Überlebensstrategien für Bits and Bytes
12 Jahre elektronische Archivierung im Siemens-Archiv
Ein Erfahrungsbericht

Ute Schiedermeier

Anfänge der Digitalisierung

Seit Anfang der 1990er Jahre nahm im Siemens-Archiv die Zahl von Anfragen erheblich zu. Was war zu tun?

Die technischen Vorraussetzungen der Informationsverarbeitung waren so weit gediehen, dass nützliche Konkordanzen komfortabel erstellt werden konnten. Allerdings fehlte die nötige Erschließungstiefe. Ein schneller Zugang zu Nachschlagewerken war also nicht die Lösung.

Erst die Retrodigitalisierung (scannen und Volltexterschließung von Papierdokumenten) - ermöglicht durch den Einsatz eines neuen Archiv-Informationssystems – half bei der Lösung der anstehenden Probleme. Das System sollte dem vom Hause gewohnten hohen IT-Standard entsprechen. Die Erwartung der externen Nutzer an die Qualität der Informationsbereitstellung war hoch.

1993 wurden 150.000 Seiten Rundschreiben und Organisationsunterlagen sowie 24.000 Seiten Dokumente der Zentrale gescannt. Diese Seiten durchliefen eine OCR-Erkennung zur Volltextsuche und wurden zusätzlich zum Volltext im TIFF-Format abgelegt. Alle Dokumente wurden zusätzlich mit Datenbankinformationen – den Indexwerten – versehen. Parallel wurden 46.000 Fotos des so genannten Handarchivs gescannt und grob erschlossen.

Das Archiv sollte in der Lage sein, bei aktueller Fragestellung die jeweils relevanten Dokumente in kürzester Zeit zu liefern. Die Praxis hat gezeigt, dass wir diesen Anspruch erfüllen konnten. Jenen positiven Erfahrungen folgten weitere Retrodigitalisierungsprojekte.

Erste Datenübernahmen

Schon 1993 wurde im Siemens-Archiv beschlossen, als Langzeit-Speicherformat nur TIFF-, JPEG- und das ASCII-Format zuzulassen. 1995 wurden erstmals elektronische Daten auf Datenträgern an uns geliefert. Diese TIFF-, JPEG- oder ASCII-Dateien konnten direkt über eine Systemschnittstelle übernommen und mit entsprechenden Indexwerten verknüpft werden.

Komfortable Datenkonvertierung 1997

Die Artenvielfalt nahm zu. Immer schneller behaupten sich neue Formate auf dem Markt. Vereinzelnd war schon zu vernehmen: „Erscheint zukünftig nur noch digital“! Die Verweilzeiten auf einer Inter-/Intranetseite wurden immer kürzer. Jetzt hieß es schnell handeln, um einem drohenden Datenverlust entgegenzuwirken. Die Lösung bestand in einer Importschnittstelle, die in der Lage war, alle gängigen Formate in unsere Standardformate zu konvertieren. Parallel war eine komfortable Vergabe von Attributen möglich.

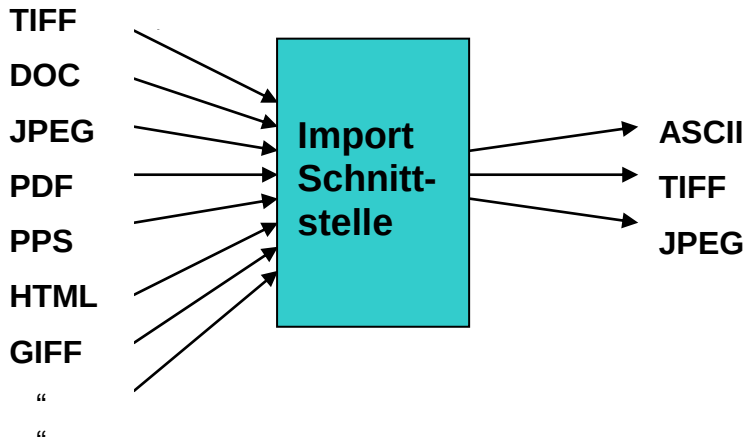

Abbildung 1: Importschnittstelle zur Datenkonvertierung

Damals wie heute werden folgende Dokumentarten im Siemens-Archiv elektronisch erfasst. Das heißt, dass entweder der Inhalt digitalisiert wird oder die Daten schon digital zu uns gelangen

