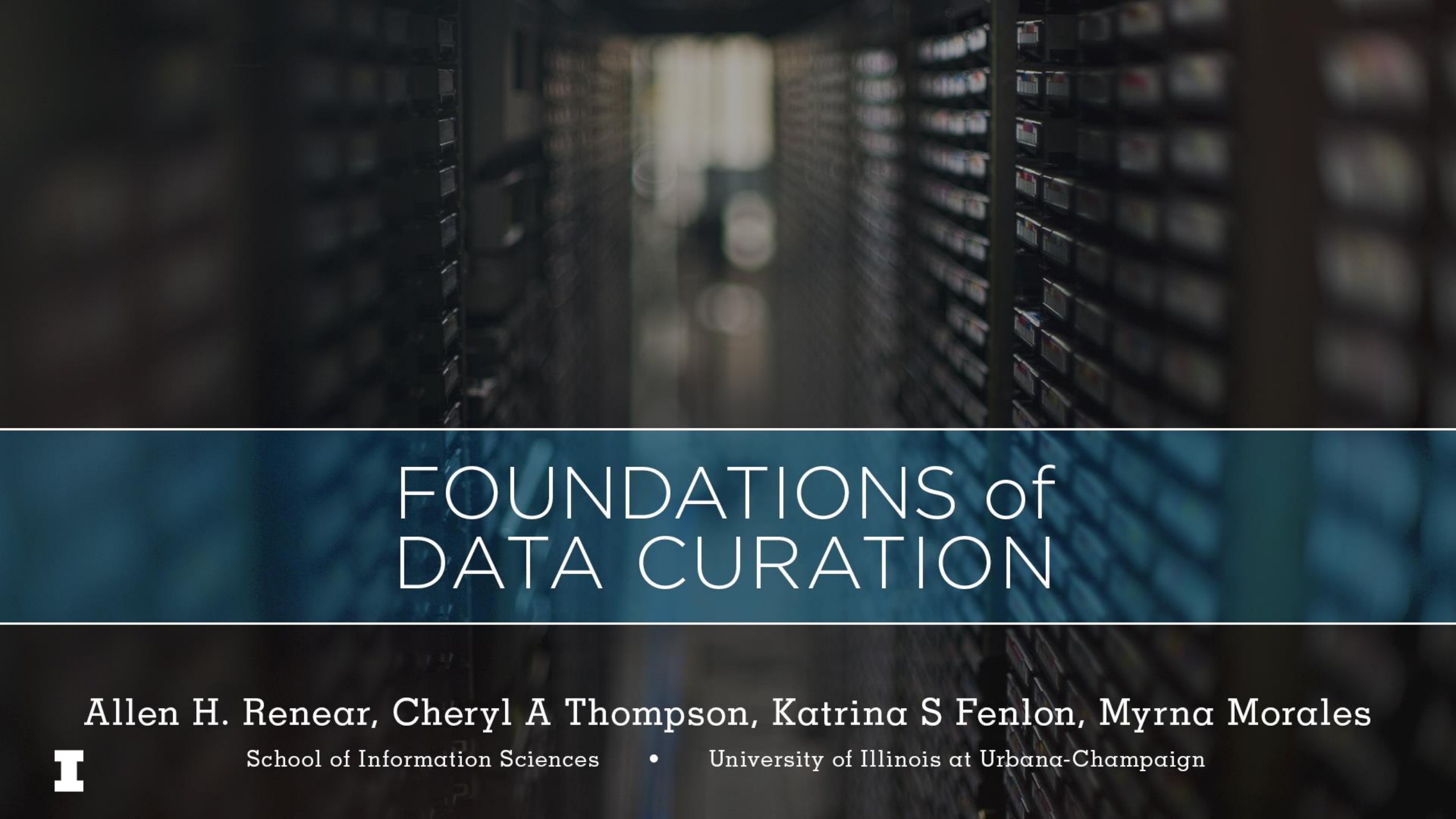

FOUNDATIONS of DATA CURATION

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales

School of Information Sciences • University of Illinois at Urbana-Champaign


DATA
CONCEPTS


2
THE IDENTITY PROBLEM

The Identity Problem

Identity problems in data curation

Identity problems and representation levels

Identity problems in data curation

Archiving:	Is this dataset already in the archive?
Preservation:	Was the information preserved in the new file format?
Security:	Has this dataset been tampered with?
Authentication:	Is this the data we think it is?
Reproducibility:	Does this XML file have the same information as that JSON file?
Provenance:	Were these datasets derived from the same data?
Conversions:	Does the converted file have the same data as the original?

and on and on. . .

> "… there are an unknown number of transformations that are invariant in the sense of preserving the scientific meaning . . . different scientific communities use different tools that require different representations.

Ruth Duerr, National Snow and Ice Data Center

Data Conservancy wiki, December 2010

Same, different, same, different. (But same/different what?)

Consider conversions*

[DC to ISO-Bib, TEI P2 to TEI P3, mzData to mzML, JSON to XML . . . and on . .]

Some conversions are simple format changes
some involve a change in model type
some have schema integration challenges
[and some have profound heterogeneity problems]

In a successful conversion we'd probably say
"the data is the same . . . it is only in a different format, encoding, etc.

So in a successful conversion something changes;

and something remains the same.

But what exactly changes? And what remains the same?

*Transformations, transcodings, etc.

Identity problems

Two biologists, Jill and John, used the same data

What does that mean?
And how can we tell?

Compare:

Two biologists, Jill and John, used the same statistician.

Identity and representation levels

Consider two files with the same data

but relational tables in one case

and RDF triples in another

Same data, different representations

Identity and representation levels

Consider two files with

… same data and the same RDF triples,

but an XML serialization in one case,

and an N3 serialization in the other

Identity and representation levels

Consider two files

with the same data, same RDF triples, same N3 serialization,

but an ASCII character encoding in one case
vs an EBCDIC encoding in another

Identity and representation levels

How many of these levels are there ?

How do we name, define, and manage them?

How can they be identified and re-identified?

Identity conditions

So, underlying many data curation issues is the problem of identity.

Is x is the same [data, document, text, image . . .] as y?

The conceptual question: What do we mean?

The operational question: How do we tell?

FOUNDATIONS OF DATA CURATION (IS531)

Allen H. Renear, Cheryl A Thompson, Katrina S Fenlon, Myrna Morales
School of Information Sciences
University of Illinois at Urbana-Champaign

Includes material adapted from work by Carole Palmer, Melissa Cragin, David Dubin, Karen Wickett, Bertram Ludaescher, Ruth Duerr and Simone Sacchi.

Comments and corrections to: renear@illinois.edu.