Assessing Migration Risk for Scientific Formats

Chris Frisz, Sam Waggoner, and Geoffrey Brown

Indiana University Bloomington

Presented 7 December 2011

Overview

- Introduction
  - Motivation
  - Hypothesis
  - Approach
- Background
  - Data set used
  - Formats studied
  - Conversion issues encountered
- Tools written
- Results and discussion
- Conclusions

Motivation

- Many migration tools exist for converting from obsolete to standard data formats.

Motivation

- Many migration tools exist for converting from obsolete to standard data formats.
- Mismatches in source and target formats introduce risk for migration.

Motivation

- Many migration tools exist for converting from obsolete to standard data formats.
- Mismatches in source and target formats introduce risk for migration.
- Automatic tools often fail silently when converting inconsistent features.

Motivation (cont.)

- Creating migration tools is hard.

Motivation (cont.)

- Creating migration tools is hard.
- Development often requires large programs written over a long time.

Motivation (cont.)

- Creating migration tools is hard.
- Development often requires large programs written over a long time.
- Migration is easier using existing tools.

Hypothesis

- Where migration tools already exist, they work well on the majority of data files despite differences in formats.

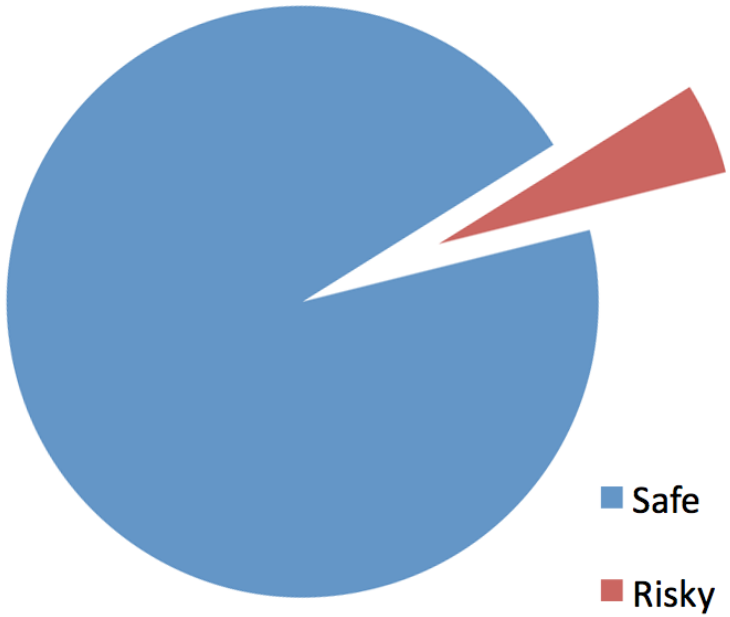
Hypothesis

- Where migration tools already exist, they work well on the majority of data files despite differences in formats.
- The remainder of the files can be identified for rarely-used, risky features.

Hypothesis

- Where migration tools already exist, they work well on the majority of data files despite differences in formats.
- The remainder of the files can be identified for rarely-used, risky features.
- Data files are separated into many that are "safe" to migrate versus a few that are "risky."

Hypothesis (in visual form)



Approach

- Wrote simple and fast analysis tools to categorize files by migration risk through deep inspection.

Approach

- Wrote simple and fast analysis tools to categorize files by migration risk through deep inspection.
- Identified 4 scientific formats with migration risks from a data set of U.S. Government documents.

Approach

- Wrote simple and fast analysis tools to categorize files by migration risk through deep inspection.
- Identified 4 scientific formats with migration risks from a data set of U.S. Government documents.

Approach

- Wrote simple and fast analysis tools to categorize files by migration risk through deep inspection.
- Identified 4 scientific formats with migration risks from a data set of U.S. Government documents.

Found that the vast majority of files show few to no migration risks.

- This comes with some caveats.

Format Overview

- Lotus 1-2-3
  - A formerly popular spreadsheet program migratable to Excel with some calculation differences.

Format Overview

- Lotus 1-2-3
  - A formerly popular spreadsheet program migratable to Excel with some calculation differences.
- CDF and netCDF
  - Array-based data formats with common roots but evolved with some different data representation and encoding features.

Format Overview

- Lotus 1-2-3
  - A formerly popular spreadsheet program migratable to Excel with some calculation differences.
- CDF and netCDF
  - Array-based data formats with common roots but evolved with some different data representation and encoding features.
- HDF
  - Hierarchical format for relating data artifacts that underwent significant changes from version 4 to 5.

Data Set

- Set of 2747 CD-ROM images from the United States Government Printing Office.

Data Set

- Set of 2747 CD-ROM images from the United States Government Printing Office.
- Thirty-six (36) images contained 14,022 Lotus 1-2-3, version 1 files.
- Sixty-eight (68) images contained 61,247 CDF files.
- Four (4) images contained 3,162 netCDF files.
- Two (2) images contained 2,213 HDF files.

Data Set (cont.)

- Lotus 1-2-3 files published from many different U.S. agencies:
  - CDC
  - Census Bureau
  - Dept. of Education
  - Office of Business and Management
- CDF and HDF files primarily from NASA.
- NetCDF files came from University of Maine, Dept. of Climatology.

Formats – Lotus 1-2-3

- Primary spreadsheet application used in the 1980s and early 1990s, but was supplanted by Microsoft Excel.

Formats – Lotus 1-2-3

- Primary spreadsheet application used in the 1980s and early 1990s, but was supplanted by Microsoft Excel.
- Microsoft provided conversion from 1-2-3 to Excel through 2003.

Formats – Lotus 1-2-3

- Primary spreadsheet application used in the 1980s and early 1990s, but was supplanted by Microsoft Excel.
- Microsoft provided conversion from 1-2-3 to Excel through 2003.
- Differences between the formats were documented by Microsoft and retrieved from knowledge base articles.

