# Compression des données et archivage : le binôme du futur = Datenkompression und Archivierung : das Binom der Zukunft

Autor(en): **Vandergheynst, Pierre / Gillioz, Stéphane**

Objekttyp: **Article**

Zeitschrift: **Arbido**

Band (Jahr): **- (2008)**

Heft 4: **Informationswissenschaft: die Instrumente der Zukunft = Information documentaire: les outils du futur = Scienze della informazione: gli strumenti di domani**

PDF erstellt am: **01.09.2024**

Persistenter Link: https://doi.org/10.5169/seals-769801

# I. Tendances générales, recherches et projets
## *Allgemeine Tendenzen, Forschungen und Projekte*

# Compression des données et archivage: le binôme du futur

Pierre Vandergheynst, professeur à l'EPFL, en collaboration avec Stéphane Gillioz, rédaction *arbido*

**Les ondes radio, les circuits téléphoniques et les câbles d'ordinateurs véhiculent quotidiennement des quantités astronomiques d'informations numériques. Or, comment les référencer si les professionnels de l'information documentaire, entre autres, veulent pouvoir les archiver et les usagers les utiliser? Un double défi attend les chercheurs: la compression et l'indexation.**

### Les faits

Lorsqu'on parle de «quantités astronomiques» d'informations numériques véhiculées chaque jour par les différents modes de transmission (TV, téléphones, internet, caméras de surveillance, etc.), ce n'est pas une figure de style, loin s'en faut! Et le futur – proche! – va nous permettre de le vérifier à la puissance n.

Le livre blanc édité par l'IDC en mars 2007 (1) constate que la quantité d'informations numériques créée, saisie et transmise en 2006 était de $1{,}288 \times 10$ puissance 18 bytes. Ce qui correspond à 161 exabytes ou 161 billions de gigabytes; autrement dit environ 3 millions de fois l'information contenue dans tous les livres écrits depuis le début des temps. Mais le meilleur est encore à venir, puisque, toujours selon le rapport de l'IDC, le volume d'informations sera multiplié par 6 d'ici à 2010...

Se pose dès lors la question suivante: lorsque l'on sait que 95% de ces données ne sont pas structurées, comment les référencer? Or, la réponse à cette question est de toute première importance pour les professionnels de l'information documentaire qui seront appelés à utiliser les nouveaux outils que les scientifiques sont en train de mettre au point dans ce domaine.

### Compresser, puis indexer

La solution comporte deux étapes: 1) il faut d'abord compresser, puis 2) indexer. La difficulté est de taille, puisqu'il s'agit de comprimer les données tout en les structurant «sémantiquement». On connaît déjà des formats de compression comme MPEG, ZIP, JPEG et, plus récent, JPEG2000 (voir encadré), mais ils ne sont encore que des embryons de solutions face au défi que représentent les volumes de données à valoriser.

Prenons par exemple les archives du Festival de Montreux, donc pour l'essentiel des données son et image. L'EPFL se charge actuellement de la numérisation de l'archivage de ce fonds. Mais comment accéder à l'information voulue dans des délais raisonnables? La réponse est sur toutes les lèvres: par recherche «sémantique».

### Le défi de la recherche «sémantique»

Certes, mais ici aussi le défi est de taille. Les contenus sont de toute première importance dans ce contexte. Or, l'on sait que ces contenus comprennent du son, du texte, de l'image et de la vidéo. Il faut donc rechercher sur différents types de données. La recherche que l'on propose actuellement est indépendante d'un type de données à un autre. La solution réside donc dans l'intégration de ces données, afin qu'une recherche ciblée soit possible.

Autre exemple: les meetings virtuels, qui sont de plus en plus fréquents et qui seront certainement appelés à se multiplier à l'avenir si l'on considère l'explosion des coûts de déplacement due à la pénurie croissante des énergies non renouvelables. L'archivage de ces meetings (politiques, scientifiques, associatifs, sportifs, culturels) sera donc indispensable et nécessitera des solutions au niveau de la compression des données et de leur stockage qui n'existent pas encore. Le fameux «binôme du futur» sur lequel des milliers de chercheurs se penchent actuellement de par le monde...

### Conclusion

La tâche est donc titanesque pour les chercheurs et il faudra encore du temps avant que les professionnels de l'information documentaire puissent disposer d'outils leur permettant de fournir à leurs clients des prestations dignes de ce nom en matière de fonds audio visuels.

*La norme JPEG2000*

**JPEG2000 est un nouveau système de codage d'image utilisant l'état de l'art des techniques de compression et basé sur la transformée en ondelettes. Son architecture devrait être appropriée à un grand nombre d'applications depuis les appareils photos numériques jusqu'à l'imagerie médicale et d'autres secteurs clé. Le codage comporte des informations sur le contenu ainsi qu'une indexation primaire.**

### Références:

(1) *The Expanding Digital Universe. A Forecast of Worldwide Information Growth Through 2010*, sous la direction de John F. Gantz, mars 2007

**Contact: pierre.vandergheynst@epfl.ch**

# Datenkompression und Archivierung: das Binom der Zukunft

Pierre Vandergheynst, Professor an der EPFL, in Zusammenarbeit mit Stéphane Gillioz, Redaktion *arbido*

