Collection Structure

Breakout Session

B. R. Barkstrom

ESIP 2014 Winter Meeting

January 9, 2014

Background and Goals

- Background
  - Sept. telecon disclosed confusion about how to handle inventory granularity.
  - Telecon session time limits and phone connections made developing group understanding difficult.
- Goals of This Session
  1. Clarify categories of stable objects in an archive inventory
  2. Clarify inventory accounting approach for stable objects
  3. Suggest policies for inventory accounting of *indistinguishable* objects
  4. Suggest an approach to identify *scientifically equivalent* data in different files
  5. Suggest policies for inventory accounting of *scientifically equivalent* data

Background and Goals

- Background
  - Sept. telecon disclosed confusion about how to handle inventory granularity.
  - Telecon session time limits and phone connections made developing group understanding difficult.
- Goals of This Session
  1. Clarify categories of stable objects in an archive inventory
  2. Clarify inventory accounting approach for stable objects
  3. Suggest policies for inventory accounting of *indistinguishable* objects
  4. Suggest an approach to identify *scientifically equivalent* data in different files
  5. Suggest policies for inventory accounting of *scientifically equivalent* data

Background and Goals

- Background
  - Sept. telecon disclosed confusion about how to handle inventory granularity.
  - Telecon session time limits and phone connections made developing group understanding difficult.
- Goals of This Session
  1. Clarify categories of stable objects in an archive inventory
  2. Clarify inventory accounting approach for stable objects
  3. Suggest policies for inventory accounting of *indistinguishable* objects
  4. Suggest an approach to identify *scientifically equivalent* data in different files
  5. Suggest policies for inventory accounting of *scientifically equivalent* data

Background and Goals

- Background
  - Sept. telecon disclosed confusion about how to handle inventory granularity.
  - Telecon session time limits and phone connections made developing group understanding difficult.
- Goals of This Session
  1. Clarify categories of stable objects in an archive inventory
  2. Clarify inventory accounting approach for stable objects
  3. Suggest policies for inventory accounting of *indistinguishable* objects
  4. Suggest an approach to identify *scientifically equivalent* data in different files
  5. Suggest policies for inventory accounting of *scientifically equivalent* data

Background and Goals

- Background
  - Sept. telecon disclosed confusion about how to handle inventory granularity.
  - Telecon session time limits and phone connections made developing group understanding difficult.
- Goals of This Session
  1. Clarify categories of stable objects in an archive inventory
  2. Clarify inventory accounting approach for stable objects
  3. Suggest policies for inventory accounting of *indistinguishable* objects
  4. Suggest an approach to identify *scientifically equivalent* data in different files
  5. Suggest policies for inventory accounting of *scientifically equivalent* data

Background and Goals

- Background
  - Sept. telecon disclosed confusion about how to handle inventory granularity.
  - Telecon session time limits and phone connections made developing group understanding difficult.
- Goals of This Session
  1. Clarify categories of stable objects in an archive inventory
  2. Clarify inventory accounting approach for stable objects
  3. Suggest policies for inventory accounting of *indistinguishable* objects
  4. Suggest an approach to identify *scientifically equivalent* data in different files
  5. Suggest policies for inventory accounting of *scientifically equivalent* data

Background and Goals

- Background
  - Sept. telecon disclosed confusion about how to handle inventory granularity.
  - Telecon session time limits and phone connections made developing group understanding difficult.
- Goals of This Session
  1. Clarify categories of stable objects in an archive inventory
  2. Clarify inventory accounting approach for stable objects
  3. Suggest policies for inventory accounting of *indistinguishable* objects
  4. Suggest an approach to identify *scientifically equivalent* data in different files
  5. Suggest policies for inventory accounting of *scientifically equivalent* data

Discussion Etiquette

- Need to hear from everyone
- Time allotted as equally as possible
- Will need to decide on next actions of group

Terminology for Objects

Definition (Stable Object)

Object that persists *more* than one month.

Definition (Archive Collection)

A set of stable objects.

Definition (Inventory System)

An accounting system to track stable objects in an archive.

Definition (Inventoried Object)

A stable object tracked by an archive's inventory system.

Terminology for Objects

Definition (Stable Object)

Object that persists *more* than one month.

Definition (Archive Collection)

A set of stable objects.

Definition (Inventory System)

An accounting system to track stable objects in an archive.

Definition (Inventoried Object)

A stable object tracked by an archive's inventory system.

Terminology for Objects

Definition (Stable Object)

Object that persists *more* than one month.

Definition (Archive Collection)

A set of stable objects.

Definition (Inventory System)

An accounting system to track stable objects in an archive.

Definition (Inventoried Object)

A stable object tracked by an archive's inventory system.

Terminology for Objects

Definition (Stable Object)

Object that persists *more* than one month.

Definition (Archive Collection)

A set of stable objects.

Definition (Inventory System)

An accounting system to track stable objects in an archive.

Definition (Inventoried Object)

A stable object tracked by an archive's inventory system.

Object Classification Terminology

Definition (Attribute)

Property of an object. Decision that an object has a particular attribute must be *independently verifiable*.

Definition (Category or Formal Concept)

Object subset with a consistent set of attributes that distinguish this subset from other subsets.

Definition (Taxonomic Classification)

An algorithm to identify which formal concepts in an archive collection include an object. The algorithm tests whether the object has the attributes of the formal concept.

Object Classification Terminology

Definition (Attribute)

Property of an object. Decision that an object has a particular attribute must be *independently verifiable*.

Definition (Category or Formal Concept)

Object subset with a consistent set of attributes that distinguish this subset from other subsets.

Definition (Taxonomic Classification)

An algorithm to identify which formal concepts in an archive collection include an object. The algorithm tests whether the object has the attributes of the formal concept.

Object Classification Terminology

Definition (Attribute)

Property of an object. Decision that an object has a particular attribute must be *independently verifiable*.

Definition (Category or Formal Concept)

Object subset with a consistent set of attributes that distinguish this subset from other subsets.

Definition (Taxonomic Classification)

An algorithm to identify which formal concepts in an archive collection include an object. The algorithm tests whether the object has the attributes of the formal concept.

Preliminary Object Classification for an Archive Collection

- Physical Objects
  - Geological Specimens
  - Fossilized Paleontological Specimens
  - Biological Specimens
  - Physical Documents
    - Informal
    - Formal
- Digital Objects
  - Data Files
  - Documentation Files

Preliminary Object Classification for an Archive Collection

- Physical Objects
  - Geological Specimens
  - Fossilized Paleontological Specimens
  - Biological Specimens
  - Physical Documents
    - Informal
    - Formal
- Digital Objects
  - Data Files
  - Documentation Files

Preliminary Object Classification for an Archive Collection

- Physical Objects
  - Geological Specimens
  - Fossilized Paleontological Specimens
  - Biological Specimens
  - Physical Documents
    - Informal
    - Formal
- Digital Objects
  - Data Files
  - Documentation Files

Preliminary Object Classification for an Archive Collection

- Physical Objects
  - Geological Specimens
  - Fossilized Paleontological Specimens
  - Biological Specimens
  - Physical Documents
    - Informal
    - Formal
- Digital Objects
  - Data Files
  - Documentation Files

Example 1 - A Geological Specimen?