Quelle	Dokument	Format	Volltext	Image
Papierform, E-Mail	**Text-Dokumente, Drucksachen**	**TIFF**	**X**	**X**
Papierform, Intranet, E-Mail	**Rundschreiben**	**TIFF**	**X**	**X**
Papierform, E-Mail	**Chroniken**	**TIFF**	**X**	**X**
Papierform, Intranet, E-Mail	**Zeitschriften**	**TIFF**	**X**	**X**
Papierform, Datei	**Presseinformationen**	**TIFF**	**X**	**X**
Papierform, Datei	**Pressebilder und Text**	**JPEG 72 dpi JPEG 300 dpi TIFF 300 dpi**	**X**	**X**
Papierform, Datei, CD, DVD	**Bilder**	**JPEG 72 dpi JPEG 300 dpi TIFF 300 dpi**		**X**
Digitalkamera, Filmkamera	**Fotos von Exponaten**	**JPEG 72 dpi JPEG 300 dpi**		**X**
Papierform, Word-Texte	**Historische Briefe und Transkriptionen**	**TIFF**	**X**	**X**
Papierform	**Vom Zerfall bedrohte Unterlagen**	**TIFF**	**X**	**X**

Tabelle 1: Übersicht über genutzte Archivierungsformate

Warum wird nicht das ganze Archiv elektronisch erschlossen?

Unsere Bestände umfassen rund 4 Regalkilometer Akten, 400.000 Fotos, 7.000 Bilder, 15.000 Exponate und 2.500 Filme. Auf der einen Seite ist eine Digitalisierung aller Bestände technisch bei uns nicht möglich und auf der anderen Seite stehen die Kosten in keinem Verhältnis zur Nutzung.

Folgende Auswahlkriterien gelten für künftige digitale Bestände:

1) Inhaltliche Bedeutung (z.B. Unterlagen der Zentrale)
2) Nachfragehäufigkeit (wie Chroniken, Zusammenfassungen)

3) Notwendigkeit – Massendaten, die bei uns aus personellen Gründen nicht einzeln erfasst werden können und ohne OCR nicht erschließbar sind (z.B. Rundschreiben oder Zeitschriften)
4) Konservatorische Gründe (Unterlagen aus dem 19. Jahrhundert oder vom Zerfall bedrohte Dokumente)
5) Arbeitsersparnis – Via Intranet bzw. firmeneigenem Datennetz können andere Archive, Agenturen, Beteiligungen bzw. weitere Mitarbeiter selbst auf freigegebene Daten zugreifen oder Daten downloaden
6) Standortunabhängigkeit – Direkter Zugriff auf Bestände in Archiven siemensweit; beispielsweise befinden sich Exponate im Außenlager.

Migration und Strukturoptimierung

Im letzten Jahr haben wir ein neues Archiv-Informationssystem in Betrieb genommen, das die wachsenden technischen und organisatorischen Anforderungen der nächsten Jahre erfüllen soll. Wegen eines damit verbundenen Datenbankwechsels musste der gesamte Datenbestand migriert werden. Dieser umfasste 240.000 Datensätzen mit rund einer Million Seiten.

Wir sahen der Migration mit gemischten Gefühlen entgegen, war dies doch die Bewährungsprobe für unsere Format- und Verzeichnungsstrategie. Wo lagen wir richtig, wo falsch, geht etwas unwiederbringlich verloren und wie stringent haben wir gearbeitet?

Migrationsablauf

Die Altdaten mussten aus einer Fulchrom-Datenbank in eine Microsoft SQL-Datenbank überführt werden. Das neue System bietet eine Import-Schnittstelle für Daten, die im csf-Format vorliegen. (csf = comma separated file)

Die Indexbegriffe werden durch Komma (oder ein anderes eindeutiges Zeichen) getrennt dargestellt. Am Schluss werden die Formatdateien – in unserem Fall TIFF und JPEG - angehängt. Das gemeinsame Element ist der eindeutige Schlüssel des Datensatzes, der den Dateinamen der Bilddateien bildet. Dieses Zurückgehen auf einen einfachen Nenner ist leicht nachvollziehbar. Der Migrationsprozess läuft aber vergleichsweise langsam ab.

Wir beauftragten unseren Systembetreuer mit der Erstellung der 240.000 csf-Zeilen. Dazu benötigte dieser etwa 40 Nächte. Tagsüber wollten wir das Altsystem für Recherchearbeiten weiternutzen. Die Migrationsfortschritte wurden ständig überwacht: Anzahl der Dokumente auf WORM, Quersummenchecks, Plausibilitätskontrollen und Fehler wurden parallel protokolliert.