Formats – Lotus 1-2-3 – Conversion issues

- Operations calculated differently
  - @MOD
  - @VLOOKUP
  - @HLOOKUP

Formats – Lotus 1-2-3 – Conversion issues (cont.)

- Exponentiation (^) and unary negation (-) differ in order of operations.

Formats – Lotus 1-2-3 – Conversion issues (cont.)

- Exponentiation (^) and unary negation (-) differ in order of operations.
  - Exponentiation was evaluated first in Lotus 1-2-3.

Formats – Lotus 1-2-3 – Conversion issues (cont.)

- Exponentiation (^) and unary negation (-) differ in order of operations.
  - Exponentiation was evaluated first in Lotus 1-2-3.
  - Negation was evaluated first in Excel.

Formats – Lotus 1-2-3 – Example

- In Lotus 1-2-3:
  -4^2

Formats – Lotus 1-2-3 – Example

- In Lotus 1-2-3:
  -4^2

Formats – Lotus 1-2-3 – Example

- In Lotus 1-2-3:
  $-4^2 = -16$

Formats – Lotus 1-2-3 – Example

- In Lotus 1-2-3:
  $-4^2 = -16$
- In Excel:
  -4^2

Formats – Lotus 1-2-3 – Example

- In Lotus 1-2-3:
  $-4^2 = -16$
- In Excel:
  -4^2

Formats – Lotus 1-2-3 – Example

- In Lotus 1-2-3:
  $-4^2 = -16$
- In Excel:
  $-4^2 = 16$

Formats – Lotus 1-2-3 – Example

- In Lotus 1-2-3:
  $-4^2 = -16$
- In Excel:
  $-4^2 = 16$
- Traditional mathematical order of operations favors Lotus.

Formats – Lotus 1-2-3 – Conversion issues (cont.)

- Comparison/logical operators (i.e. = or #and#) and string concatenation (&) also differ in order of operations.

Formats – Lotus 1-2-3 – Conversion issues (cont.)

- Comparison/logical operators (i.e. = or #and#) and string concatenation (&) also differ in order of operations.
  - Comparison and logical operators were evaluated first in Lotus 1-2-3.

Formats – Lotus 1-2-3 – Conversion issues (cont.)

- Comparison/logical operators (i.e. = or #and#) and string concatenation (&) also differ in order of operations.
  - Comparison and logical operators were evaluated first in Lotus 1-2-3.
  - Concatenation was evaluated first in Excel.

Formats – Lotus 1-2-3 – Conversion Issues – Example

- In Lotus 1-2-3:
  “Fo” & “o” = “Foo”

Formats – Lotus 1-2-3 – Conversion Issues – Example

- In Lotus 1-2-3:
  “Fo” & “o” = “Foo”

Formats – Lotus 1-2-3 – Conversion Issues – Example

- In Lotus 1-2-3:
  "Fo" & "o" = "Foo" $\rightarrow$ **False**

Formats – Lotus 1-2-3 – Conversion Issues – Example

- In Lotus 1-2-3:
  "Fo" & "o" = "Foo" $\rightarrow$ **False**
- In Excel:
  "Fo" & "o" = "Foo"

Formats – Lotus 1-2-3 – Conversion Issues – Example

- In Lotus 1-2-3:
  "Fo" & "o" = "Foo" $\rightarrow$ **False**
- In Excel:
  "Fo" & "o" = "Foo"

Formats – Lotus 1-2-3 – Conversion Issues – Example

- In Lotus 1-2-3:
  "Fo" & "o" = "Foo" $\rightarrow$ **False**
- In Excel:
  "Fo" & "o" = "Foo" $\rightarrow$ **True**

Formats – CDF and netCDF

- CDF and netCDF are both file formats utilized for multidimensional data.

Formats – CDF and netCDF

- CDF and netCDF are both file formats utilized for multidimensional data.
- Often used to represent image, climate, and elevation data.

Formats – CDF/netCDF Layout

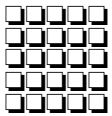
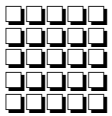
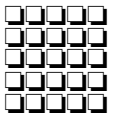
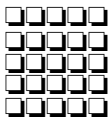
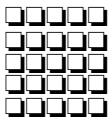
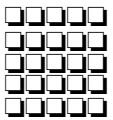
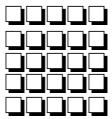
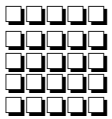
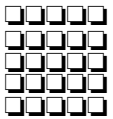
Record Number	rVariable 1	rVariable 2	. . .	rVariable n
1				
2				
3				

Image courtesy of NASA/Goddard Space Flight Center

Formats – CDF/netCDF Layout

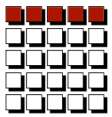
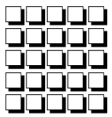
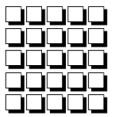
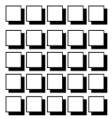
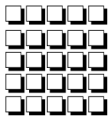
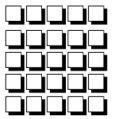
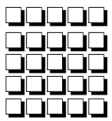
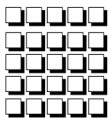
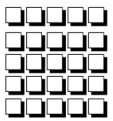
Record Number	rVariable 1	rVariable 2	. . .	rVariable n
1				
2				
3				

