Data: Types & Roles

INFX 551
Foundation of Data Curation

Agenda

- Data - Type vs Role distinctions; Structured vs Unstructured Data
- Curation - The premise of data curation; Workflows for Curation

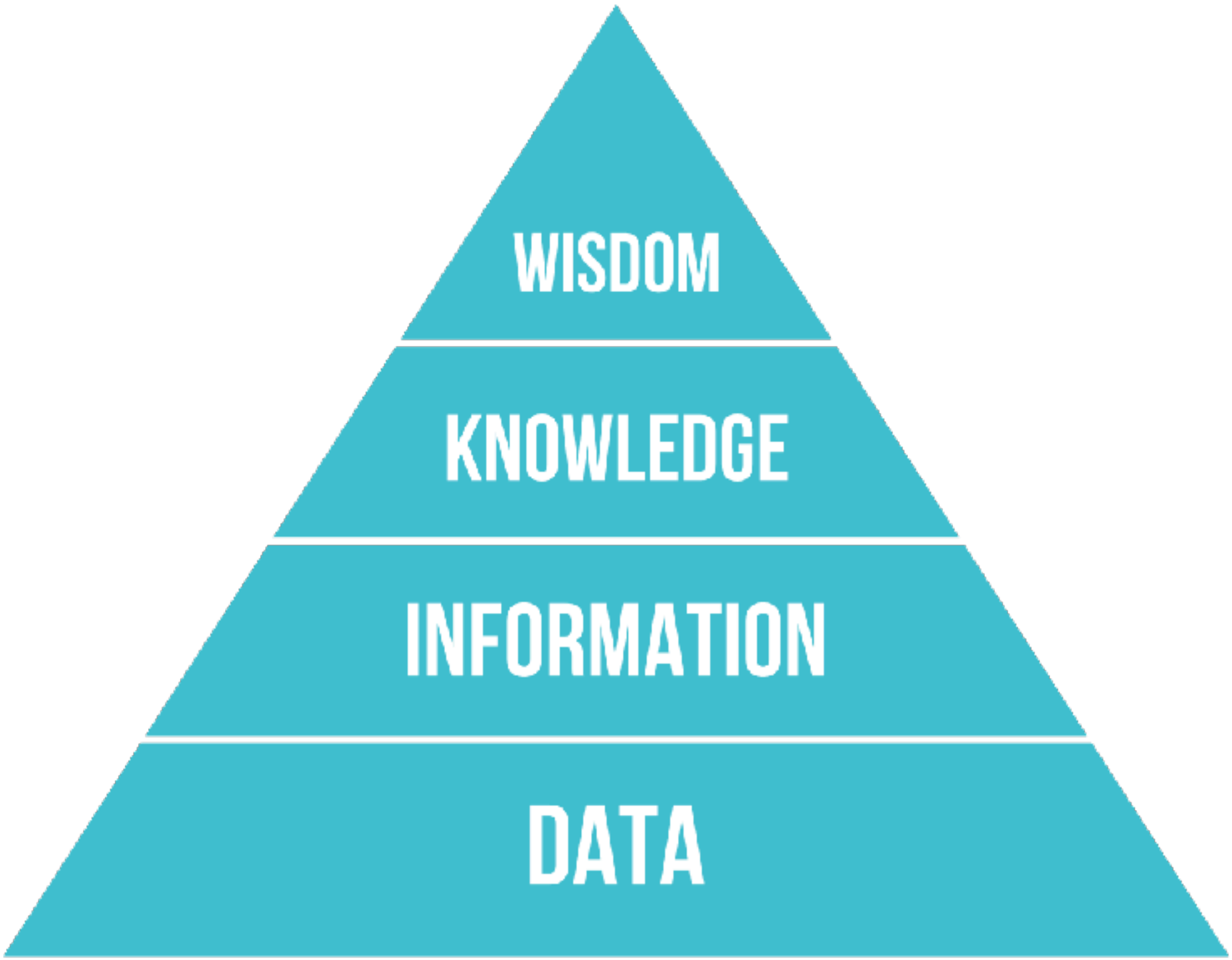
Course Outline

This course consists of three modules.
Each module consists of three topics.

Data	Data Systems	Policy, Privacy, & Ethics
Types & Roles	Repositories	Policy
Documentation	Preservation	Privacy
Standardization	Cost Models	Ethics

For this course, let's assume that….

Data Curation is the active and ongoing management of **data** throughout a lifecycle of use, including reuse unanticipated contexts.


WISDOM
KNOWLEDGE
INFORMATION
DATA

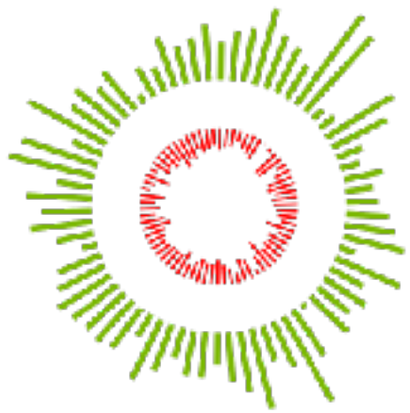
Research Data

"The data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise research observations, findings or outcomes, including primary materials and analysed data."