How would I identify an object as a geological specimen?

- Attributes
  - Rock or soil sample (has mineral grains)
  - Originally found in nature
  - Not made in lab
  - Encoded human language information not part of object
- Objects Included in Category
  - Rock in display case
  - Fragment from a rock core sample
- Objects Excluded from Category
  - Manufactured Diamond
  - Marble slab chiseled with Lincoln's *Gettysburg Address*

Example 1 - A Geological Specimen?

How would I identify an object as a geological specimen?

- Attributes
  - Rock or soil sample (has mineral grains)
  - Originally found in nature
  - Not made in lab
  - Encoded human language information not part of object
- Objects Included in Category
  - Rock in display case
  - Fragment from a rock core sample
- Objects Excluded from Category
  - Manufactured Diamond
  - Marble slab chiseled with Lincoln's *Gettysburg Address*

Example 1 - A Geological Specimen?

How would I identify an object as a geological specimen?

- Attributes
  - Rock or soil sample (has mineral grains)
  - Originally found in nature
  - Not made in lab
  - Encoded human language information not part of object
- Objects Included in Category
  - Rock in display case
  - Fragment from a rock core sample
- Objects Excluded from Category
  - Manufactured Diamond
  - Marble slab chiseled with Lincoln's *Gettysburg Address*

Example 2 - A Digital File Containing Data?

How would I identify an object as a digital file containing Earth science data?

- Attributes
  - Intangible object with an abstract representation as a finite array of bits
  - Bit array interpretable as serialized sequence of interpretation tokens
  - Encoded information recorded by *human observation* of phenomenon or event in Earth's environment
  - Encoded information recorded by *automated instrument measurement* of phenomenon or event in Earth's environment

Example 2 - A Digital File Containing Data?

How would I identify an object as a digital file containing Earth science data?

- Attributes
  - Intangible object with an abstract representation as a finite array of bits
  - Bit array interpretable as serialized sequence of interpretation tokens
  - Encoded information recorded by *human observation* of phenomenon or event in Earth's environment
  - Encoded information recorded by *automated instrument measurement* of phenomenon or event in Earth's environment

Example 2 - A Digital File Containing Data?

How would I identify an object as a digital file containing Earth science data?

- Attributes
  - Intangible object with an abstract representation as a finite array of bits
  - Bit array interpretable as serialized sequence of interpretation tokens
  - Encoded information recorded by *human observation* of phenomenon or event in Earth's environment
  - Encoded information recorded by *automated instrument measurement* of phenomenon or event in Earth's environment

Example 2 - A Digital File Containing Data?

How would I identify an object as a digital file containing Earth science data?

- Attributes
  - Intangible object with an abstract representation as a finite array of bits
  - Bit array interpretable as serialized sequence of interpretation tokens
  - Encoded information recorded by *human observation* of phenomenon or event in Earth's environment
  - Encoded information recorded by *automated instrument measurement* of phenomenon or event in Earth's environment

Example 2 - A Digital File Containing Data?

How would I identify an object as a digital file containing Earth science data?

- Attributes
  - Intangible object with an abstract representation as a finite array of bits
  - Bit array interpretable as serialized sequence of interpretation tokens
  - Encoded information recorded by *human observation* of phenomenon or event in Earth's environment
  - Encoded information recorded by *automated instrument measurement* of phenomenon or event in Earth's environment

Example 2 - A Digital File Containing Data? (cont'd.)

- Attributes (cont'd.)
  - Encoded information recorded as output from an algorithm or network of algorithms ingesting digital data from human observation or automated instrument measurement of Earth's environment
- Example Objects
  - GHCN Monthly Precipitation File: `v1.prcp` obtained 2010/2/27:3:52 PM EST from NCDC
- Objects Excluded
  - EOSDIS Metadata Database [transient state; probably not Earth observation data]

Example 2 - A Digital File Containing Data? (cont'd.)

- Attributes (cont'd.)
  - Encoded information recorded as output from an algorithm or network of algorithms ingesting digital data from human observation or automated instrument measurement of Earth's environment
- Example Objects
  - GHCN Monthly Precipitation File: `v1.prcp` obtained 2010/2/27:3:52 PM EST from NCDC
- Objects Excluded
  - EOSDIS Metadata Database [transient state; probably not Earth observation data]

Example 2 - A Digital File Containing Data? (cont'd.)

- Attributes (cont'd.)
  - Encoded information recorded as output from an algorithm or network of algorithms ingesting digital data from human observation or automated instrument measurement of Earth's environment
- Example Objects
  - GHCN Monthly Precipitation File: `v1.prcp` obtained 2010/2/27:3:52 PM EST from NCDC
- Objects Excluded
  - EOSDIS Metadata Database [transient state; probably not Earth observation data]

Example 2 - A Digital File Containing Data? (cont'd.)

- Attributes (cont'd.)
  - Encoded information recorded as output from an algorithm or network of algorithms ingesting digital data from human observation or automated instrument measurement of Earth's environment
- Example Objects
  - GHCN Monthly Precipitation File: `v1.prcp` obtained 2010/2/27:3:52 PM EST from NCDC
- Objects Excluded
  - EOSDIS Metadata Database [transient state; probably not Earth observation data]

Example 2 - A Digital File Containing Data? (cont'd.)