Image courtesy of NASA/Goddard Space Flight Center

Formats – CDF/netCDF Layout

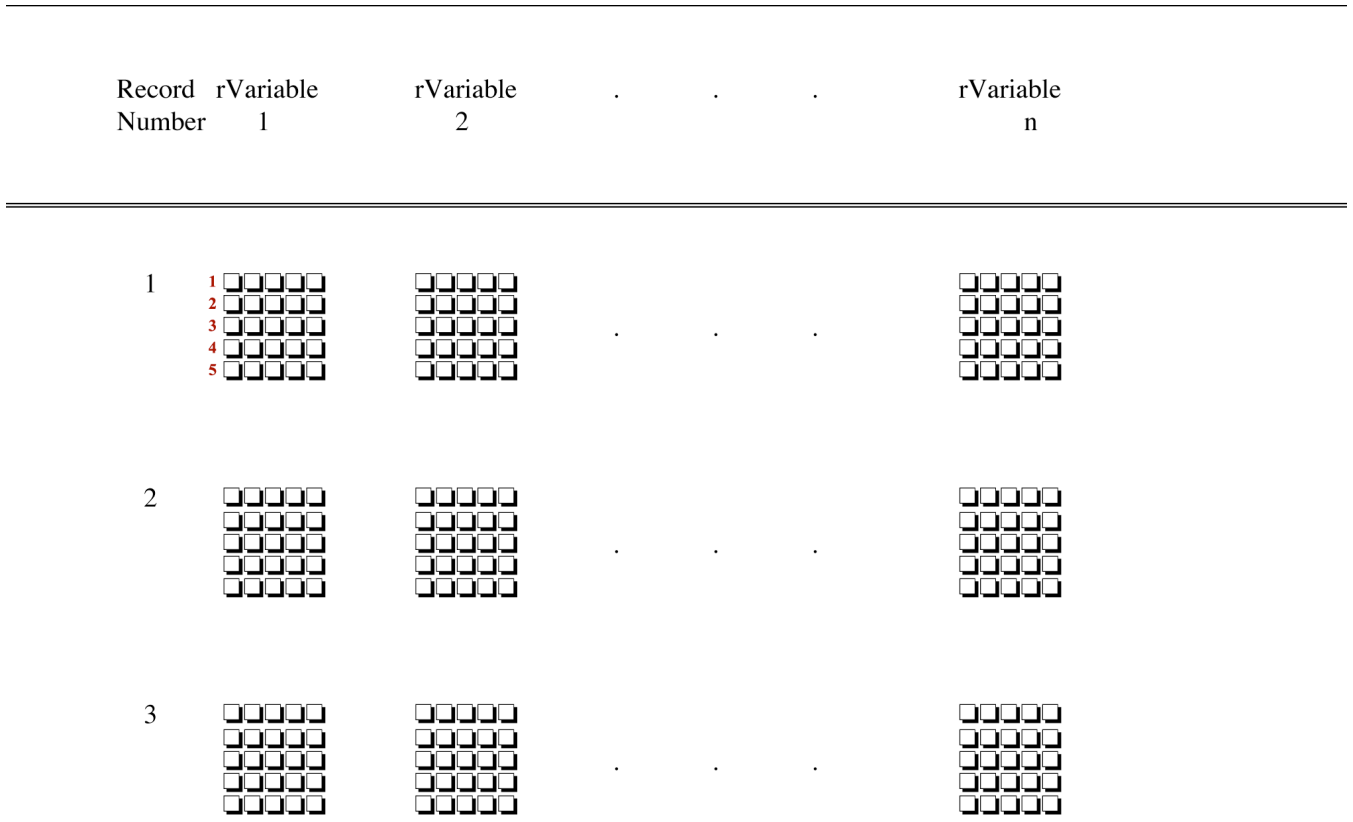


Image courtesy of NASA/Goddard Space Flight Center

Formats – CDF/netCDF Layout

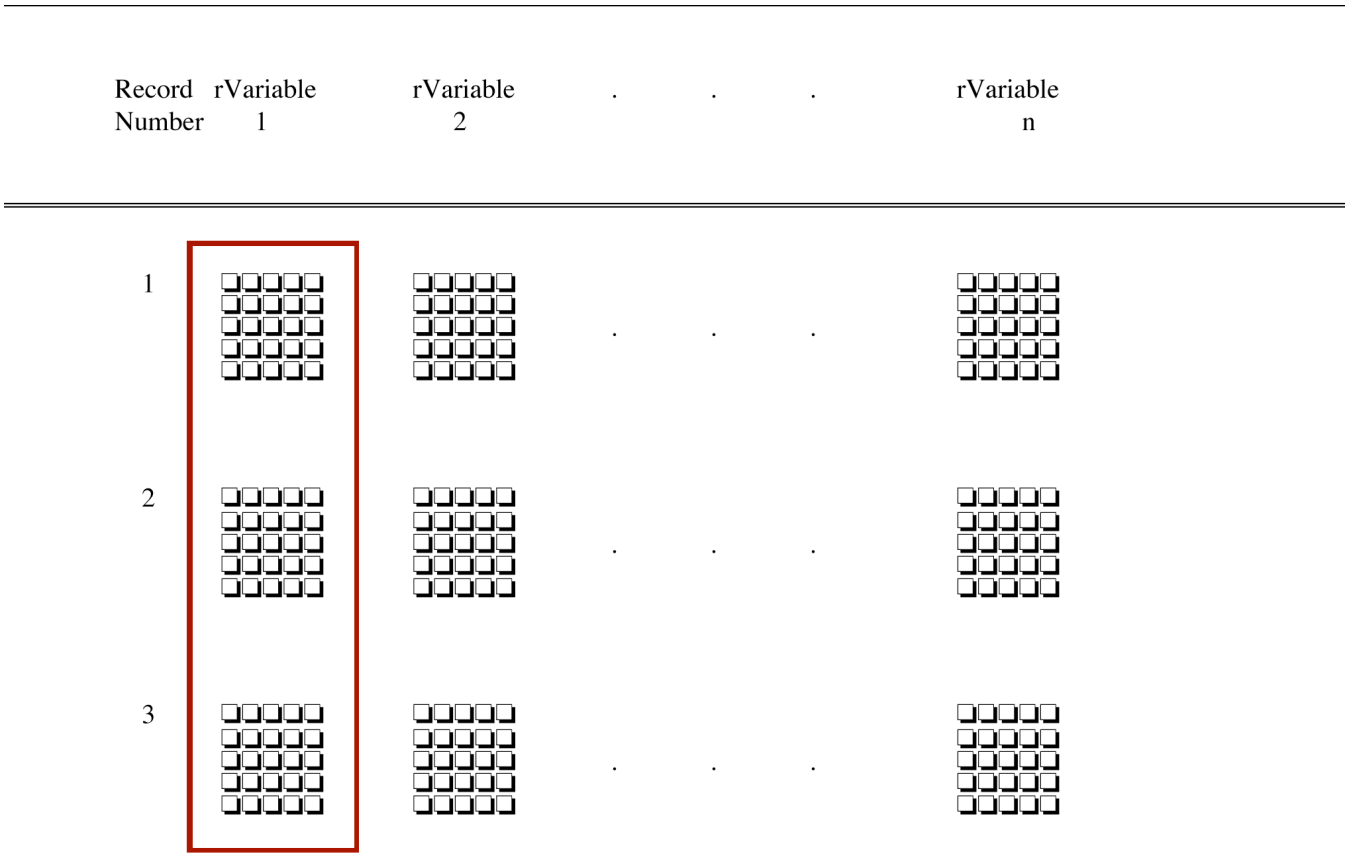


Image courtesy of NASA/Goddard Space Flight Center

Formats – CDF/netCDF – Background

- CDF originally developed by NASA.

Formats – CDF/netCDF – Background

- CDF originally developed by NASA.
- NetCDF developed later by NCAR based on the CDF.

Formats – CDF/netCDF – Background

- CDF originally developed by NASA.
- NetCDF developed later by NCAR based on the CDF.
- Both formats still currently supported.

Formats – CDF/netCDF – Background (cont.)

- Separate development allowed for evolution of different features.

Formats – CDF/netCDF – Background (cont.)

- Separate development allowed for evolution of different features.
- Overall functionality remained similar.

Formats – CDF/netCDF – Background (cont.)

- Separate development allowed for evolution of different features.
- Overall functionality remained similar.
- Primary conversion path between CDF and netCDF was through NASA's Data Translation Web Service (DTWS).

Formats – CDF – Conversion Issues

- Features present in CDF, not in netCDF:

Formats – CDF – Conversion Issues

- Features present in CDF, not in netCDF:
  - Multi-file format for organizing variables into different files.

Formats – CDF – Conversion Issues

- Features present in CDF, not in netCDF:
  - Multi-file format for organizing variables into different files.
  - Native-mode encoding for faster data access on particular system architectures.

Formats – CDF – Conversion Issues

- Features present in CDF, not in netCDF:
  - Multi-file format for organizing variables into different files.
  - Native-mode encoding for faster data access on particular system architectures.
  - Epoch data type for high-resolution time data.

Formats – CDF – Conversion Issues

- Features present in CDF, not in netCDF:
  - Multi-file format for organizing variables into different files.
  - Native-mode encoding for faster data access on particular system architectures.
  - Epoch data type for high-resolution time data.
- Multi-file and native-mode differences were identified in CDF documentation.

Formats – CDF – Conversion Issues

- Features present in CDF, not in netCDF:
  - Multi-file format for organizing variables into different files.
  - Native-mode encoding for faster data access on particular system architectures.
  - Epoch data type for high-resolution time data.
- Multi-file and native-mode differences were identified in CDF documentation.
- Epoch data type mismatch was discovered through DTWS source code review.

Formats – netCDF – Conversion Issues

- Features present in netCDF, not in CDF:

Formats – netCDF – Conversion Issues

- Features present in netCDF, not in CDF:
  - Descriptive named dimensions usable for data access

Formats – netCDF – Conversion Issues

- Features present in netCDF, not in CDF:
  - Descriptive named dimensions usable for data access
  - Support for up 32 dimensions per variable (versus CDF's 10)

Formats – netCDF – Conversion Issues

- Features present in netCDF, not in CDF:
  - Descriptive named dimensions usable for data access
  - Support for up 32 dimensions per variable (versus CDF's 10)
- Named dimensions mismatch was documented in NASA's CDF FAQ.

Formats – netCDF – Conversion Issues

- Features present in netCDF, not in CDF:
  - Descriptive named dimensions usable for data access
  - Support for up 32 dimensions per variable (versus CDF's 10)
- Named dimensions mismatch was documented in NASA's CDF FAQ.
- Maximum dimension mismatch was discovered through netCDF API code review.

Formats – HDF

- Hierarchical data format for relating and interacting with hetergenous data sets.

Formats – HDF

- Hierarchical data format for relating and interacting with hetergenous data sets.
- Organized similarly to Unix file system with Vgroups like directories and Vdata like files.

