils einer von mehreren Fremdschlüsseln enthält einen definierten Wert, die anderen Fremdschlüssel haben einen NULL-Wert.

3. Es besteht die Möglichkeit, dass einzelne Spalten in diesen Tabellen nur für Datensätze, welche aus einem Element unter einem bestimmten Überelement extrahiert wurden, ausgefüllt sind.

3.2.3 Struktur der Daten

Erste Versuche zeigen, dass insgesamt 152 Tabellen angelegt werden. Einige Tabellen aus den verschiedenen Dateien weisen eine ähnliche oder gar identische Struktur auf. Gruppieren wir die Dateien vorerst nach ihrem Namen, so erhalten wir drei Gruppen:

Gruppe 1: Dateien, die mit MIG* beginnen. Es handelt sich um acht Dateien, deren Inhalt eher Stammdatencharakter, also keine grossen Änderungen über die Zeit erfahren hatte. Mit Ausnahme der Tabelle MIGADRES.XML, deren Inhalt die in der Stadt existierenden (und möglicherweise existiert habenden) Strassen und ihre Hausnummern zu beschreiben scheint (mit einem zeitlichen Gültigkeitsbereich), beinhalten die anderen sieben Dateien themenspezifische Stadtkreise mit allen diesen zugeordneten Strassen und Hausnummern. Die Themen umfassen: Reformierte Stadtkreise (MIGKREEK.XML), Katholische Stadtkreise (MIGKREKK.XML), Politische Stadtkreise (MIGKREPK.XML), Schulkreise (MIGKRESK.XML), Verkehrskreise (MIGKREVK.XML), Volkszählkreise (MIGKREVZ.XML) und Wohnkreise (MIGKREWV.XML). Eine Nachfrage bei Stephan Wenger, Leiter des Einwohneramtes, ergab, dass diese Kreise keine Relevanz im Tagesgeschäft des Einwohneramtes haben. Die Kreise sind heute im Stadtplan bei den administrativen Einteilungen abgebildet. Die politischen Kreise (MIGREPK.XML) gehen auf die Zeit zurück, als die Stadt in drei Stimm- und Wahlkreise mit selbständigen Stimmbüros unterteilt war. Die Verkehrskreise (MIGKREVK.XML) wurden früher für die Verkehrsplanung verwendet und haben heute keine Relevanz mehr. Die Volkszählkreise (MIGKREVZ.XML) waren im Zusammenhang mit der Volkszählung ein organisatorisches Hilfsmittel, das mit der registergeführten Volkszählung seit 2000 nicht mehr benötigt wird. Die Wohnkreise (MIGKREWV.XML) haben lediglich statistische Bedeutung, die Fragen wie „Wieviele Personen sind innerhalb der Stadt von wo nach wo umgezogen“ beantworten können.

Gruppe 2: Dateien, die mit PERS* beginnen. Diese Dateien beinhalten natürliche (PERSNAT.XML) und juristische Personen (PERSJUR.XML), welche in einer Beziehung zum Einwohneramt stehen. Laut Auskunft von Stephan Wenger vom Einwohneramt, sind diese Daten nicht von Interesse für statistische oder personenbezogene Auswertungen von Einwohnerdaten.

Gruppe 3: Diese Gruppe umfasst die beiden Dateien RELATION.XML und RESIDENT.XML. Bereits die Grösse der Dateien (RELATION.XML 178 MB, RESIDENT.XML 4 GB) lässt vermuten, dass es sich bei diesen um die relevanten beiden Dateien handelt. Wir werden uns daher schwerpunktmässig mit diesen Daten befassen.

3.2.4 Erstellen von XML Schema-Dateien als Unterstützung bei der Datenanalyse[9]

Einen Überblick über die innere Struktur der XML-Dateien kann eine XML Schema-Datei (*.xsd) geben. Da die XML-Dateien ohne weitergehende Informationen abgeliefert worden waren, schlossen wir diese Lücke, indem wir die Schema-Dateien erstellten.

