ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Data Preservation

NSF

INSTITUTE of Museum and Library SERVICES

V3: The preservation⬄integration parallels

Preservation goals (again)

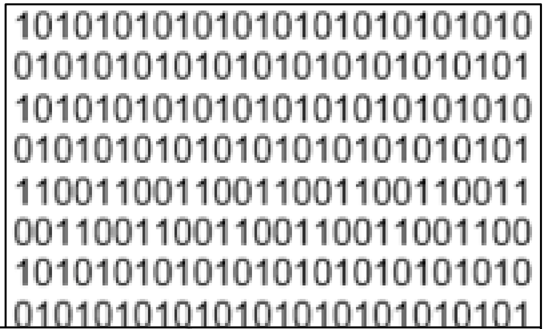
Preservation challenges appear isomorphic to *integration* challenges;
tracing the parallels helps us to better understand
both the challenges and the required interventions.

Preservation goals

Many organizations and standards
have proposed classifications of preservation goals.

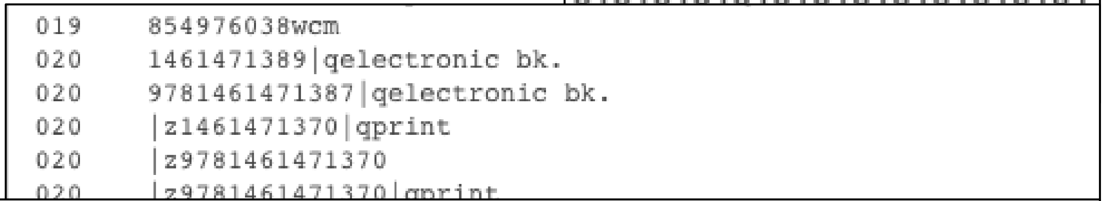
An early statement

Viable
(can be read from media)



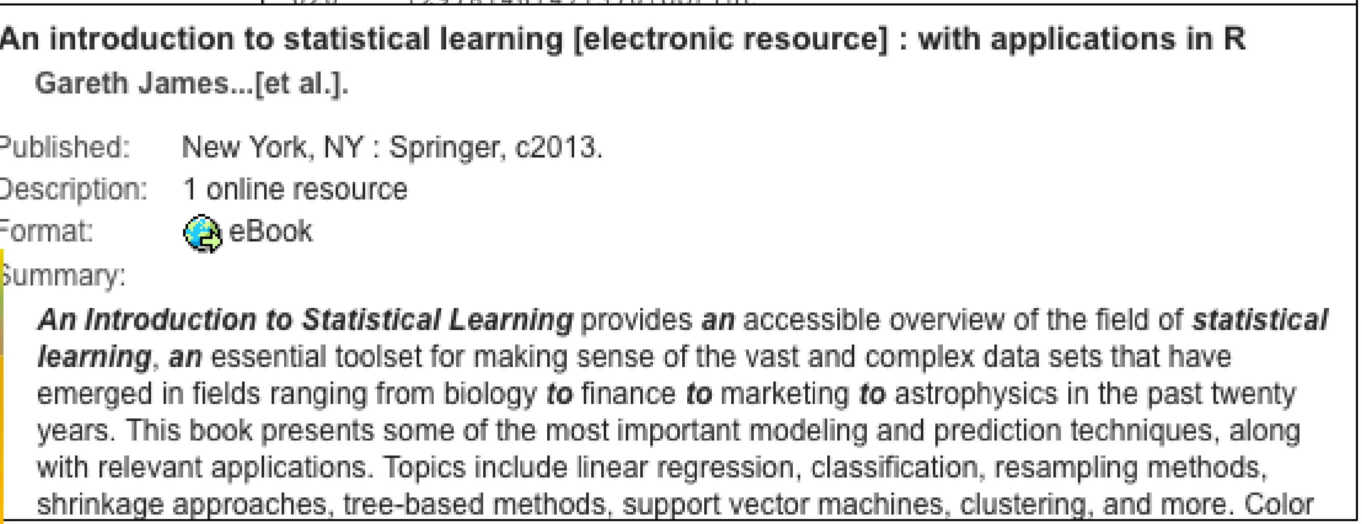
```
1010101010101010101010101010
0101010101010101010101010101
1010101010101010101010101010
0101010101010101010101010101
1100110011001100110011001100
0011001100110011001100110011
1010101010101010101010101010
0101010101010101010101010101
```

Renderable
(viewable and processable)



```
019     854976038wcm
020     1461471389|qelectronic bk.
020     9781461471387|qelectronic bk.
020     |z1461471370|qprint
020     |z9781461471370
020     |z9781461471370|qprint
```

Understandable
(interpretable)

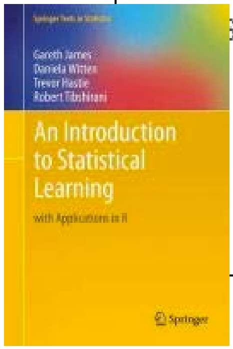


An introduction to statistical learning [electronic resource] : with applications in R
Gareth James...[et al.].