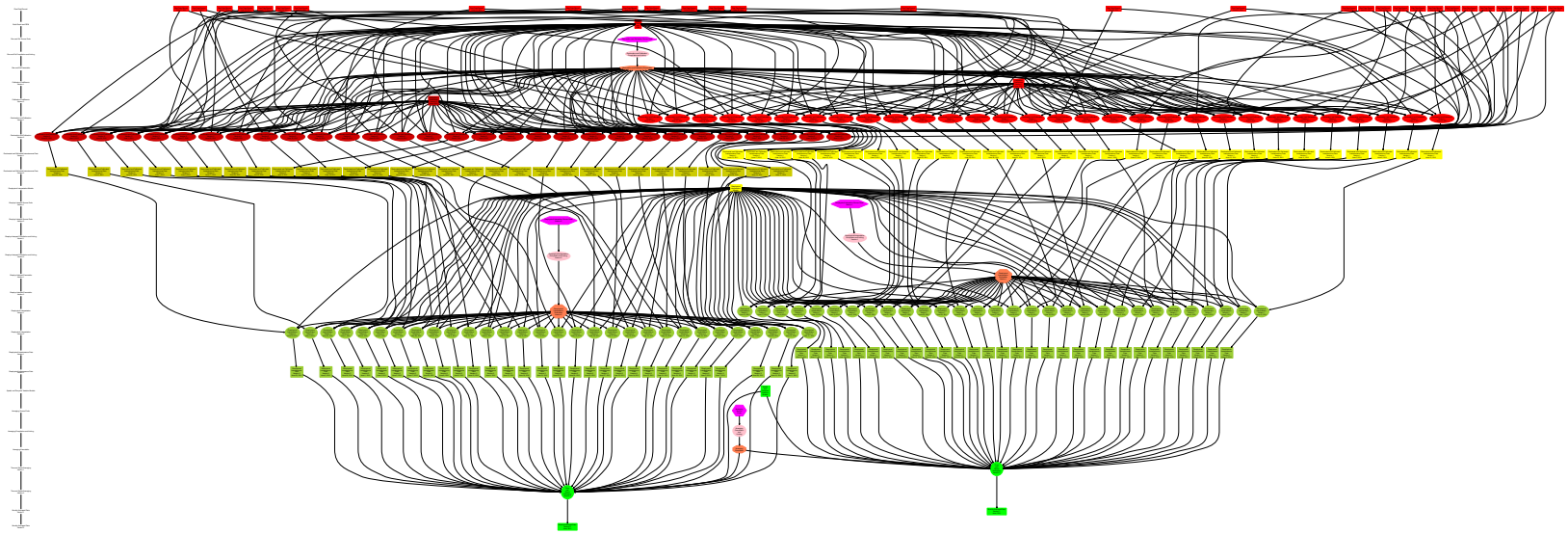
- Attributes (cont'd.)
  - Encoded information recorded as output from an algorithm or network of algorithms ingesting digital data from human observation or automated instrument measurement of Earth's environment
- Example Objects
  - GHCN Monthly Precipitation File: `v1.prcp` obtained 2010/2/27:3:52 PM EST from NCDC
- Objects Excluded
  - EOSDIS Metadata Database [transient state; probably not Earth observation data]

Example 2 - A Digital File Containing Data? (cont'd.)

- Attributes (cont'd.)
  - Encoded information recorded as output from an algorithm or network of algorithms ingesting digital data from human observation or automated instrument measurement of Earth's environment
- Example Objects
  - GHCN Monthly Precipitation File: `v1.prcp` obtained 2010/2/27:3:52 PM EST from NCDC
- Objects Excluded
  - EOSDIS Metadata Database [transient state; probably not Earth observation data]

Example 2 - Testing Whether a Digital File Contains Measurements in its Production Chain

Production Graph for Level 0 to 3 of One Month of Data



- Collections appear as horizontal groupings
- Red → Level 0; Yellow → L 1; Green → L 2; Cyan → L 3

Other Examples

Other categorization examples included in Handout:

- Fossilized Paleontological Specimen
- Biological Specimen
- Informal Physical Document
- Formal Physical Document
- Digital Documentation File

For categorization of other examples, see <ExampleDefinitions.pdf>

Classification Concepts Discussion

Key Questions:

1. The taxonomic categorization aims at being able to have independently verifiable categories, regardless of disciplinary community or language base. *Do the examples suggest that this goal is a feasible approach to categorization*?
2. *How could you improve the preliminary categorization*?

Reminder:

- *Identify missing categories, as well as missing attributes or tests for categories*.
- *Provide written comments with such suggestions*.

Bookkeeping Concepts

Concept Use Double-entry bookkeeping as basis for the inventory management system.

Definition (Account)

Object that tracks state of one or more inventoried objects.

Definition (Journal)

Uses balanced transactions to track inventoried object state changes.

Definition (Ledger)

Historical log of state changes for inventoried objects.

Definition (Chart of Accounts)

Hierarchical list of all accounts in the inventory *by category*.

Bookkeeping Concepts

Concept Use Double-entry bookkeeping as basis for the inventory management system.

Definition (Account)

Object that tracks state of one or more inventoried objects.

Definition (Journal)

Uses balanced transactions to track inventoried object state changes.

Definition (Ledger)

Historical log of state changes for inventoried objects.

Definition (Chart of Accounts)

Hierarchical list of all accounts in the inventory *by category*.

Bookkeeping Concepts

Concept Use Double-entry bookkeeping as basis for the inventory management system.

Definition (Account)

Object that tracks state of one or more inventoried objects.

Definition (Journal)

Uses balanced transactions to track inventoried object state changes.

Definition (Ledger)

Historical log of state changes for inventoried objects.

Definition (Chart of Accounts)

Hierarchical list of all accounts in the inventory *by category*.

Bookkeeping Concepts

Concept Use Double-entry bookkeeping as basis for the inventory management system.

Definition (Account)

Object that tracks state of one or more inventoried objects.

Definition (Journal)

Uses balanced transactions to track inventoried object state changes.

Definition (Ledger)

Historical log of state changes for inventoried objects.

Definition (Chart of Accounts)

Hierarchical list of all accounts in the inventory *by category*.

Bookkeeping Concept Cont'd.

Observations

- It seems natural to track each stable object with an inventory account.
- Usually, each account has a name and an account number, rather like an OID label.
- Accountants usually keep track of the value of goods for which the organization is responsible. It seems reasonable to track the *number of objects* in the account if an account tracks more than one object.

Bookkeeping Concept Cont'd.

Observations

- It seems natural to track each stable object with an inventory account.
- Usually, each account has a name and an account number, rather like an OID label.
- Accountants usually keep track of the value of goods for which the organization is responsible. It seems reasonable to track the *number of objects* in the account if an account tracks more than one object.

Bookkeeping Concept Cont'd.

Observations

- It seems natural to track each stable object with an inventory account.
- Usually, each account has a name and an account number, rather like an OID label.
- Accountants usually keep track of the value of goods for which the organization is responsible. It seems reasonable to track the *number of objects* in the account if an account tracks more than one object.

