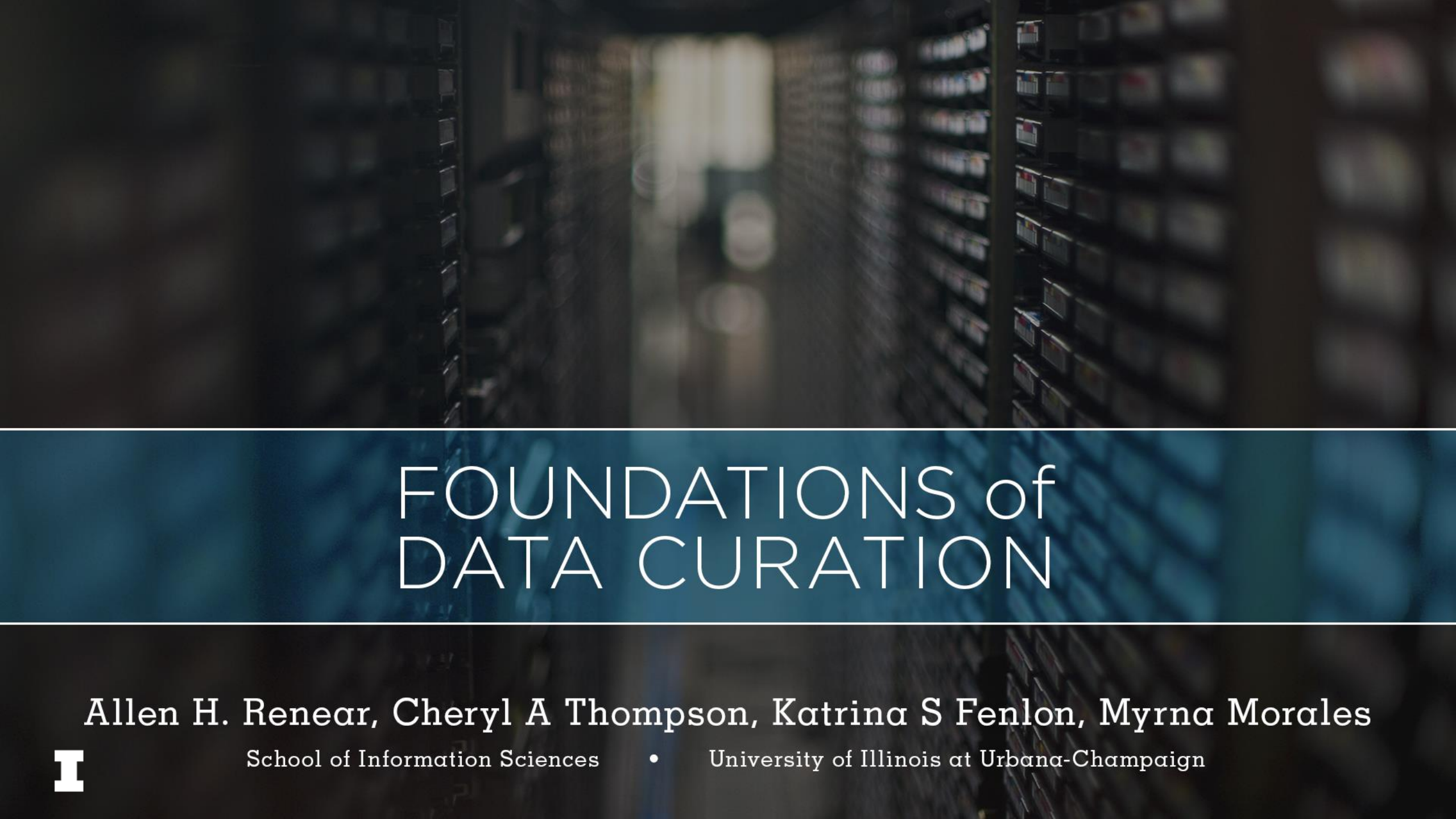


FOUNDATIONS of DATA CURATION

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

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


DATA
CONCEPTS

1

WHAT IS DATA?
A FIRST ATTEMPT

What Is Data? A First Attempt.

Some dictionary definitions.

These are useful, but not formal enough, and conceal problems.

An empirical approach to the question.

This yields some additional insights,

but no rigor, and plenty of variation, and inconsistencies.

Scientists still insist formal definitions are needed.

So we aren’t going to give up.

The Question

What is data?

???

No, we want to know *what* is data, not *who* is Data.

What we want for an answer

What we are after here is a *formal definition* of data.

And one that is part of a *conceptual model (or ontology) of data concepts*.

This is not just of theoretical interest.

> A conceptual model for data concepts would provide a rigorous formal foundation for the design of systems supporting all aspects of data curation.

Data, some lexicographical definitions

information, especially information organized for analysis
-- American Heritage Dictionary

factual information (such as measurements and statistics) used as a basis for reasoning, discussion, or calculation.
-- Merriam Webster Dictionary

a collection of facts from which conclusions may be drawn
-- Wordnet (Princeton)

a collection of observations . . .
-- [common]

a collection of organized information, usually the result of experience, observation or experiment, … may consist of numbers, words, or images, particularly as measurements or observations …

-- State of Maryland, Department of Information Technology

Are we done?

No, these definitions useful, but they are much too casual for rigorous modeling.

