Addendum zum Standard eCH-0165: SIARD-Formatspezifikation

Dokument

Titel	Addendum zum Standard SIARD-Formatspezifikation
eCH-Nummer	eCH-0165
Addendum zu	Standard Version V1.0
Reifegrad	Implementiert
Sprachen	Deutsch (Original)

Status

Dokument	Genehmigt; Abgelöst; Aufgehoben
Ausgabedatum	2015-03-03

Autor

Fachgruppe	Digitale Archivierung
Kontaktperson	
Name Vorname	Martin Kaiser
Organisation	KOST Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen
E-Mail	martin.kaiser@kost.admin.ch
Telefon	+41 79 464 08 60
Herausgeber	Verein eCH, Mainaustrasse 30, Postfach, 8034 Zürich T 044 388 74 64, F 044 388 71 80 www.ech.ch / info@ech.ch

Zusammenfassung

In der Erstfassung der SIARD-Spezifikation wurde die ZIP-Komprimierung ausdrücklich untersagt. Die entsprechenden Algorithmen waren damals noch patentrechtlich geschützt. Das hat sich inzwischen geändert, und es besteht kein Grund mehr, auf die Vorteile der Datenkomprimierung, gerade beim Archivieren grosser Datenbanken, zu verzichten.

Inhaltsverzeichnis

1 Einleitung

Beim Definieren von Standards kann es sein, dass gewisse Fragen und Probleme erst bei der Implementierung und Anwendung auftreten oder entdeckt werden. Das Addendum hält Ergänzungen und Präzisierungen fest, die sich aus diesem Umstand ergeben und sonst erst mit der nächsten Version des Standards veröffentlicht werden könnten.

Einträge im Addendum fliessen in die nächste Version des betroffenen Standards ein.

1.1 Grundlegendes

Bei der ersten Spezifikation des SIARD-Formates war der beim Erstellen eines ZIP-Archives gängige Komprimierungsalgorithmus *Deflate* noch mit einem Patent belastet. Weil andere *Public-Domain*-Komprimierungsalgorithmen damals noch keine grosse Verbreitung hatten und in der Archivwelt generell ein Vorbehalt gegenüber Komprimierungsalgorithmen bestand (wegen patentrechtlicher Bedenken und aus Angst vor zu grosser Komplexität), wurde auf die Komprimierung generell verzichtet (siehe A_4.1-1).

Inzwischen ist aber diese patentrechtliche Einschränkung schon seit einiger Zeit weggefallen.

1.2 Grosse Datenmengen als Herausforderung

Die Anwendung von SIARD zur Archivierung von grossen Datenbanken und CSV-Sammlungen hat gezeigt, dass hier durch Datenkomprimierung für die Archive ein enormes Sparpotential brach liegt. SIARD-Dateien komprimieren wie alle XML-basierten Formate unglaublich gut, das heisst in der Regel um den Faktor 9 bis 10.[1]

Zudem wird das Handling entsprechender SIARD-Dateien im Archivsystem wesentlich vereinfacht, wenn die zu archivierenden Dateien um den Faktor 10 kleiner sind: Es werden dann in der Regel keine Dateigrössen-Einschränkungen von Seiten Betriebs- oder Dateisystem verletzt.

1.3 Deflate Algorithmus

Deflate ist ein Algorithmus zur verlustlosen Datenkompression, der von Phil Katz für das ZIP-Archivformat entwickelt wurde. Bei *Deflate* handelt es sich um eine Kombination des Lempel-Ziv-Storer-Szymanski-Algorithmus und der Huffman-Kodierung. LZSS basiert auf dem Lempel-Ziv-Welch Algorithmus. Das den LZW-Algorithmus betreffende US-Patent 4.558.302 ist am 20. Juni 2003 nach 20 Jahren ausgelaufen. Die entsprechenden europäischen, kanadischen und japanischen Patente folgten im Juni 2004.

[1] Gerade Gedächtnisinstitutionen mit grösseren digitalen Sammlungen haben inzwischen bemerkt, dass die Kosten für die Speicherung schnell alle andern Kostenpositionen übertreffen können.

2 Lösung

2.1 *Deflate* Komprimierung zulassen

Es wird darum in Paragraph A_4.1-1 *Deflate*-Komprimierung erlaubt.

Im Augenblick scheint es nicht sinnvoll, andere Komprimierungsverfahren als die seit Anbeginn für ZIP definierte *Deflate*-Komprimierung zuzulassen[2]. *Deflate* ist mit Abstand am weitesten verbreitet und wird heute auch bei den gängigen XML basierten Formaten (z.B. Microsoft DOCX oder Open-Document ODF) verwendet.