[9] Vgl. Kurzprotokoll Seite 12, Kapitel 6.2 unter „Erstes Problem“, wo die Erstellung von Schema-Dateien gewünscht ist. Da der Hersteller der Anwendung VRSG EK laut eigenen Angaben nicht (mehr) über die Schema-Dateien zu den XML Dateien verfügt, wurde ein anderer Weg gewählt.

Dafür griffen wir auf ein Projekt zurück, dessen Ziel die Generierung von XML Schema-Dateien aus bestehenden XML-Dateien ist: Trang[10]. Wir erstellten damit die Schema-Dateien und verifizierten damit gleich, ob sie die XML-Dateien als gültig erkennen. Zwar sind auch Schema-Dateien nicht unbedingt leichte Kost für Ungeübte, sie verdichten jedoch die Struktur der XML-Datei so effizient, dass es bedeutend leichter ist, sich mit ihnen zu befassen als mit den XML-Dateien selber. Die Validierung war für alle XML-Dateien erfolgreich. Damit die Codes aus der Ursprungsdatenbank im Zusammenhang mit dem Projekt verfügbar sind[11], wurden sie zusätzlich von Hand als Kommentare in die Schema-Dateien eingearbeitet. Die Schema-Dateien befinden sich in der Projektstruktur im Ordner xsd.

3.2.5 Codes aus der Ursprungsdatenbank

Eine erste Datenanalyse an der Datenbank zeigte, dass es zahlreiche Spalten in Tabellen gibt, welche Codes enthalten, deren Bedeutung nicht unmittelbar herauszufinden war. Nachfragen bei der VRSG, dem Lieferanten der XML-Dateien, halfen[12], dass wenigstens ein wesentlicher Teil der Codes und deren Aufschlüsselung nachgeliefert wurde, allerdings nicht mehr kostenfrei. Die Codes sind in der sich im Projektverzeichnis doc befindlichen Datei Resident_GRDM_Mapping.pdf aufgelistet und wurden aus Zweckmässigkeitsgründen in die entsprechenden Schema-Dateien eingearbeitet. Nicht dokumentierte Codes, also Codes, welche zwar in der XML-Datei Verwendung finden, für die aber kein Eintrag in der von der VRSG gelieferten Codeliste vorhanden ist, wurden ebenfalls in die Schema-Dateien eingearbeitet, allerdings mit dem Vermerk „Nicht dokumentiert, da aus Vorgängersystem von VRSG EK übernommen“[13].

3.3 Archivierung

Die erstellte Datenbank (siehe Kapitel 3.2.2 Verarbeitung der Dateien zwecks Datenanalyse) wurde anschliessend mit dem Werkeug SIARD des Bundesarchivs in ein archivtaugliches Format extrahiert, welches zusammen mit dieser Dokumentation im Archiv abgelegt wurde. Zur Überprüfung wurde die SIARD Datei wieder in eine Datenbank (Zielsystem: MySQL) zurückgespielt, was ca. 120 Stunden dauerte.

4 Zusammenfassung

Die Ziele des KOST-Projektes EDat 14-025 konnten erreicht werden:

1. Eine umfassende Evaluation des Ist-Zustandes zeigt, dass in den Verwaltungen der beteiligten Stadtarchive verschiedene Systeme im Einsatz sind, deren Gemeinsamkeit der Mindestsatz an Daten ist, wie sie im

[10] Siehe Website http://www.thaiopensource.com/relaxng/trang.html, wo das Projekt beschrieben ist. Die Seite http://www.thaiopensource.com/relaxng/trang-manual.html beschreibt die Verwendung der Bibliothek, die von https://code.google.com/archive/p/jing-trang/downloads heruntergeladen werden kann (Stand: 10.04.2017).
[11] Siehe Kapitel 3.2.5 Codes aus der Ursprungsdatenbank
[12] Wie sich bei der weiteren Verarbeitung herausstellte, sind zahlreiche vorhandene Codes nicht mehr dokumentiert. Da sie ausserdem nur bedingt selbsterklärend sind, werden sie wohl kaum für Auswertungen verwendet werden können.
[13] Um was für ein System es sich gehandelt hat, ob es ein analoges (z.B. Kartei) oder digitales System war, ist nicht bekannt.