Formats – HDF layout

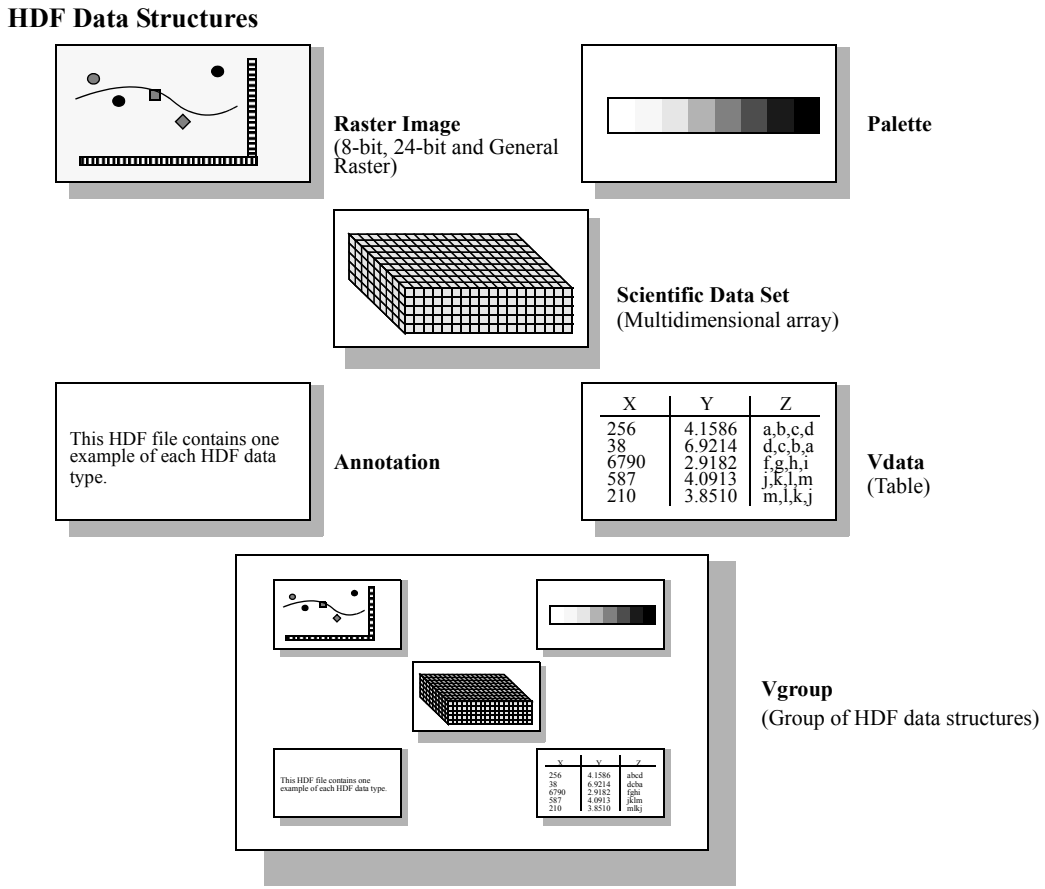


Image courtesy of the HDF Group.

Formats – HDF – Background

- Developed by the National Center for Supercomputing Applications.

Formats – HDF – Background

- Developed by the National Center for Supercomputing Applications.
- Support provided by the HDF Group.

Formats – HDF – Background

- Developed by the National Center for Supercomputing Applications.
- Support provided by the HDF Group.
- Most recent version was HDF5.

Formats – HDF – Background (cont.)

- Previous versions were backwards compatible.

Formats – HDF – Background (cont.)

- Previous versions were backwards compatible.
- HDF5 drastically changed data model and broke backwards compatibility.

Formats – HDF – Background (cont.)

- Previous versions were backwards compatible.
- HDF5 drastically changed data model and broke backwards compatibility.
- HDF Group provided both conversion API and automatic tool.

Formats – HDF – Conversion Issues

- Merging Vgroups with elements sharing the same name resulted in renaming of one element.

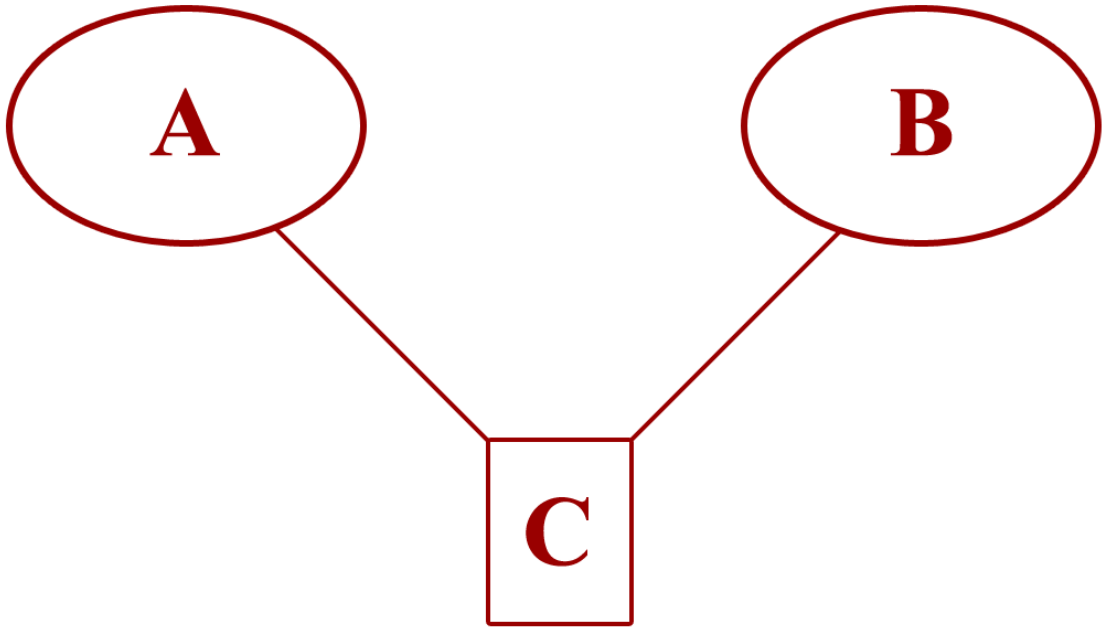
Formats – HDF – Conversion Issues

- Merging Vgroups with elements sharing the same name resulted in renaming of one element.
  - This was only relevant for manual conversion.