2.2 Bestehende Programme

Wichtige bestehenden Programme und Tools für den Umgang mit SIARD-Format (SIARD-Suite des Schweizerischen Bundesarchives und KOST-Val) benutzen beide die Open-Source ZIP64 Library der Enter AG, welche bereits *Deflate* der Untergruppen Normal *(-en compression)* und Maximum *(-exx/-ex compression)* uneingeschränkt unterstützt.

Eine komprimierte SIARD Datei kann aber auch mit jedem gängigen ZIP Programm und der Komprimierung *Deflate* (Normal oder Maximum) aus einer unkomprimierten SIARD Datei erzeugt werden. Wenn das ZIP-Programm keine Auswahl des Komprimierungsalgorithmus zur Verfügung stellt, ist der *Default* immer *Deflate*.

2.3 Lösung im folgenden Release

Diese Anpassung wird in den folgenden Release übernommen. A_4.1-1 lautet dann:

ID	Beschreibung Anforderung	M/K
A_4.1-1	Die SIARD-Datei wird als ein einziges unkomprimiertes oder deflate komprimiertes ZIP-Archiv gemäss der von der Firma PkWare publizierten Spezifikation, Version 6.3.2 gespeichert.	M

[2] BZIP2, das in den 90er Jahren als patentfreie Alternative zu *Deflate* entwickelt wurde, hat nicht dieselbe Verbreitung erlangt.

3 Haftungsausschluss/Hinweise auf Rechte Dritter

eCH-Standards, welche der Verein eCH dem Benutzer zur unentgeltlichen Nutzung zur Verfügung stellt, oder welche eCH referenziert, haben nur den Status von Empfehlungen. Der Verein eCH haftet in keinem Fall für Entscheidungen oder Massnahmen, welche der Benutzer auf Grund dieser Dokumente trifft und / oder ergreift. Der Benutzer ist verpflichtet, die Dokumente vor deren Nutzung selbst zu überprüfen und sich gegebenenfalls beraten zu lassen. eCH-Standards können und sollen die technische, organisatorische oder juristische Beratung im konkreten Einzelfall nicht ersetzen.

In eCH-Standards referenzierte Dokumente, Verfahren, Methoden, Produkte und Standards sind unter Umständen markenrechtlich, urheberrechtlich oder patentrechtlich geschützt. Es liegt in der ausschliesslichen Verantwortlichkeit des Benutzers, sich die allenfalls erforderlichen Rechte bei den jeweils berechtigten Personen und/oder Organisationen zu beschaffen.

Obwohl der Verein eCH all seine Sorgfalt darauf verwendet, die eCH-Standards sorgfältig auszuarbeiten, kann keine Zusicherung oder Garantie auf Aktualität, Vollständigkeit, Richtigkeit bzw. Fehlerfreiheit der zur Verfügung gestellten Informationen und Dokumente gegeben werden. Der Inhalt von eCH-Standards kann jederzeit und ohne Ankündigung geändert werden.

Jede Haftung für Schäden, welche dem Benutzer aus dem Gebrauch der eCH-Standards entstehen ist, soweit gesetzlich zulässig, wegbedungen.

4 Urheberrechte

Wer eCH-Standards erarbeitet, behält das geistige Eigentum an diesen. Allerdings verpflichtet sich der Erarbeitende sein betreffendes geistiges Eigentum oder seine Rechte an geistigem Eigentum anderer, sofern möglich, den jeweiligen Fachgruppen und dem Verein eCH kostenlos zur uneingeschränkten Nutzung und Weiterentwicklung im Rahmen des Vereinszweckes zur Verfügung zu stellen.

Die von den Fachgruppen erarbeiteten Standards können unter Nennung der jeweiligen Urheber von eCH unentgeltlich und uneingeschränkt genutzt, weiterverbreitet und weiterentwickelt werden.

eCH-Standards sind vollständig dokumentiert und frei von lizenz- und/oder patentrechtlichen Einschränkungen. Die dazugehörige Dokumentation kann unentgeltlich bezogen werden.

Diese Bestimmungen gelten ausschliesslich für die von eCH erarbeiteten Standards, nicht jedoch für Standards oder Produkte Dritter, auf welche in den eCH-Standards Bezug genommen wird. Die Standards enthalten die entsprechenden Hinweise auf die Rechte Dritter.