Registerharmonisierungsgesetz definiert sind und deren Historisierung mehr oder weniger vollständig ist.

2. Eine Ablieferungsvereinbarung für die digital vorliegenden Daten aus einem Einwohnerregister stellt sicher, dass eine Archivierung der Originaldaten aufgrund von definierten Ereignissen erfolgen kann.
3. Die im Rahmen der Phase 2 erfolgte Extraktion der Daten aus dem Quellsystem wurde durch den Hoster der Anwendung, die VRSG, vorgenommen. Die Daten geben den Stand vom 31. Oktober 2014 wieder, dem Stichdatum für die Übernahme der Daten in das neue System Loganto des Herstellers. Sie wurden dem Stadtarchiv auf Anfrage angeboten und im Zustand geliefert, wie sie für die Migration in das neue System erstellt worden waren. Es handelt sich dabei um 12 XML-Dateien, ohne weitere Dokumentation des Herstellers. Vom Einwohneramt geliefert wurden Screenshots der Bildschirmmasken, wie sie an einem Beispiel ausgesehen haben. Vom Hersteller nachgeliefert wurden auf Nachfrage hin im System verwendete Codes, wobei sich herausstellte, dass nur für einen Teil der Codes Erklärungen vorhanden sind.
4. Die Migration in ein archivtaugliches Format erfolgte über Umwege: Zuerst wurden die XML-Dokumente mit einem Java-Programm in eine relationale Datenbank eingelesen (vgl. dazu Kapitel 3.2.2, Seite 6). Aus den XML-Dateien wurden mit einem Fremdprogramm (vgl. Anmerkung 10, Seite 9) ausserdem XSD Schema-Dateien erstellt, in die die nachgelieferten Codes als Kommentare eingearbeitet wurden, damit sie zu einem späteren Zeitpunkt zu Rate gezogen werden können. Aus der Datenbank wurde anschliessend mit der SIARD Suite[14] eine SIARD-Datei erstellt. Bei dieser Gelegenheit wurde noch ein Bug in der Suite aufgespürt, und es wurden den Verantwortlichen Erweiterungsvorschläge gemacht.
5. Der Import in das digitale Archiv sowie ein Export aus demselben waren erfolgreich.
6. Nach der Archivierung in das SIARD-Format wurde aus der SIARD-Datei testweise wieder eine Datenbank erstellt. Die Erstellung der Datenbank dauerte ca. 120 Stunden.

5 Schlussfolgerungen

Wie die detaillierte Beschreibung des Archivierungsprozesses zeigt, war der Aufwand für die Archivierung extrem hoch. Wenn davon ausgegangen wird, dass für die Archivierung weiterer Fachanwendungen mit einem ähnlichen Aufwand zu rechnen ist, dann muss festgestellt werden, dass dieser Weg nicht beschritten werden sollte. Zu gross sind der betriebene Aufwand und die damit verbundenen Risiken eines Datenverlustes (durch Datenverfälschung oder –verlust). Stattdessen sollten die Energien darauf verwendet werden, die Hersteller solcher Systeme zu verpflichten, entsprechende Archivierungsschnittstellen zu implementieren, die die Anforderungen der Archive erfüllen.

[14] Die SIARD Suite wird vom Schweizerischen Bundesarchiv gepflegt, weiterentwickelt und kostenlos zur Verfügung gestellt. Sie ist nicht quelloffen. Siehe auch: https://www.bar.admin.ch/bar/de/home/archivierung/tools---hilfsmittel/siard-suite.html).

Während des Projekts wurden auch erste Abfragen durchgeführt, welche die Informationen wiedergeben, wie sie im Einwohneramt benötigt werden. Diese Abfragen werden im Anschluss an das Projekt noch vervollständigt. Sie zeigen aber auch die Grenzen solcher individuellen Lösungen auf, da deren Erstellung ebenfalls sehr zeitraubend ist. Eine durch den Hersteller eines Systems implementierte Archivierungsschnittstelle müsste auch diesen Aspekt berücksichtigen.