Inventory Management Discussion

Key Questions:

1. *What policy seems most appropriate for accounting for an archive's inventoried objects*?
   - Each stable object should have its own account **OR**
   - Stable object replicas should be placed in the same account
2. *What recommendations do we have about the organization of the chart of accounts*?
   - Some accounts may need a deep hierarchy: Rock core well accounts contain box accounts, which contain rock core fragment accounts.
   - The US Government has a Standard General Ledger with a chart of accounts. ▸ see <ChartOfAccountsUSSGL.html>

Replication

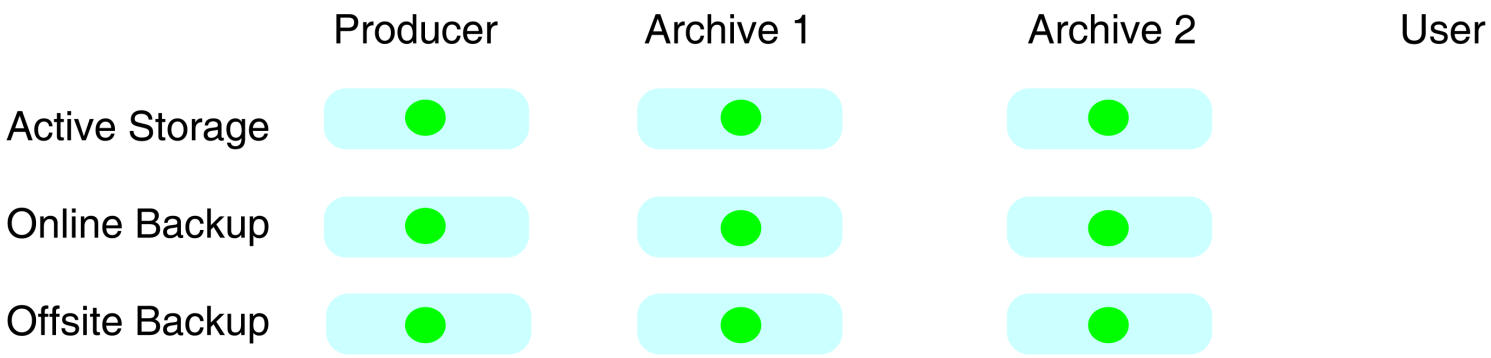
- Software often copies digital files from one location to another.
- This type of software copy produces digital files that are *identical* based on the copies having the same cryptographic digests.
- We can call such copies *replicas*.

Definition (Replica of a Digital Object)

Digital object that has the same cryptographic digest as another digital object.

Test Case: Replicas in Several Locations

Case 1: Data Producer, Archives, and User



Does it matter to a user which replica he or she obtains?

Discussion of Replica Inventory Accounting

Key Questions:

1. *How should an archive handle replicas*?
   - Each inventoried object should have its own account **OR**
   - Inventoried object replicas should belong in the same account.
2. Users can treat replicas as identical copies. *Why should we have a policy that assumes there is only one source of replicated digital files or replicated documents*?

Scientific Equivalence

A Tentative Definition of Scientific Equivalence

- Two sets of data values are *scientifically equivalent* if a knowledgeable person could use produce the same result or inference from either set.

Some Sources of Divergent Results:

- Failing to use an item correctly because the user couldn't understand what it was (mislabeling a measurand)
- Using an item incorrectly because the units were incorrect, e.g. mistaking a temperature in K for a temperature in F
- Drawing an incorrect conclusion or inference because of production input errors or misunderstanding of observation sampling patterns

Scientific Equivalence

A Tentative Definition of Scientific Equivalence

- Two sets of data values are *scientifically equivalent* if a knowledgeable person could use produce the same result or inference from either set.

Some Sources of Divergent Results:

- Failing to use an item correctly because the user couldn't understand what it was (mislabeling a measurand)
- Using an item incorrectly because the units were incorrect, e.g. mistaking a temperature in K for a temperature in F
- Drawing an incorrect conclusion or inference because of production input errors or misunderstanding of observation sampling patterns

Scientific Equivalence

A Tentative Definition of Scientific Equivalence

- Two sets of data values are *scientifically equivalent* if a knowledgeable person could use produce the same result or inference from either set.

Some Sources of Divergent Results:

- Failing to use an item correctly because the user couldn't understand what it was (mislabeling a measurand)
- Using an item incorrectly because the units were incorrect, e.g. mistaking a temperature in K for a temperature in F
- Drawing an incorrect conclusion or inference because of production input errors or misunderstanding of observation sampling patterns

Scientific Equivalence

A Tentative Definition of Scientific Equivalence

- Two sets of data values are *scientifically equivalent* if a knowledgeable person could use produce the same result or inference from either set.

Some Sources of Divergent Results:

- Failing to use an item correctly because the user couldn't understand what it was (mislabeling a measurand)
- Using an item incorrectly because the units were incorrect, e.g. mistaking a temperature in K for a temperature in F
- Drawing an incorrect conclusion or inference because of production input errors or misunderstanding of observation sampling patterns

Representation Information and Formatting

Not all bits in a digital file contain data. Some bits may provide information on to how to interpret bit sequences.

Definition (Representation Information)

The information that maps a Data Object into more meaningful concepts. [OAIS RM, p. 1-14]

Bit Arrays, Token Arrays, and Token Strings

A user's program can interpret a digital file in three important ways:

1. As an array of bits
2. As an array of tokens, with each token having
   - A number of bits
   - An interpretation of the bits in the token
3. As an array of (an array of tokens) – a token string

Bit Array Interpretations

Bit Arrays, Token Arrays, and Token Strings

A user's program can interpret a digital file in three important ways:

1. As an array of bits
2. As an array of tokens, with each token having
   - A number of bits
   - An interpretation of the bits in the token
3. As an array of (an array of tokens) – a token string

Bit Array Interpretations

Bit Arrays, Token Arrays, and Token Strings

A user's program can interpret a digital file in three important ways:

1. As an array of bits
2. As an array of tokens, with each token having
   - A number of bits
   - An interpretation of the bits in the token
3. As an array of (an array of tokens) – a token string

Bit Array Interpretations

Bit Arrays, Token Arrays, and Token Strings

A user's program can interpret a digital file in three important ways:

1. As an array of bits
2. As an array of tokens, with each token having
   - A number of bits
   - An interpretation of the bits in the token
3. As an array of (an array of tokens) – a token string

Bit Array Interpretations

Bit Arrays, Token Arrays, and Token Strings (cont'd.)

- If the bits in a token are interpreted as *text*, then the token interpretation is a *character*: 'a' or 'A'
- If the bits in a token are interpreted as *numerical values*, then the token interpretation is a *number*: '0' or '3' (using an ASCII character value to represent the number)
- If the tokens in a token string are interpreted as a *character string*, then the token string interpretation is a *word* or a *phrase*: 'Who' or 'Who am I?'
- If the tokens in a token string are *numerical values*, then the sequence of tokens may be an array or list or even more complicated data structures.

Bit Arrays, Token Arrays, and Token Strings (cont'd.)

- If the bits in a token are interpreted as *text*, then the token interpretation is a *character*: 'a' or 'A'
- If the bits in a token are interpreted as *numerical values*, then the token interpretation is a *number*: '0' or '3' (using an ASCII character value to represent the number)
- If the tokens in a token string are interpreted as a *character string*, then the token string interpretation is a *word* or a *phrase*: 'Who' or 'Who am I?'
- If the tokens in a token string are *numerical values*, then the sequence of tokens may be an array or list or even more complicated data structures.

Bit Arrays, Token Arrays, and Token Strings (cont'd.)