**Radiowellen, Fernsprechleitungen und Computerkabel transportieren jeden Tag astronomische Mengen von digitalen Daten. Wie soll man diese Daten referenzieren, wenn u.a. Profis der Informationsdokumentation diese Daten archivieren und «gewöhnliche» Nutzer sie nutzen wollen? Die Forschungsgemeinde erwartet eine doppelte Herausforderung: einerseits die Kompression, andererseits die Indexierung.**

## Fakten

Wenn die Rede ist von «astronomischen Datenmengen», die tagtäglich mit verschiedenen Geräten und Medien (TV, Telefon, Internet, Überwachungskameras etc.) transportiert werden», so ist das keine Übertreibung – die (nahe!) Zukunft wird uns zeigen, dass «astronomisch» noch um den Faktor n zunehmen wird. Das von der IDC im März 2007 herausgegebene Weissbuch (1) hält fest, dass die Gesamtsumme von digitalen Daten, die 2006 produziert wurde, $1{,}288 \times 10$ hoch 18 Bytes beträgt, das sind 161 Exabytes oder 161 Billionen Gigabytes; oder mit anderen Worten drei Millionen Mal die Information, die in sämtlichen je geschriebenen Büchern enthalten ist. Das Beste kommt aber noch: Gemäss dem Bericht der IDC wird diese Informationsmasse bis zum Jahr 2010 noch um den Faktor 6 anwachsen.

Damit drängt sich folgende Frage auf: Man weiss, dass 95% der Daten nicht strukturiert sind – wie soll man sie also referenzieren? Die Antwort auf diese Frage ist für die Berufsleute aus dem Bereich Informationsdokumentation von entscheidender Wichtigkeit: Sie werden mit unter den Ersten sein, welche die von den Forschern gegenwärtig zu diesem Zweck entwickelten neuen Instrumente anwenden werden.

## Komprimieren, dann indexieren

Die Lösung umfasst zwei Schritte: 1) Zuerst müssen die Daten komprimiert und dann 2) indexiert werden. Die damit verbundenen Schwierigkeiten haben es in sich, geht es doch darum, die Daten zu komprimieren, indem man sie «semantisch» strukturiert. Komprimierungsformate wie MPEG, ZIP, JPEG und, neueren Datums, JPEG2000 (siehe Kasten) sind bereits bekannt, sie sind aber zurzeit angesichts der gigantischen Datenmengen, die es zu verarbeiten gilt, noch nicht mehr als «Lösungsembryonen».

Werfen wir beispielsweise einen Blick auf die Archive des Jazzfestivals Montreux. Dabei handelt es sich mehrheitlich um Ton- und Bilddaten. Die EPFL ist zurzeit mit der Archivierung dieser Daten beschäftigt. Wie soll man innert nützlicher Zeit Zugriff auf exakt jene Daten erhalten, die man sucht? Die Antwort ist in aller Munde: mittels «semantischer» Abfrage.

## Die Herausforderung semantische Abfrage

Auch diese Lösung weist zahlreiche Fallstricke auf. Die Inhalte sind in diesem Zusammenhang sehr wichtig. Nun ist aber bekannt, dass die Inhalte Töne, Text, Bild und Video umfassen. Man muss also in verschiedenen Datentypen suchen. Die Suche, die gegenwärtig vorgeschlagen wird, ist unabhängig von der Art und Weise der Daten. Die Lösung heisst also Integration von Daten – erst mit integrierten Daten wird eine zielgerichtete Suche möglich.

Ein anderes Beispiel: virtuelle Sitzungen. Immer häufiger werden Sitzungen virtuell durchgeführt. Diese Tendenz wird sich angesichts der explodierenden Kosten für nicht erneuerbare Energien und damit für örtliche Verschiebungen künftig noch akzentuieren. Die Archivierung dieser Sitzungen (Politik, Wissenschaft, Verbände, Sport, Kultur) wird damit unumgänglich und verlangt nach Lösungen im Bereich Datenkompression und Lagerung/Speicherung. Entsprechende Lösungen sind zurzeit noch nicht greifbar. Über die Lösung für dieses berüchtigte «Binom der Zukunft» beugen sich heute in der ganzen Welt Heerscharen von Forschern ...

## Schlussfolgerung

Die Aufgabe der Forscher hat titanische Ausmasse. Es wird noch eine gewisse Zeit dauern, bis die Berufsleute aus dem Bereich Informationsdokumentation über Instrumente verfügen werden, die es ihnen ermöglichen, ihrer Kundschaft Dienstleistungen im Bereich audiovisuelle Bestände anbieten können, die diesen Namen auch verdienen.

**Anmerkung:**

(1) The Expanding Digital Universe. A Forecast of Worldwide Information Growth Through 2010, unter der Leitung von John F. Gantz, März 2007.

**Kontakt: pierre.vandergheynst@epfl.ch**

*Die Norm JPEG2000*

**JPEG2000 ist ein neues Bildcodierungssystem, das die modernsten Komprimierungstechniken anwendet und auf der Transformierung in Wavelets aufbaut. Die Systemarchitektur ist für eine Vielzahl von Anwendungen (von digitalen Fotoapparaten bis hin zu medizinischen Bildgebungsverfahren und anderen Schlüsselbereichen) geeignet. Die Codierung umfasst Informationen über den Inhalt sowie eine primäre Indexierung.**