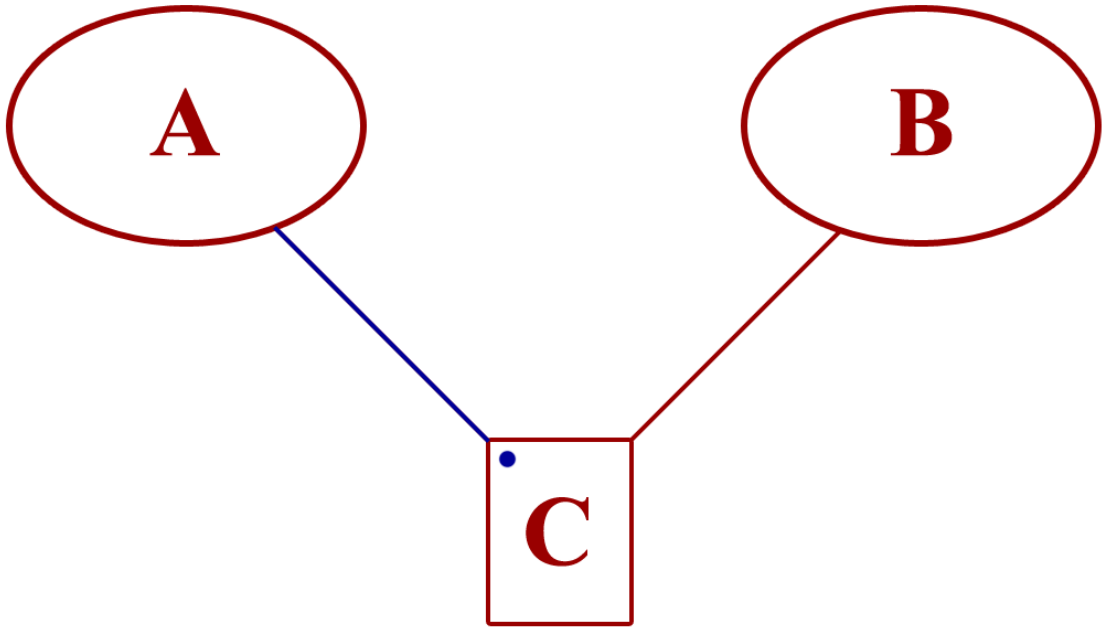
Formats – HDF – Conversion Issues

- Merging Vgroups with elements sharing the same name resulted in renaming of one element.
  - This was only relevant for manual conversion.
- Data object shared between Vgroups were copied on conversion.

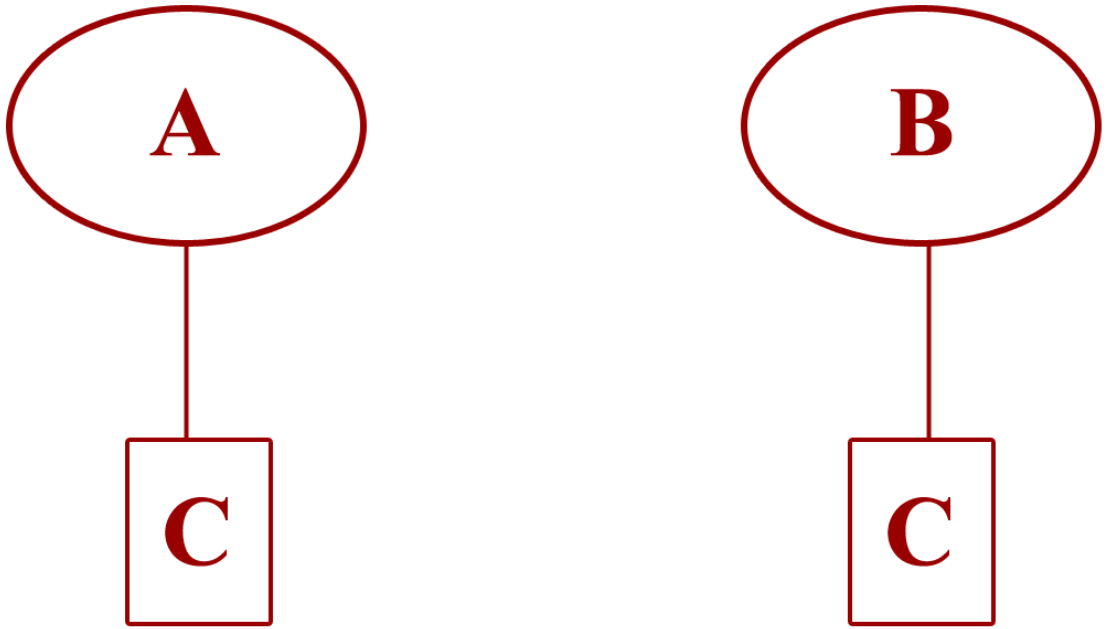
Formats – HDF – Conversion Issues – Example



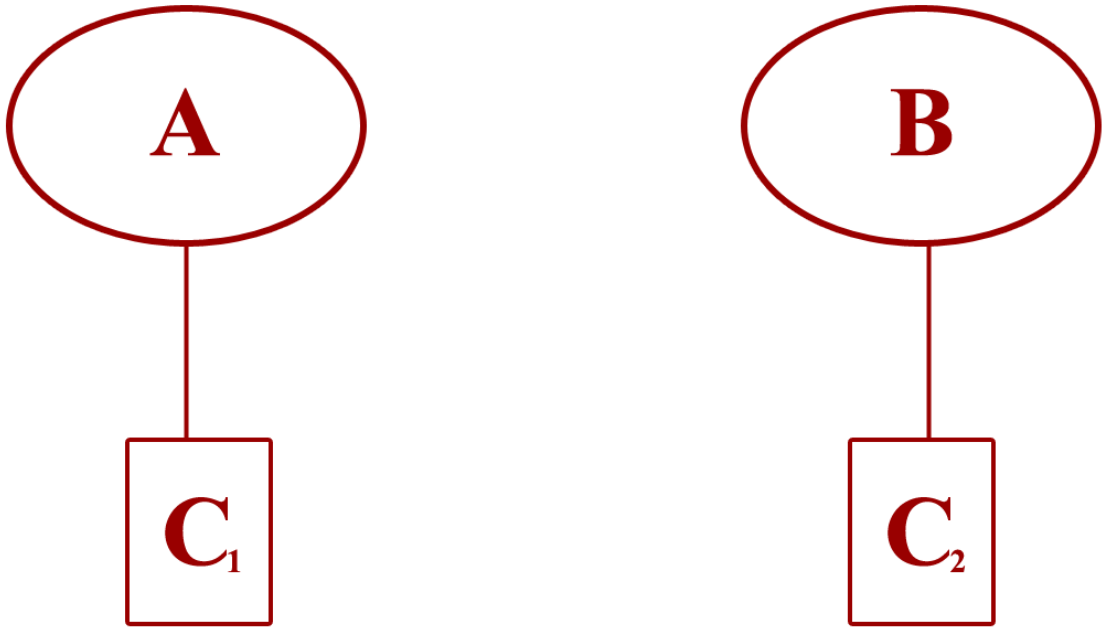
Formats – HDF – Conversion Issues – Example