- If the bits in a token are interpreted as *text*, then the token interpretation is a *character*: 'a' or 'A'
- If the bits in a token are interpreted as *numerical values*, then the token interpretation is a *number*: '0' or '3' (using an ASCII character value to represent the number)
- If the tokens in a token string are interpreted as a *character string*, then the token string interpretation is a *word* or a *phrase*: 'Who' or 'Who am I?'
- If the tokens in a token string are *numerical values*, then the sequence of tokens may be an array or list or even more complicated data structures.

Bit Arrays, Token Arrays, and Token Strings (cont'd.)

- If the bits in a token are interpreted as *text*, then the token interpretation is a *character*: 'a' or 'A'
- If the bits in a token are interpreted as *numerical values*, then the token interpretation is a *number*: '0' or '3' (using an ASCII character value to represent the number)
- If the tokens in a token string are interpreted as a *character string*, then the token string interpretation is a *word* or a *phrase*: 'Who' or 'Who am I?'
- If the tokens in a token string are *numerical values*, then the sequence of tokens may be an array or list or even more complicated data structures.

Methods of Formatting Digital Files

There are at least three ways of providing the tokens that interpret the bit array:

- Embed Formatting in Source Code – classic FORTRAN or C FORMAT statements
- Embed Labels for Partitioning Bit Array in the File – XML
- Provide an External Input Stream for Partitioning and Interpretation – Data Stream input to XML SAX document
- **How the file receives its interpretation should not influence the interpretation of its data values!**

Methods of Formatting Digital Files

There are at least three ways of providing the tokens that interpret the bit array:

- Embed Formatting in Source Code – classic FORTRAN or C FORMAT statements
- Embed Labels for Partitioning Bit Array in the File – XML
- Provide an External Input Stream for Partitioning and Interpretation – Data Stream input to XML SAX document
- **How the file receives its interpretation should not influence the interpretation of its data values!**

Methods of Formatting Digital Files

There are at least three ways of providing the tokens that interpret the bit array:

- Embed Formatting in Source Code – classic FORTRAN or C FORMAT statements
- Embed Labels for Partitioning Bit Array in the File – XML
- Provide an External Input Stream for Partitioning and Interpretation – Data Stream input to XML SAX document
- **How the file receives its interpretation should not influence the interpretation of its data values!**

Methods of Formatting Digital Files

There are at least three ways of providing the tokens that interpret the bit array:

- Embed Formatting in Source Code – classic FORTRAN or C FORMAT statements
- Embed Labels for Partitioning Bit Array in the File – XML
- Provide an External Input Stream for Partitioning and Interpretation – Data Stream input to XML SAX document
- **How the file receives its interpretation should not influence the interpretation of its data values!**

Methods of Formatting Digital Files

There are at least three ways of providing the tokens that interpret the bit array:

- Embed Formatting in Source Code – classic FORTRAN or C FORMAT statements
- Embed Labels for Partitioning Bit Array in the File – XML
- Provide an External Input Stream for Partitioning and Interpretation – Data Stream input to XML SAX document
- **How the file receives its interpretation should not influence the interpretation of its data values!**

Permutations of Token String Ordering

Text and Numerical Values treat permuted order of tokens differently:

- Character Strings - *permutations scramble meaning* [‘One’ is not the same as ‘enO’]
- Numerical Value Token Strings - *permutations do not necessarily scramble meaning as long as the permutation mapping is available*

Permutations of Token String Ordering

Text and Numerical Values treat permuted order of tokens differently:

- Character Strings - *permutations scramble meaning* [‘One’ is not the same as ‘enO’]
- Numerical Value Token Strings - *permutations do not necessarily scramble meaning as long as the permutation mapping is available*

Permutations of Token String Ordering

Text and Numerical Values treat permuted order of tokens differently:

- Character Strings - *permutations scramble meaning* [‘One’ is not the same as ‘enO’]
- Numerical Value Token Strings - *permutations do not necessarily scramble meaning as long as the permutation mapping is available*

Character Strings as Labels

Possible Labels for a Measurement Parameter (or Measurand Label)

- `OCEAN COLOR`
- `Ocean Color`
- `ocean color`
- `Ocean Colour`
- `OceanColor`

- Human users seem likely to understand these labels as roughly equivalent
- Computers seem likely to treat them as distinct – unless the software that uses them adds the complications needed to handle them as aliases

Character Strings as Labels

Possible Labels for a Measurement Parameter (or Measurand Label)

- `OCEAN COLOR`
- `Ocean Color`
- `ocean color`
- `Ocean Colour`
- `OceanColor`

- Human users seem likely to understand these labels as roughly equivalent
- Computers seem likely to treat them as distinct – unless the software that uses them adds the complications needed to handle them as aliases

Character Strings as Labels

Possible Labels for a Measurement Parameter (or Measurand Label)

- `OCEAN COLOR`
- `Ocean Color`
- `ocean color`
- `Ocean Colour`
- `OceanColor`

- Human users seem likely to understand these labels as roughly equivalent
- Computers seem likely to treat them as distinct – unless the software that uses them adds the complications needed to handle them as aliases

User Dialects and Character Strings

- There are many lists of labels:
  - Essential Climate Variables
  - Names of parameters from GCMD, Unidata CF Profile, NPOESS IORD Environmental Data Records
  - Tags from various standards, such as ISO 19115, FGDC, MathML, . . .
- A combined list of names for measureands shows little commonality - the number of exact matches out of about 2,000 names is less than 10

To access the spreadsheet with these values,  see <AlphabetizedParameterValidsV5.ods>

Discussion on Label Equivalence

Key Questions:

1. *What should we recommend about case sensitivity and string delimiters* (or semantic heterogeneity in general)?
2. *Do we have a "best practices" suggestion on how an archive should handle the divergence between human pattern matching and computer pattern matching in text?*
   - Insist users learn a controlled vocabulary.
   - Automatically expand terms in such situations as query keyword expansion.
   - Provide hints or alternatives that the user can interpret.
   - Other approaches?

Discussion on Label Equivalence

Key Questions:

1. *What should we recommend about case sensitivity and string delimiters* (or semantic heterogeneity in general)?
2. *Do we have a "best practices" suggestion on how an archive should handle the divergence between human pattern matching and computer pattern matching in text?*
   - Insist users learn a controlled vocabulary.
   - Automatically expand terms in such situations as query keyword expansion.
   - Provide hints or alternatives that the user can interpret.
   - Other approaches?

Discussion on Label Equivalence

Key Questions:

1. *What should we recommend about case sensitivity and string delimiters* (or semantic heterogeneity in general)?
2. *Do we have a "best practices" suggestion on how an archive should handle the divergence between human pattern matching and computer pattern matching in text*?
   - Insist users learn a controlled vocabulary.
   - Automatically expand terms in such situations as query keyword expansion.
   - Provide hints or alternatives that the user can interpret.
   - Other approaches?