Grundsätzlich lässt sich sagen, dass zwar eine Archivierung einer Datenbank ohne Archivierungsschnittstelle möglich ist, der damit verbundene Aufwand (vor allem wenn noch berücksichtigt wird, dass diese Arbeit von jedem Archiv gemacht werden muss) aber unverantwortlich hoch ist.

Während das Einwohneramt uns mit Rat und Tat zur Seite stand, war die Zusammenarbeit mit dem Hersteller eher schwierig. Es kann nicht davon ausgegangen werden, dass Anwendungen, die in Verwaltungen benützt werden, gut dokumentiert sind. Wie sich im Projekt gezeigt hat, konnte oder wollte der Hersteller der Software technische Informationen nicht oder, wie bei den an sich unproblematischen Codes, nur nach hartnäckigem Nachfragen weitergeben. Es muss davon ausgegangen werden, dass auch bei anderen Fach- oder Datenbankanwendungen kein bis wenig Interesse des Herstellers besteht, Informationen zur Software herauszugeben.

6 Anhang

6.1 1. Entwurf einer Vereinbarung zwischen Einwohneramt und Archiv

Vereinbarung zwischen dem Einwohneramt und dem Stadtarchiv bezüglich Datenhaltung der Einwohnerdaten

1. Zweck dieser Vereinbarung ist die langfristige Sicherstellung der Datenbestände des Einwohneramtsregisters.

2. Diese Vereinbarung präzisiert die in der Fristenliste des Staatsarchivs St.Gallen auf Seite 7 aufgelisteten Daten mit ihren Aufbewahrungsfristen:

Einwohnerwesen	Aufbewahrungsfrist
Aufenthaltsregister	dauernd
Karteien (usw.)	dauernd
Niederlassungsregister	dauernd
Einwohnerstatistik, laufende Fortschreibungen	dauernd
Gesuche von Ausländern um Niederlassung oder Aufenthalt	10 Jahre

Ziel ist die Klärung der Verantwortlichkeiten zwischen Einwohneramt und Stadtarchiv.

3. Das Einwohneramt bietet dem Stadtarchiv bei folgenden Ereignissen den Altbestand an:

   a. Bei Änderungen von Gesetzen, Verordnungen, Reglementen etc., welche zu einer Änderung in der Datenhaltung und –pflege führen, so dass mit Verlusten von bereits bestehenden Daten zu rechnen ist.

   b. Bei einer Migration des Einwohneramtsregisters und den damit verbundenen Daten.

   c. Bei strukturellen Änderungen am System, durch die vorhandene Daten verändert werden oder verloren gehen können.

4. Die regelmässige Ablieferung von aktuellen Daten ist nicht vorgesehen, da diese vollständig und historisiert im produktiven System verfügbar sind.

6.2 Kurzprotokoll KOST Projekt 14-025 Edat vom 2. Februar 2017

(von Martin Kaiser)

Treffen mit C. Eugster zum Thema „in welcher Form sollen die Daten aus dem Einwohnerregistersystem archiviert werden?“

Ausgangslage:

Die Daten des Einwohnerregisters St. Gallen wurden 2014 aus VRSG-EK nach Loganto migriert (beides von VRSG)

Seit 2014 werden VRSG-EK und Loganto parallel betrieben, Loganto als leading System.

VRSG-EK wird für die Datenlieferungen an diverse andere Fachapplikationen vorläufig weiter benötigt.

VRSG-EK basiert auf ADABAS von Software AG, eine nicht relationale Datenbank, eine Archivierung in SIARD ist also nicht möglich.

Für diese Migration wurden die Daten aus ADABAS in zwölf XML-Dateien exportiert.

Jede XML Datei enthält in einer hierarchischen Struktur den Inhalt vieler Tabellen und Records.

Die XML-Dateien sind well-formed aber ohne Schema, sie wurden offenbar zur Datenmigration VRSG-EK Loganto verwendet und stellen den Datenbestand 31. Oktober 2014 dar. Die grösste Datei ist 3.8 GB und enthält die eigentlichen Personendaten.

Wie im Projekt 14-025 EDat definiert, wurden die Daten dem Stadtarchiv angeboten und übernommen.