Formats – HDF – Conversion Issues – Example



Formats – HDF – Conversion Issues – Example



Formats – HDF – Conversion Issues

- Merging Vgroups with elements sharing the same name resulted in renaming of one element.
  - This was only relevant for manual conversion.
- Data object shared between Vgroups were copied on conversion.
- Unnamed data objects were given default names

Formats – HDF – Conversion Issues

- Merging Vgroups with elements sharing the same name resulted in renaming of one element.
  - This was only relevant for manual conversion.
- Data object shared between Vgroups were copied on conversion.
- Unnamed data objects were given default names

The HDF Group documented all of these issues for the HDF4-to-HDF5 conversion API and automated tool.

Tools – Lotus 1-2-3

- We wrote a C program to traverse 1-2-3 files and parse formulas.

Tools – Lotus 1-2-3

- We wrote a C program to traverse 1-2-3 files and parse formulas.
- It identified presence of @MOD, @VLOOKUP, or @HLOOKUP in formulas.

Tools – Lotus 1-2-3

- We wrote a C program to traverse 1-2-3 files and parse formulas.
- It identified presence of @MOD, @VLOOKUP, or @HLOOKUP in formulas.
- The program also conservatively reported presence of both exponentiation and negation or logical/comparison operators and string concatenation.

Tools – Lotus 1-2-3 (cont.)

- Tool consisted of approximately 500 lines.

Tools – Lotus 1-2-3 (cont.)

- Tool consisted of approximately 500 lines.
- Processed our entire data set in less than 15 mintues.

Tools – CDF and netCDF

- We wrote C programs for each CDF and netCDF.

Tools – CDF and netCDF

- We wrote C programs for each CDF and netCDF.
- CDF program consisted of 300 lines using the version 3.3.0 API from NASA.

Tools – CDF and netCDF

- We wrote C programs for each CDF and netCDF.
- CDF program consisted of 300 lines using the version 3.3.0 API from NASA.
- NetCDF program was 150 lines using the version 4.1.3 API from Unidata.

Tools – CDF and netCDF

- We wrote C programs for each CDF and netCDF.
- CDF program consisted of 300 lines using the version 3.3.0 API from NASA.
- NetCDF program was 150 lines using the version 4.1.3 API from Unidata.
- Processed entire 61,000-file data set in 55 minutes.

Tools – CDF and netCDF

- We wrote C programs for each CDF and netCDF.
- CDF program consisted of 300 lines using the version 3.3.0 API from NASA.
- NetCDF program was 150 lines using the version 4.1.3 API from Unidata.
- Processed entire 61,000-file data set in 55 minutes.
- NetCDF tool exhibited similar performance.

Tools – HDF

- Yet again, wrote a C program.

Tools – HDF

- Yet again, wrote a C program.
- Written in 900 lines using the 4.2.6 API from the HDF Group.

Tools – HDF

- Yet again, wrote a C program.
- Written in 900 lines using the 4.2.6 API from the HDF Group.
- This tool was longer because of large number of interfaces.

Tools – HDF

- Yet again, wrote a C program.
- Written in 900 lines using the 4.2.6 API from the HDF Group.
- This tool was longer because of large number of interfaces.
- Processed all HDF files in our data set within 1.5 minutes.

Results – Lotus 1-2-3

- We ran our analysis tool on 14,022 version 1 files.

Results – Lotus 1-2-3

- We ran our analysis tool on 14,022 version 1 files.
- It detected a single file containing 7 formulas with possible order of operations mismatches between 1-2-3 and Excel.

Results – Lotus 1-2-3 (cont.)

Example formula from the file:

@IF($EJ$85=“NA”, +“ ”&$EJ$85,+$EJ$85)

Results – Lotus 1-2-3 (cont.)

Example formula from the file:

@IF($EJ$85=“NA”, +“ ”&$EJ$85,+$EJ$85)

- The other six also followed this form.

Results – Lotus 1-2-3 (cont.)

Example formula from the file:

@IF($EJ$85=“NA”, +“ ”&$EJ$85,+$EJ$85)

- The other six also followed this form.
- Logical comparison and string concatenation appeared in the same formula, but would not conflict if converted to Excel.

Discussion – Lotus 1-2-3

- The vast majority of files can be converted conventially without risk.

Discussion – Lotus 1-2-3

- The vast majority of files can be converted conventially without risk.
- Only a few files may require a more robust conversion process or by-hand translation.

Discussion – Lotus 1-2-3

- The vast majority of files can be converted conventially without risk.
- Only a few files may require a more robust conversion process or by-hand translation.
- All 14,022 files in our data set could have been converted without risk after manually verifying a single file.

Results – CDF

- Our tool ran on 61,247 CDF version 2 files.

Results – CDF

- Our tool ran on 61,247 CDF version 2 files.
- 14,574 (23.8%) files with no potential conversion risk to netCDF.

Results – CDF

- Our tool ran on 61,247 CDF version 2 files.
- 14,574 (23.8%) files with no potential conversion risk to netCDF.
- 46,669 (76.2%) utilized the Epoch data type.

