ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG ~~on Requirements~~ on Genome Compression and Storage (m40200)

M. Golebiewsky (HITS), J. Delgado (UPC),
M. Mattavelli (EPFL)

Joint AHG with ISO/IEC TC276

Mandates

1. To disseminate the information on the activities jointly carried out by ISO/IEC JTC 1/SC 29/WG 11 and ISO TC 276/WG 5 to other ISO TC and other organizations.
2. To make available on line the MPEG genome data to be used for core experiments (N16726).
3. To carry out the core experiments described in document N16724 based on the results of the answers to the Call for Proposals.
4. To contribute to WD editing.
5. To further validate and possibly extend the benchmark framework for quasi-lossless compression of sequencing quality values defined in N16727 and the identified metrics.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

AHG Activity

- Mainly focused around:
  - Editing the 2 WDs for Part 1 (Transport and Storage of Genomic Information) and Part 2 (Compression of Genomic Information)
  - Running the 2 Core Experiments
- Coordination activities with TC276 and other ISO committees
- More than 80 emails on the reflector and several tenths among CE participants
- 18 Input documents

Summary of CE 2 “Quality Values”

- 4 Participants
- 5 Input documents (M40325, M40222, M40223, M40224, M40292)
- 4 Coding modes can be unified in a decoding syntax (1 lossless, 3 quantized)
- Conclusion
  - Defined decoding syntax goes to WD
  - Syntax elements goes to CE 5 Entropy Coding
  - Performance results need to be validated after entropy coding technology is selected and “Quality Values” compression is applied to classified data.
  - CE 2 is closed.

Summary of CE 5 “Entropy Coding”

- 3 participants reported results on CE 5 (M40270, M40276, M40470, M40272)
- Summary of results:
  - Descriptors streams have been generated and distributed
  - Initial improvements of about 5% in compression reported
  - Significantly faster in speed (t.b.q)
- Conclusion/Consensus:
  - Continue experiments with read descriptors from more test items
  - Add experiments on Quality Values syntax elements
  - Finalize the syntax
  - Add experiments on read names syntax elements

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Input Documents

No.	Title	Authors
M40270	Core Experiment 5 on Genome Compression	Claudio Alberti
M40276	Wellcome Trust Sanger Institute submission to CE 5 on Genome Compression	James Bonfield, Claudio Alberti
M40470	Submission to CE 5 Genome Compression	Claudio Alberti, Filippo Medri
M40271	Use cases for an efficient genomic information representation	N. Guex, C. Iseli, T. Schuepbach, I. Xenarios, D. Renzi, G. Zoia, C. Alberti, M. Mattavelli, J. Delgado, S. Llorente, D. Naro
M40277	Unified representation of genomic reads	Claudio Alberti, Jan Voges, Giorgio Zoia
M40309	Proposal for a new class of read pairs where only one read is mapped	Giorgio Zoia, Daniele Renzi,
M40310	Indexing tools syntax for Genomic Information Transport and Storage	Giorgio Zoia, Daniele Renzi
M40312	Labeling syntax for genomic regions	Daniele Renzi, Giorgio Zoia

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Input documents

No.	Title	Authors
M40313	Genomic transport format and depacketization process flow	Daniele Renzi, Giorgio Zoia
M20382	Reported inconsistencies in aligned genomic data	Claudio Alberti, Giorgio Zoia, Mikel Hernaez
M40222	Unified representation of sequencing quality values	Jan Voges, Claudio Alberti, Mikel Hernaez, Tom Paridaens, James Bonfield, Paolo Ribeca, Jaime Delgado
M40494	Genomic Information Representation. Proposal for Part 3 on Protection, Application Programming Interfaces and Metadata	Jaime Delgado, Silvia Llorente, Daniel Naro, Claudio Alberti
M40224	Core Experiment 2 on Genomic Information Representation results LUH/Stanford/UIUC	Jan Voges, Mikel Hernaez
M40292	Core Experiment 2 on Genomic Information Representation results Stanford/UIUC	Mikel Hernaez

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Input documents

No.	Title	Authors
M40223	Summary of Core Experiment 2 on Genomic Information Representation	Jan Voges
M40272	CE5: Results of syntax compression based on CABAC	Tom Paridaens, Glenn Van Wallendael, Wesley De Neve, Peter Lambert
M40235	CE2 lossy compression evaluation with Crumble	James Bonfield
M40494	Genomic Information Representation. Proposal for Part 3 on Protection, Application Programming Interfaces and Metadata	Jaime Delgado, Silvia Llorente, Daniel Naro, Claudio Alberti

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Recommendations

- Establish the dissemination and information exchange platform with the support of TC276 to include information on other ISO TCs genomic related activities
- Close CE 2
- Update “Use Case Document”
- Continue with the CE 5 extending the scope to QVs syntax elements as defined by CE 2 and read names syntax elements
- Consider a new CE 6 to be able to support multiple alignments and spliced alignments
- Continue the editing of WD of Part 1 with the submitted inputs
- Continue the editing of WD of Part 2 with the current results of CE 2