Published:	New York, NY : Springer, c2013.
Description:	1 online resource
Format:	eBook

Summary:

An Introduction to Statistical Learning provides ***an*** accessible overview of the field of ***statistical learning***, ***an*** essential toolset for making sense of the vast and complex data sets that have emerged in fields ranging from biology ***to*** finance ***to*** marketing ***to*** astrophysics in the past twenty years. This book presents some of the most important modeling and prediction techniques, along with relevant applications. Topics include linear regression, classification, resampling methods, shrinkage approaches, tree-based methods, support vector machines, clustering, and more. Color



See OCLC/RLG Working Group on Preservation Metadata. Image credits: UNC libraries.

And a longer one*

Viable	can be read (correctly) from media
Renderable	can be (correctly) viewed, processed, executed
Understandable	can be (correctly) understood
Authenticatable	can be (correctly) determined to be what it purports to be
Identifiable	can be (correctly) identified and re-identified

And more can be added of course: findable, conformant . . . etc..

Following our own definition of preservation we would emphasize that
each of these five not only must be achievable,
but the user must have *justified confidence* in the result.

*See: PREMIS https://www.loc.gov/standards/premis/

The data integration⇔preservation isomorphism

In many respects the challenges of data *preservation* are the same as the challenges of data *integration*,

only across *time* rather than *space*

.

Let’s take a look at data preservation through that lens,
we will be able to apply many things we’ve already discussed.

(As always, metadata is a key instrument for achieving our objectives)

The role of physical media in communication with the future

Physical objects are essential for *communication with the future* — because we live in a world of space, time, energy, matter, and causality,

And so attending to those physical objects is necessary to *ensure* that communication.

But the individual objects involved need not persist throughout the entire temporal interval:
Communication with the future is achieved via *a chain of overlapping physical objects*

Still, functionality of physical objects must be maintained during relevant intervals in the chain
.

Storage media, and associated hardware (drives, cables, etc.) must be protected against

kinetic damage, chemical decay, electromagnetic insult, etc.,

as well as theft and loss.

For physical media the ensuring of communication is accomplished with
policies and procedures,
the management of associated physical environments;
accurate, complete, and computer readable documentation

[Associated digital objects (APIs, embedded systems, operating systems) must also be preserved.]

Communicating encoding

The decoding of every level of encoding must be reliable.

From reading 1s and 0s off media, to mapping the bitstream to bytes then integers, then characters or other meaningful symbolic units.

[With UTF and Unicode some things are easier than they were, but nothing here can be taken for granted.]

As always we ensure the communication of encoding like we ensure all communication:

through documentation, particularly computer-processable metadata that uses standard vocabularies and a standard serialization syntax.

Usually the most important role of this metadata is to indicate the various data standards being used at different levels of encoding, or to indicate extensions, restrictions, subsetting, or other modifications of standards.

Communicating the syntax

A schema that identifies the structure of the data statements and documents the controlled vocabulary of attributes, identifiers, and values is essential.

> This schema should be computer processable so it can be used to validate the structure of the data, to supply data types, and to configure processing tools and applications.
>
> Here again we ensure communication through documentation, particularly computer-processable metadata that uses standard vocabularies and a standard serialization syntax.
>
>> And here too the most important role of this metadata is often to indicate the various data standards being used.

Communicating the propositions

This is what it is all about: *Ensuring the reliable communication of what is being asserted*

All the prior attention given to media, encoding, and syntax helps. But there is more to do.

The constituents of propositions are *relationships* and *entities* indicated in the controlled vocabulary for the top level logical syntax. These must be identified, explained, documented.

> For this a formal model or ontology is useful.
>
> The schema for this ontology should be computer processable so it can be used for validation, data integration, and management of transformation to alternative syntaxes.
>
> Again we are ensuring communication through documentation, particularly computer-processable metadata with standard vocabularies and standard serialization syntax.
>
> > And here too the most important role of this metadata is often to indicate the various data standards being used (here standard ontologies, perhaps modified).
>
> However at this level natural language prose will also be essential:
>
> > *temperature*? *location*? *county*? *race*? species?
> >
> > > . . . such things cannot be explained by formal language alone.

Turtles?

[No, not the RDF serialization language]

As with data integration documentation
there is also the regress problem for data preservation as well.

Metadata for data is itself data,
and must also be preserved . . .
and that will require, among other things, metadata for metadata

Where to stop is a practical matter.

And the foundation will be natural language prose.

This is another reason why existing shared standards are so important.

They already back up mathematical formalities with natural language,
and their existence and use creates communities of shared understanding.