Die so entstandenen csf-Dateien haben wir nicht gleich in das neue System eingelesen, sondern haben vorab eine umfangreiche Strukturoptimierung und Datenpflege durchgeführt: Unsere Datensätze gehen bis auf das Jahr 1969 zurück. Mehrere Generationen von Mitarbeitern und Systemen waren an der Bestandserschließung beteiligt. Der Zustand der Indexwerte war suboptimal; die Schreibweise zum Teil sehr uneinheitlich. In monatelanger Kleinarbeit haben hier Mitarbeiter und Werkstudenten Abhilfe geschaffen. Excel-Listen mit hunderttausenden von Indexbegriffen wurden korrigiert und anschließend mit den csf-Dateien abgeglichen. Das Abgleichen erfolgte automatisiert und dauerte pro Datenkollektion nur wenige Minuten.

Außerdem wurden neue Indexfelder hinzugefügt und bisher bei uns nicht vorhandene Nachschlagetabellen erzeugt. Die Inhalte nicht mehr benötigter Felder wurden aufgeteilt und die Felder anschließend nicht ins neue System übernommen. Außerdem wurde der Thesaurus angepasst. Über Protokolle konnten wir diese Arbeitsschritte überwachen.

Die jetzt modifizierten csf-Dateien wurden an das Importprogramm des Neusystems übergeben. Bei dieser Übernahme fanden standardmäßig umfangreiche Plausibilitätskontrollen statt, wie sie in unserem Altsystem nicht möglich waren. Die Übernahme der csf-Daten erfolgte archivweise. Am folgenden Tag konnten die übernommenen Daten getestet werden.

Wie haben unsere Daten die Migration überstanden?

- Nach unserer Kenntnis gingen keine Datensätze verloren
- Dateifehler konnten bereits während der jeweiligen Migrationsschritten beseitigt werden. Das betraf folgende Vorgänge:
    - Zehn Seiten der Retrodigitalisierung mussten nachgescannt werden
    - Doppelte Einträge wurden beseitigt (ca. 30)
    - Eine Bilderkollektion im JPEG-Format entsprach nicht mehr den heutigen Formatrichtlinien und musste konvertiert werden.
- Die Kosten blieben innerhalb des eng bemessenen Budgets
- Einen kompletten Rechnerausfall gab es nie; im Altsystem konnte bis zum Schluss recherchiert werden; im Neusystem konnte die Verzeichnungsarbeit nach der jeweiligen Archivübernahme sofort wieder aufgenommen werden.
- Die Übernahme der Formatdateien lief reibungslos

Fazit: Unsere Befürchtungen in Bezug auf einen möglichen Datenverlust durch die Migration waren unbegründet.

Der Datenbestand war noch nie so wertvoll, sauber und übersichtlich wie heute!

Übernahme elektronischer Daten heute

Das neue Archiv-Informationssystem bietet eine DMS-orientierte Arbeitsweise. Die elektronischen Daten werden entweder außerhalb des Systems mit Standardprodukten in unsere Formate konvertiert und über den DMS-Eingangskorb in das System geladen oder direkt in das System gescannt. Dort werden die Daten strukturiert, Archiven zugeordnet und verzeichnet.

Größere Datenmengen werden mittels des vorher genannten Importtreibers als csf-Dateien übernommen. Viele Standardprogramme bieten heute die Umwandlung in csf-Dateien oder das Ausgeben in Excel-Listen an, die wiederum in csf-Dateien gewandelt werden können. Diese Listen müssen von unseren Systembetreuern vorher genau analysiert und den gültigen Datenbankfeldern zugeordnet werden. Eine einfache Standardlösung ist bei den vielen unterschiedlichen Übernahme-Formaten zurzeit noch nicht in Sicht.

Für die Übernahme von komplexen Datenbanken, kompletten Internetauftritten oder mit Autorensystemen erzeugten Medien gibt es bei uns noch keine Lösungen.

Kundenauftrag versus Standardformat

Bisher wurden nur die Formate TIFF und JPEG gespeichert: TIFF als verlustfreies Standardformat und JPEG als Transportformat (z.B. für E-Mails).

Durch neue Software-Funktionen können Office-Formate mit abgespeichert werden, z.B. doc, ppt, xls. Wir speichern neuerdings diese Formate zusätzlich zu unseren Standardformaten ab, wohl wissend, dass diese Formate in einigen Jahren nicht mehr lesbar sein werden. Diese zusätzlichen Formate werden nur an interne Kunden weitergegeben, mit dem Ziel, diese Dokumente weiter zu verarbeiten – eine Dienstleistung, die in letzter Zeit immer wieder von uns gefordert wurde. Wir weichen damit von unseren grundsätzlichen Überlegungen ab – zumindest intern – nur unveränderbare Daten weiterzugeben.

Andererseits hoffen wir, dass uns diese parallele Speicherung von verschiedenen Formaten dabei hilft, auf zukünftige Standardarchivformate wie z.B. XML mit einem unverfälschten Ausgangsformat zu reagieren.