Mitgeliefert wurde ein Set von Screenshots der Applikation, mit einem best., immer gleichen Einwohner.

Erstes Problem

Es gibt keine weitere Strukturbeschreibung zu den XML-Dateien und der ADABAS Datenbank. Die Elemente in der XML Datei sind wohl die Feldnamen im Quellsystem und recht selbsterklärend, es ist aber mit gängigen Tools nicht möglich, ein Schema zu generieren (wegen der Dateigrösse).

➔ Christian recherchiert, ob es nicht doch ein Tool dafür gibt, vielleicht von Oracle, weil der Oracle XML-Import möglicherweise für die Migration nach Loganto Verwendung gefunden hat.

Zweites Problem

Viele Elemente enthalten nur Codewerte, z.B. Geschlecht, die Aufschlüsselung in sprechende Texte erfolgte offenbar direkt in den Masken und Reportabfragen.

➔ Die VRSG wird dazu gebeten eine Codewert Aufstellung/Tabelle nachzuliefern

Drittes Problem

Zugang und Erschliessung der Daten in XML ist praktisch nicht möglich. Christian hat dazu die XML-Dateien in einem Javaprogramm sequentiell gelesen und in eine (Apache) DERBY Datenbank geschrieben. Die hierarchische Struktur in der XML Datei hat er dabei als Relationen zwischen den Tabellen abgebildet (durch Einfügen eines künstlichen Schlüssels). In dieser Datenbank sind nun einfache SQL Abfragen möglich.

➔ Es fragt sich, ob diese Zugriffsdatenbank allenfalls durch Löschen überflüssiger Daten vereinfacht werden könnte

➔ Für eine Archivierung müsste die DERBY Datenbank z.B. in mySQL konvertiert werden (via SQL Dump/Loader File oder via JDBC Schnittstelle?)

➔ Die möglichen und sinnvollen Abfragen müssen mit archiviert werden.

Universelle Frage

Die Möglichkeit, aus einem beliebig grossen XML-File ein Schema für die Dokumentation zu erzeugen, ist ein generelles Problem für die digitale Archivierung. Christian überlegt sich, ob seine Lösung nicht (statt die Daten zu transferieren) ein XSD Schema generieren könnte.

6.3 Verarbeitungsprogramm ch.eugster.ea.Converter

Das Verarbeitungsprogramm ch.eugster.ea.Converter wurde primär dafür geschaffen, die Daten aus den XML-Dateien in eine Datenbank zu konvertieren. Bei den XML-Dateien kann es sich um beliebige valide XML-Dateien beliebiger Grösse handeln. Während des EDat Projektes wurden laufend Ergänzungen implementiert, die die Arbeit unterstützen

sollten. Da das Programm für den Eigengebrauch erstellt wurde, müsste es umfassend überarbeitet werden, wenn es der Öffentlichkeit verfügbar gemacht werden sollte. Das Programm umfasst folgende Funktionalitäten:

1. Erstellen der Datenstruktur in einer bestehenden Datenbank[15] und Füllen der Datenbank mit den Daten aus dem XML-Dateien;
2. Auflisten der erstellten Tabellen der Datenbank;
3. Anzeigen der Datenstruktur einzelner Tabellen;
4. Leeren der Datenbank (d.h. Löschen aller Tabellen und Beziehungen);
5. Exportieren der Datenbank in CSV Dateien;
6. Erstellen von XSD Schema-Dateien aus den XML-Dateien;
7. Validieren der XML-Dateien mittels der erstellten XSD Schema-Dateien;
8. Validieren der XML-Dateien gegen die vorhandenen Checksummen;
9. Konvertieren der bestehenden Datenbank in Microsoft Access Dateien (diese erlauben mit ihrem GUI leichter komplexere Abfragen zu erstellen);
10. Angefangen, aber noch nicht fertig: Vorgefertigte Abfragen, um Informationen zu einer bestimmten Person anzuzeigen.

[15] Zur Zeit sind folgende Datenbanksysteme unterstützt: Apache Derby und Microsoft SQL Server. Weitere Systeme können relativ leicht ebenfalls implementiert werden.