What we want is a conceptual model, or ontology, for data concepts

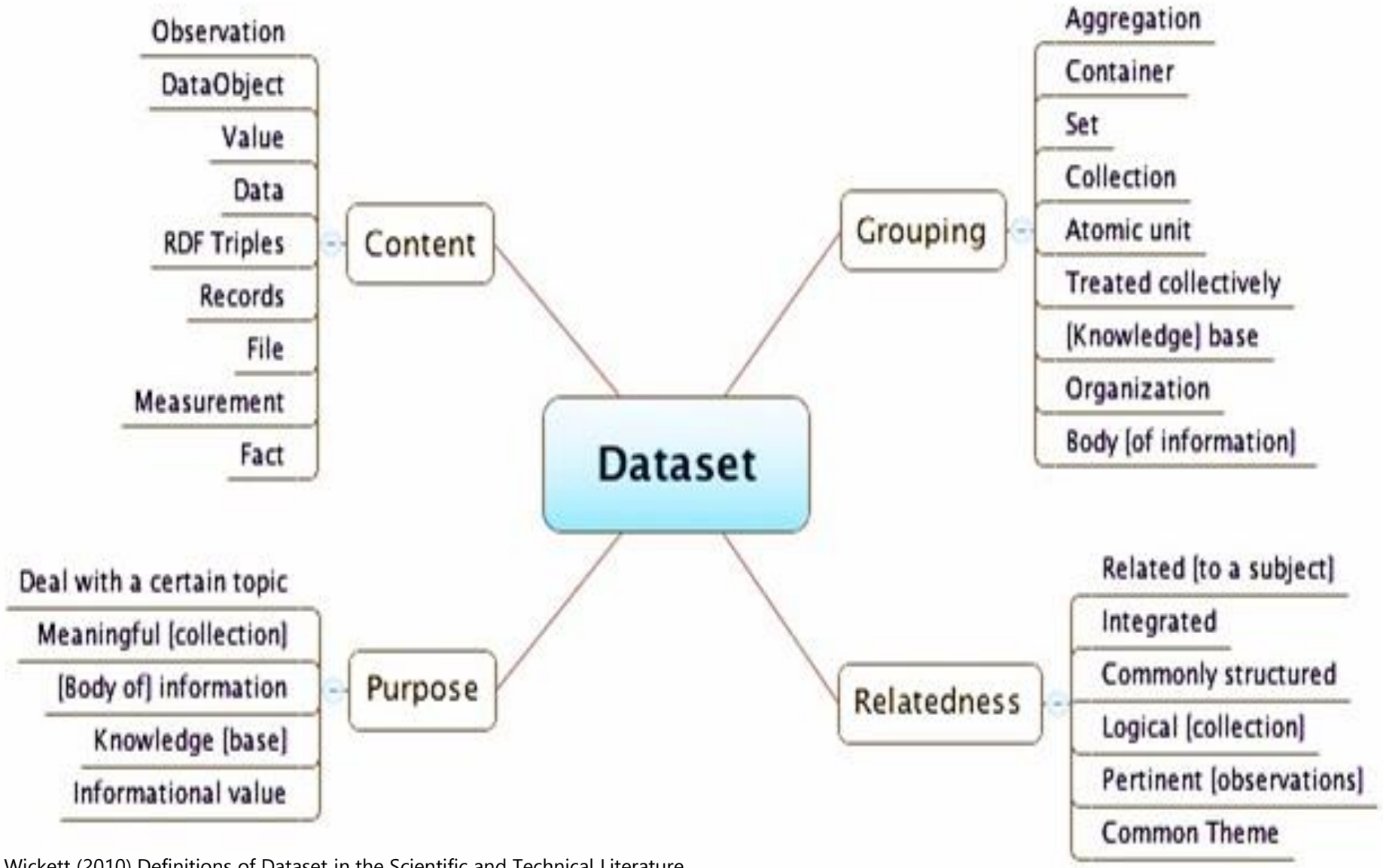
An empirical approach

Maybe we should ask a scientist

— they should have an answer, right?

[Well, actually they don’t; but let’s ask anyway. . .]

How scientists define datasets



Renear, Sacchi, Wickett (2010) Definitions of Dataset in the Scientific and Technical Literature

Fits our dictionary definitions well enough

information, especially information organized for analysis
-- American Heritage Dictionary

factual information (such as measurements and statistics) used as a basis for reasoning, discussion, or calculation.
-- Merriam Webster Dictionary

a collection of facts from which conclusions may be drawn
-- Wordnet (Princeton)

a collection of observations . . .
-- [common]

a collection of organized information, usually the result of experience, observation or experiment, … may consist of numbers, words, or images, particularly as measurements or observations …
-- State of Maryland, Department of Information Technology

Content
Grouping
Relatedness
Purpose

But is there really enough agreement for modeling?

There is too much variation here:

Purpose:	analysis, evidence, explaining, being explained
Relatedness:	same subject, same attributes, same syntax
Grouping:	set, aggregation, collection

And most challenging for modeling:

Content:	observations, values, facts, numbers, records, expressions, triples, tuples, information . . .

While we see patterns, colloquial definitions are still, from a modelling point of view, too informal, too varied, too inconsistent.

Well, maybe scientists just don't care much about definitions?

Cries from the heart. . .

"There is ambiguity in what type of object a dataset is;
with different groups of users applying different connotations

*There needs to be an explicit statement of what
the intended preservation of a dataset will imply."*

Sam Pepler, National Center for Atmospheric Science, UK

"While there has been substantial work in the IT community regarding metadata and file identifier schemas, there appears to be relatively little work on the organization of the file collections .

One symptom . . . appears in nomenclature describing collections: the terms `Data Product,' `Data Set,' and `Version' are overlaid with multiple meanings between communities."

Bruce Barkstrom, NASA

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.