Results – CDF

- Our tool ran on 61,247 CDF version 2 files.
- 14,574 (23.8%) files with no potential conversion risk to netCDF.
- 46,669 (76.2%) utilized the Epoch data type.
- 4 files used multi-file format.

Results – CDF

- Our tool ran on 61,247 CDF version 2 files.
- 14,574 (23.8%) files with no potential conversion risk to netCDF.
- 46,669 (76.2%) utilized the Epoch data type.
- 4 files used multi-file format.
- There were no files which used native encoding.

Discussion – CDF

- Use of Epoch data type was prevalent (76.2%).

Discussion – CDF

- Use of Epoch data type was prevalent (76.2%).
- CDF API included functions to convert Epochs to strings.

Discussion – CDF

- Use of Epoch data type was prevalent (76.2%).
- CDF API included functions to convert Epochs to strings.
  - DTWS tool used this method during conversion.

Discussion – CDF

- Use of Epoch data type was prevalent (76.2%).
- CDF API included functions to convert Epochs to strings.
  - DTWS tool used this method during conversion.
  - Tools for converting date string formats are widely available (i.e. Unix).

Discussion – CDF

- Use of Epoch data type was prevalent (76.2%).
- CDF API included functions to convert Epochs to strings.
  - DTWS tool used this method during conversion.
  - Tools for converting date string formats are widely available (i.e. Unix).
- Multi-file format was handled by DTWS tools, despite its rare appearance.

Discussion – CDF (cont.)

- The results indicated a minimal migration risk for converting CDF to netCDF, which supported our hypothesis.

Results – netCDF

- We ran our tool on 3,162 netCDF files.

Results – netCDF

- We ran our tool on 3,162 netCDF files.
- All files included named dimensions.

Results – netCDF

- We ran our tool on 3,162 netCDF files.
- All files included named dimensions.
  - We expected this result.

Results – netCDF

- We ran our tool on 3,162 netCDF files.
- All files included named dimensions.
  - We expected this result.
- No files included variables with more than CDF's maximum 10 dimensions.

Results – netCDF

- We ran our tool on 3,162 netCDF files.
- All files included named dimensions.
  - We expected this result.
- No files included variables with more than CDF's maximum 10 dimensions.
  - This indicated it was a rare feature.

Discussion – netCDF

- Dimensions names (present in all netCDF datasets) were not saved in conversion.

Discussion – netCDF

- Dimensions names (present in all netCDF datasets) were not saved in conversion.
- This represented actual metadata loss.

Discussion – netCDF

- Dimensions names (present in all netCDF datasets) were not saved in conversion.
- This represented actual metadata loss.
- Though raw data was preserved in conversion, this conflicted with our hypothesis.

Discussion – netCDF (cont.)

- One possible solution was to save names in a separate metadata file.

Discussion – netCDF (cont.)

- One possible solution was to save names in a separate metadata file.
- We were not aware of an existing tool to do this.

Results – HDF

- Tool ran on 352 HDF3 and 1,861 HDF4 (2,213 total) files.

Results – HDF

- Tool ran on 352 HDF3 and 1,861 HDF4 (2,213 total) files.
- 324 (14.6%) files with no conversion risks.

Results – HDF

- Tool ran on 352 HDF3 and 1,861 HDF4 (2,213 total) files.
- 324 (14.6%) files with no conversion risks.
- 1,891 (85.4%) with multiple Vgroups containing objects with the same name.

Results – HDF

- Tool ran on 352 HDF3 and 1,861 HDF4 (2,213 total) files.
- 324 (14.6%) files with no conversion risks.
- 1,891 (85.4%) with multiple Vgroups containing objects with the same name.
- 1,889 (85.4%) with data objects shared between Vgroups.

Results – HDF

- Tool ran on 352 HDF3 and 1,861 HDF4 (2,213 total) files.
- 324 (14.6%) files with no conversion risks.
- 1,891 (85.4%) with multiple Vgroups containing objects with the same name.
- 1,889 (85.4%) with data objects shared between Vgroups.
- No unnamed data objects.

Discussion – HDF

- Duplicate Vdata object names were irrelevant for automatic conversion.

Discussion – HDF

- Duplicate Vdata object names were irrelevant for automatic conversion.
- Shared object copying broke data relationships from the source files.

Discussion – HDF (cont.)

- Issues would not manifest when converting for purely archival reasons.

Discussion – HDF (cont.)

- Issues would not manifest when converting for purely archival reasons.
- This overall supported our hypothesis with a caveat.

Conclusions

- Existing conversion tools could safely convert the vast majority of files in general.

Conclusions

- Existing conversion tools could safely convert the vast majority of files in general.
- Caveats:

Conclusions

- Existing conversion tools could safely convert the vast majority of files in general.
- Caveats:
  - NetCDF-to-CDF conversion loses metadata and requires a separate solution.

Conclusions

- Existing conversion tools could safely convert the vast majority of files in general.
- Caveats:
  - NetCDF-to-CDF conversion loses metadata and requires a separate solution.
  - HDF4-to-HDF5 conversion breaks data relationships and is only completely safe for archival purposes.

Conclusions (cont.)

- The results for our data set overall supported our hypothesis.

Conclusions (cont.)

- The results for our data set overall supported our hypothesis.
- Our findings supported use of simple and fast tools for migration risk analysis

Conclusions (cont.)

- The results for our data set overall supported our hypothesis.
- Our findings supported use of simple and fast tools for migration risk analysis
- Open formats (e.g. CDF, netCDF, HDF) are easier to analyze than proprietary ones (i.e. Lotus 1-2-3).

Acknowledgements

The authors would like to gratefully acknowledge the support of the Data to Insight Center, a partnership of the School of Informatics and Computing, Digital Libraries and Pervasive Technology Institute at Indiana University. This research funded in part by a grant provided by the Lilly Endowment Inc.

Time for questions and comments