Numerical Value Equivalence

- There are many binary representations of numerical values:
  - ASCII character string (‘831’)
  - Text string (‘eight hundred thirty one’)
  - Binary representation (‘01101100111’)
- Users can transform different binary representations into scientifically equivalent numerical values

Discussion on Numerical Value Equivalence

Key Questions:

1. *How should an archive handle objects that differ in numerical format, but that have corresponding sequences of numerical value*?
   - Treat the different objects as being different.
   - Treat the different objects as being scientifically equivalent.
2. *If two objects have scientifically equivalent numerical values in the same sequence, is one more "authentic", and if so why*?
   - They are equally authentic.
   - They are not equally authentic. Authenticity could be independently verified by some algorithm operating on some records.

Note on Bitmaps

Classic Windows Bitmaps (`.bmp` files)

- Have two parts to the file
  - A header containing the array size and a color palette
  - An array of one-byte numbers
- Scientific data values are in the array
- If the array represents geolocated data,
  - geolocation is implicit in the order of the token string with the data values **OR**
  - geolocation is stored somewhere, perhaps outside the file

Note on Bitmaps

Classic Windows Bitmaps (`.bmp` files)

- Have two parts to the file
  - A header containing the array size and a color palette
  - An array of one-byte numbers
- Scientific data values are in the array
- If the array represents geolocated data,
  - geolocation is implicit in the order of the token string with the data values **OR**
  - geolocation is stored somewhere, perhaps outside the file

Note on Bitmaps

Classic Windows Bitmaps (`.bmp` files)

- Have two parts to the file
  - A header containing the array size and a color palette
  - An array of one-byte numbers
- Scientific data values are in the array
- If the array represents geolocated data,
  - geolocation is implicit in the order of the token string with the data values **OR**
  - geolocation is stored somewhere, perhaps outside the file

Scientifically Equivalent Arrays - An Example

Consider an array of surface types:

- 18 surface types
- each type represented as a one-byte number
- numbers in a global, equal-area array (each pixel has an area of about 972.9 km^2)

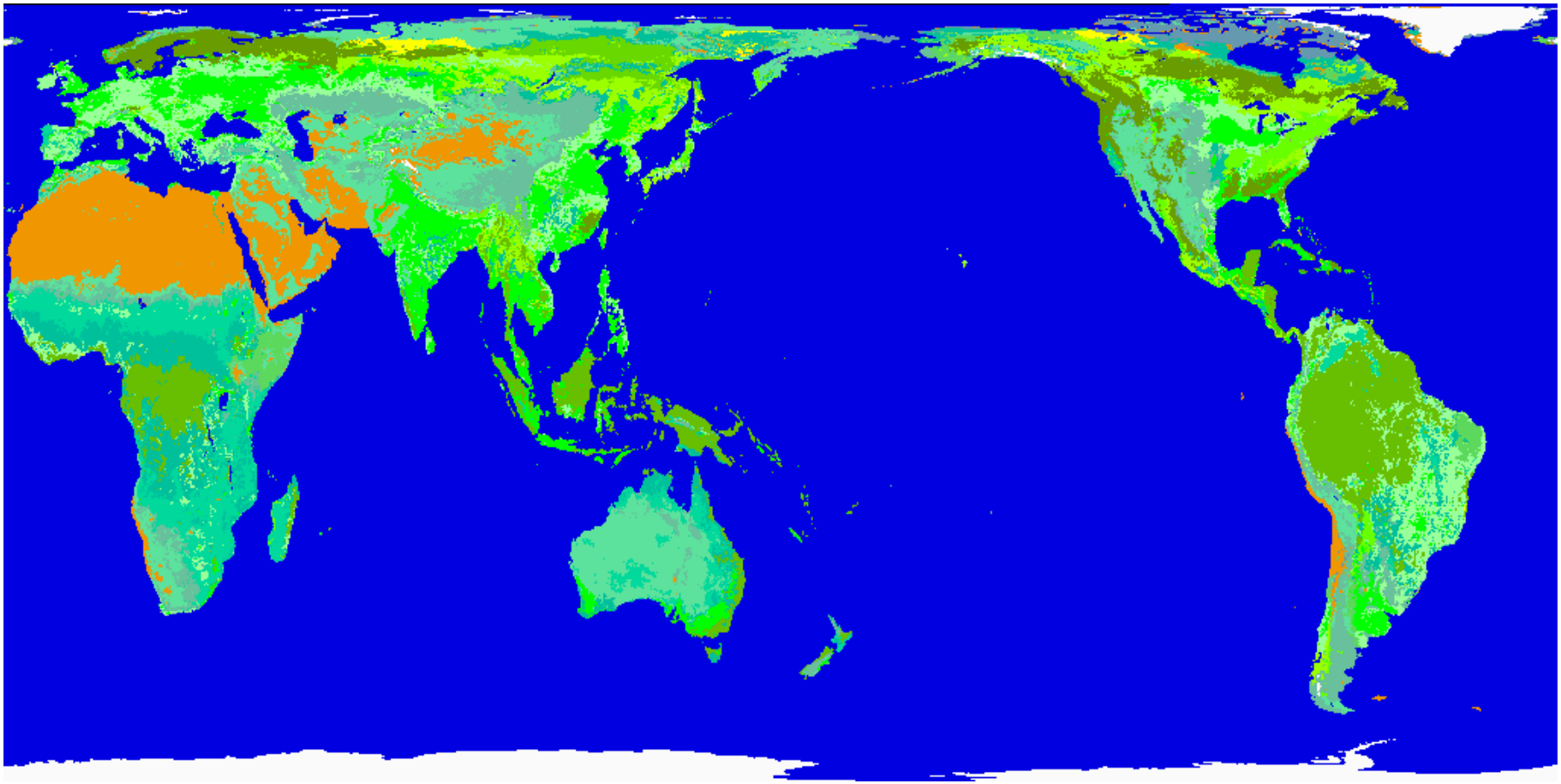
Scientific task:

- Compute the area of Deciduous Broadleaf Forest

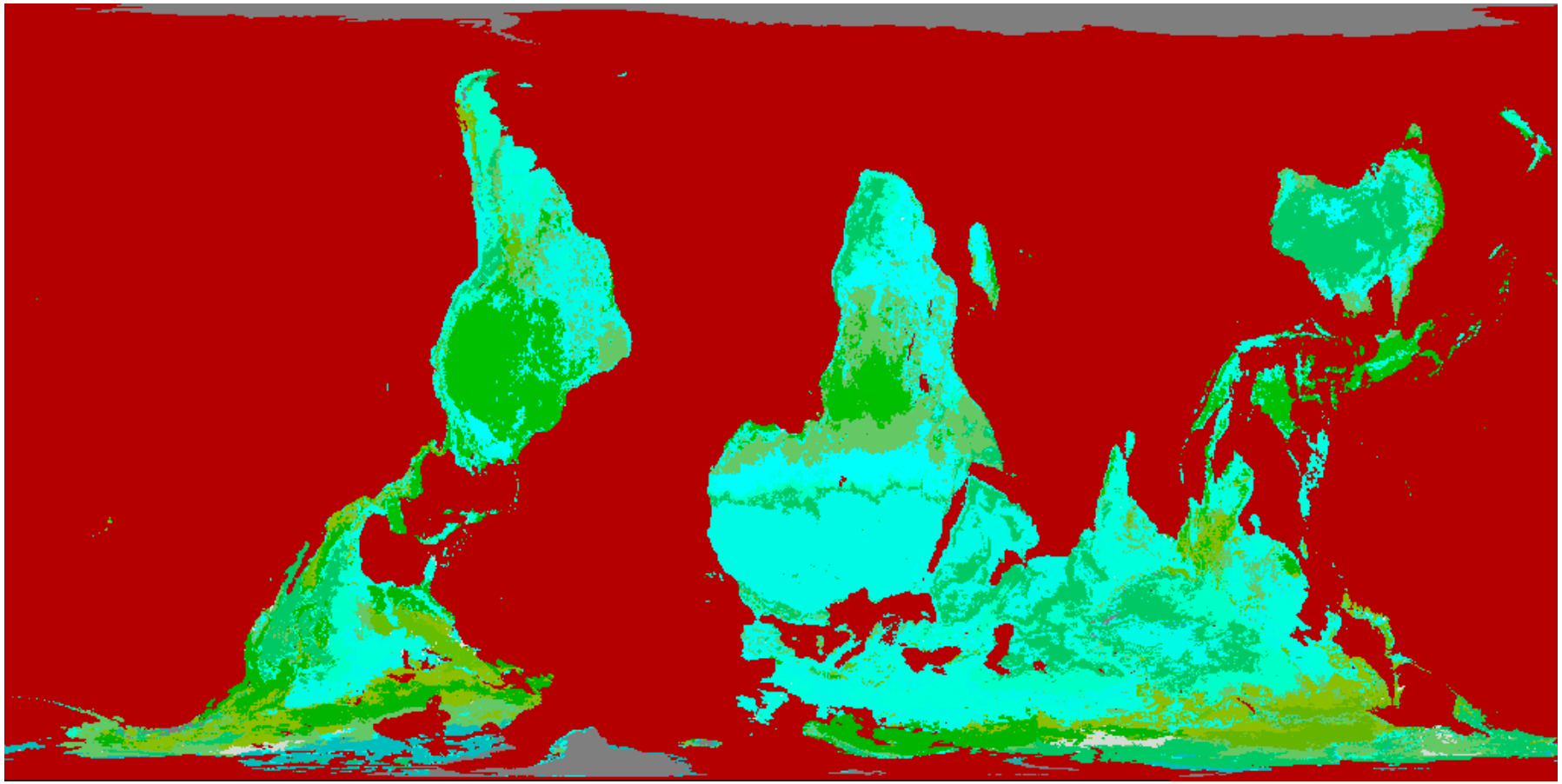
Algorithm:

1. Count number of array elements whose token value = 3, index number of Deciduous Broadleaf Forest
2. Multiply total number of elements by area of a single pixel

Example Visualization - Case 1



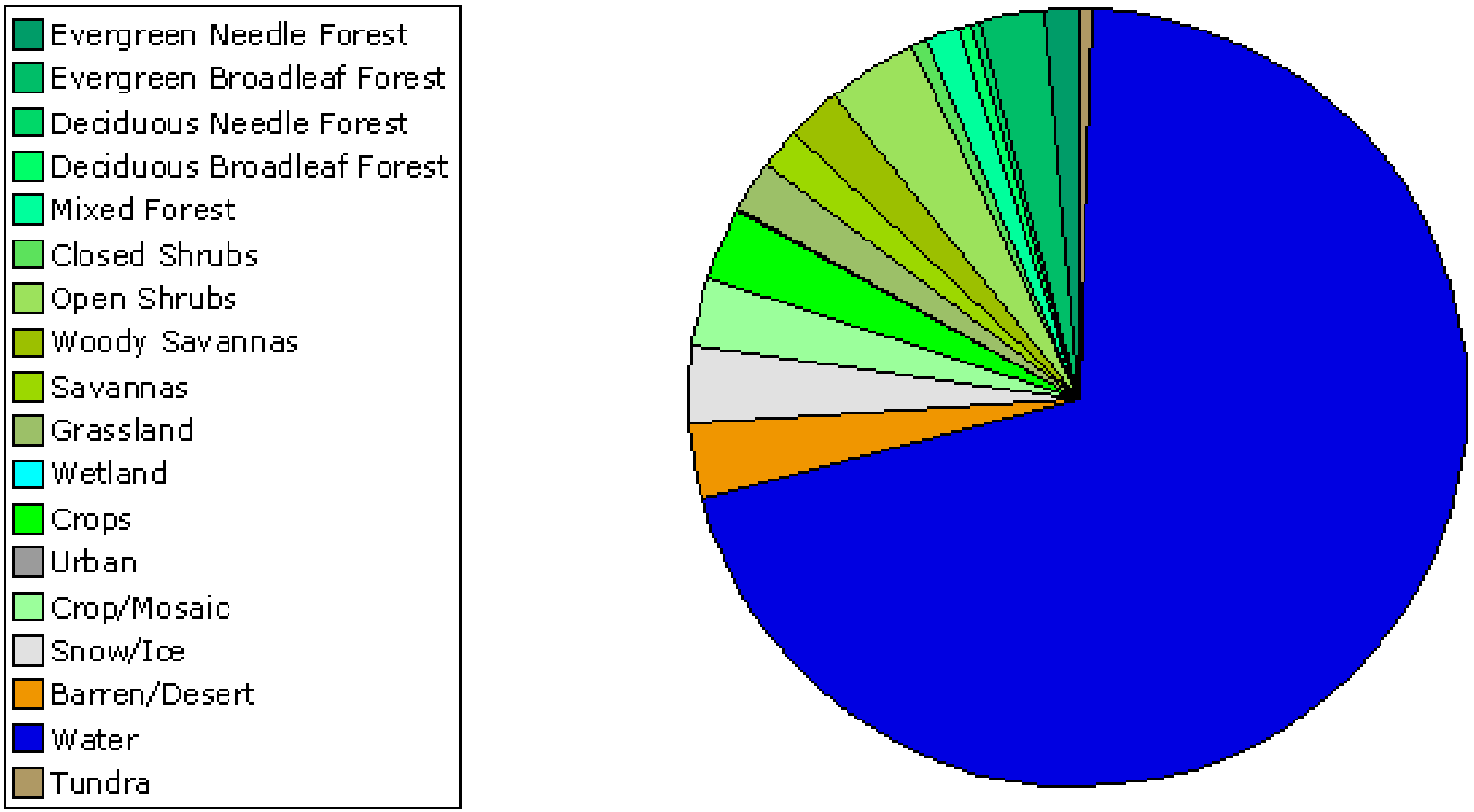
Example Visualization - Case 2



Notes on the Data in the Two Visualizations

- The array sizes are identical: 1024 in x, 512 in y
- The one-byte values that categorize the IGBP vegetation types use the same indexing scheme
- The palette in the first visualization is different from the palette in the second visualization.
- The order of byte values in the second visualization permutes the order of the byte values in the second image: **The cryptographic digest of the file with first visualization is not equal to the cryptographic digest of the file with the second visualization.**

Visualized Answer to Problem



Discussion of Permuted Numerical Token Order

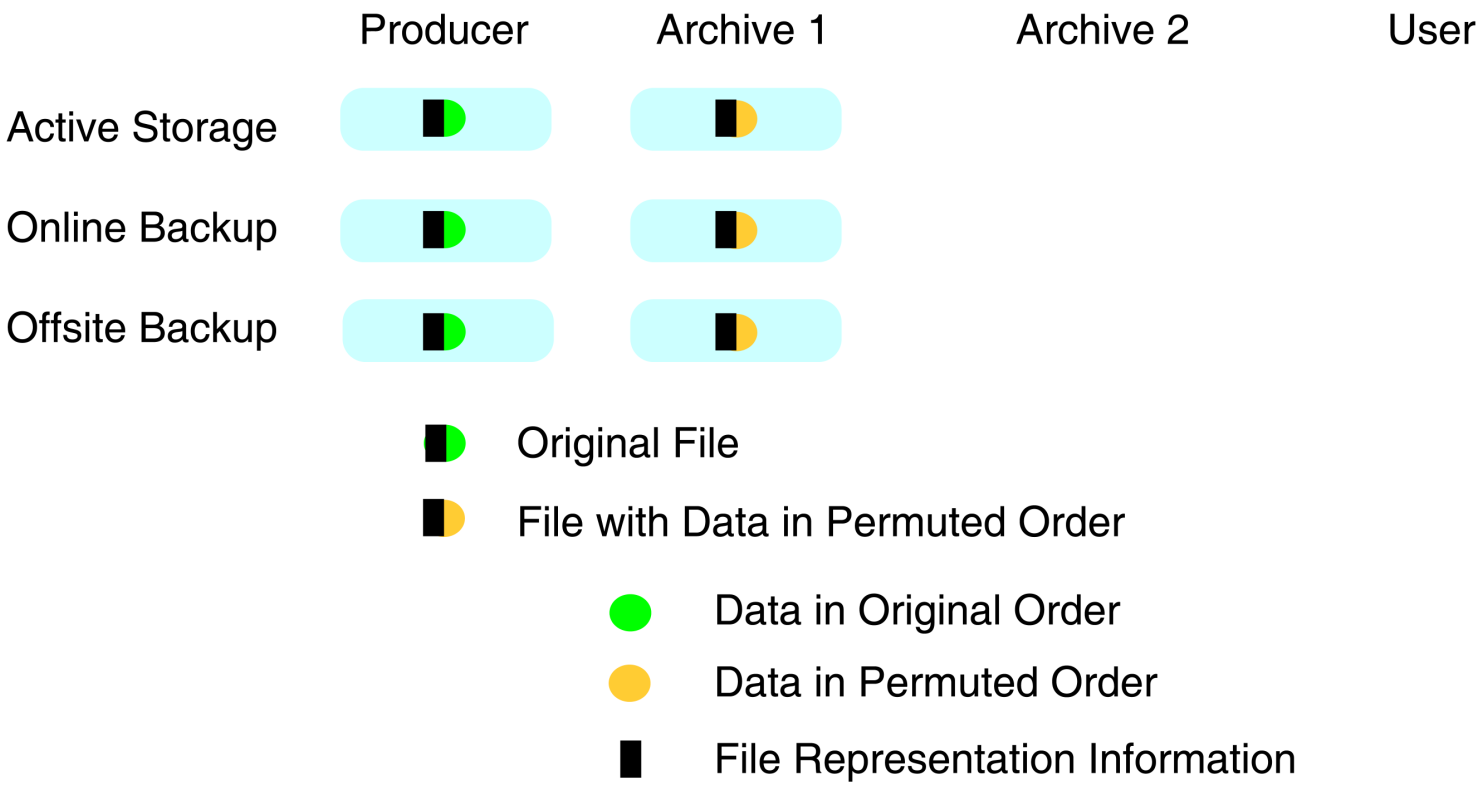
- Decide if the arrays in the two bit-mapped (.bmp) files are scientifically equivalent for using the algorithm to compute the area of the Earth covered by Deciduous Broadleaf Forest.
- Provide a chain of reasoning to justify your decision
- Create a chart of accounts for holding these files in an inventory
  - Decide if the two files belong in the same account
  - Decide how the chart of accounts should show the scientific equivalence (or lack of equivalence)
- Provide a chain of reasoning to justify your decision on the chart of accounts organization

Archives With Digital Files That are Scientifically Equivalent

- Two (or more) archives may have digital data files that have scientifically equivalent data with different numerical value formats and permuted order of equivalent values
- Users can test for equivalence by making sure
  - Each file in a pair under test has the same number of values.
  - Each pair of values in a loop through the data values has essentially the same numerical value and that the data comes from the same observation and chain of algorithms
- In the figure that follows, the data array in a green file is scientifically equivalent to the data array in an orange file.

Test Case: Archives With Scientifically Equivalent Data In Permuted Order

Case 2: Scientifically Equivalent Data



Discussion of Archive Policy Regarding Data Collections With Scientifically Equivalent Data

Key Questions:

1. *How should archives handle objects that differ in the order of numerical values, but that have scientifically equivalent data as far as users are concerned*?
   - Treat the different objects as being different.
   - Treat the different objects as being scientifically equivalent.
2. *If two objects have scientifically equivalent numerical values, is one more "authentic", and if so why*?
   - They are equally authentic.
   - They are not equally authentic. Authenticity could be independently verified (or audited) some algorithm and custodianship records.

Summary

- An archive can categorize stable objects with a taxonomic procedure.
- An archive can use standard bookkeeping to keep accounting for its object inventory.
- Considering scientific equivalence of data in objects may be more important to users than trying to emphasize uniqueness and authenticity
- Work To Be Done
  - Complete mathematical taxonomy algorithms to categorize objects with independently verifiable tests.
  - Develop quantitative techniques for judging how techniques for identifying scientifically equivalent digital files of Earth science data improve user work efficiency.