AUSTRALIAN NATIONAL DATA SERVICE

Open Data

"Open data is data that can be freely used, shared and built-on by anyone, anywhere, for any purpose."



For this course, let's assume that….

Data are various **types** of information objects playing the **role** of evidence.

Evidence of…
Infection (Public Health)
Patterns of Consumer Behavior (Business)
Weather (Atmosphere)

Types vs Roles

Structured vs Unstructured

Type vs Role
distinctions

Type:
Donald Trump is a person.

Role:
Donald Trump is POTUS

Types of Data
(by file format)

XML

Databases

Flat Files

EDI

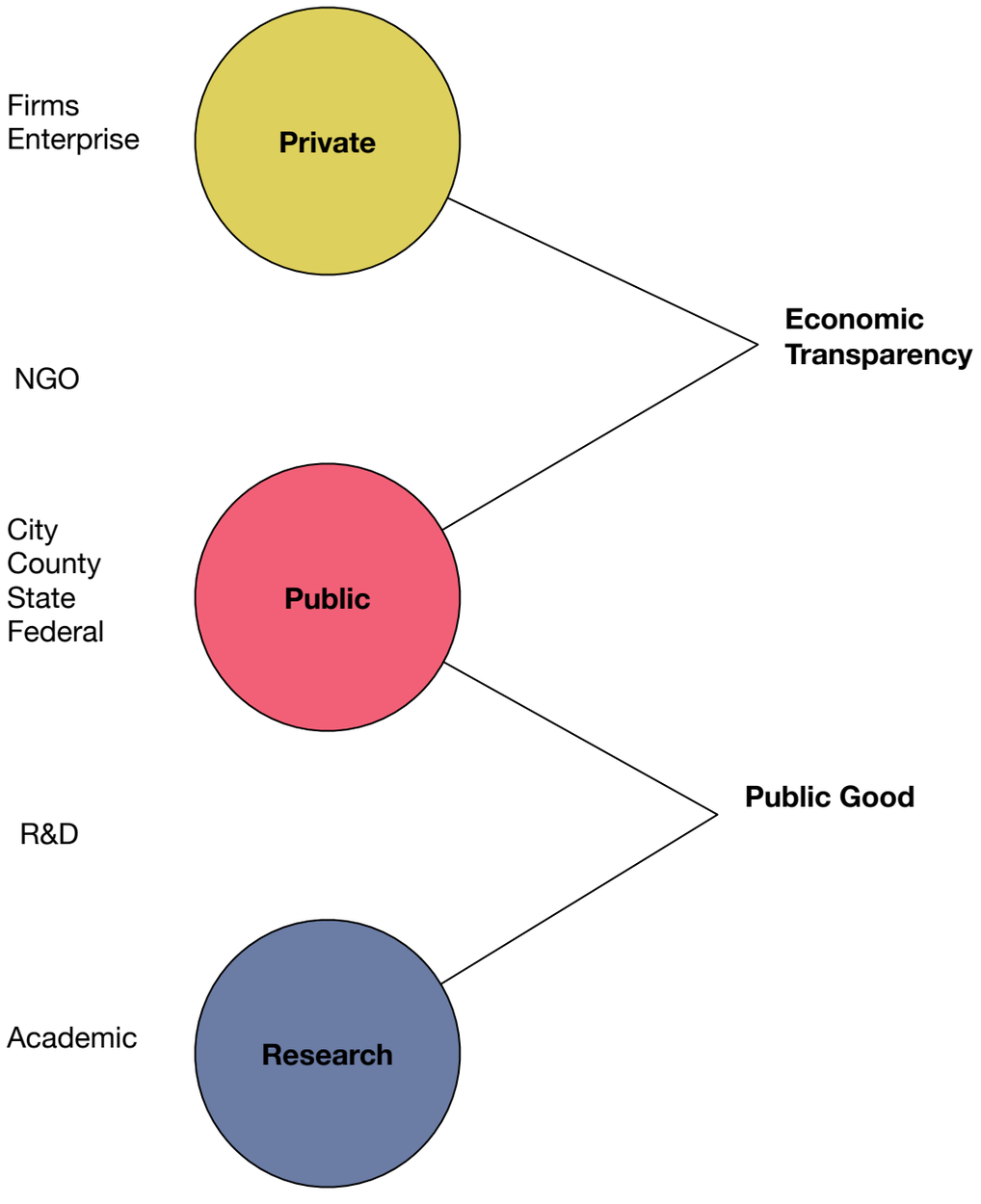
Excel

XBRL

JSON

Web Services

Types of Data
(by sector)



Roles

OPEN ACCESS Freely available online

PLOS ONE

The Role of Gender in Scholarly Authorship

Jevin D. West[1]*, Jennifer Jacquet[2], Molly M. King[3], Shelley J. Correll[4], Carl T. Bergstrom[1,4]

1 Department of Biology, University of Washington, Seattle, Washington, United States of America, 2 Environmental Studies, New York University, New York, New York, United States of America, 3 Department of Sociology, Stanford University, Stanford, California, United States of America, 4 Santa Fe Institute, Santa Fe, New Mexico, United States of America

Abstract

Gender disparities appear to be decreasing in academia according to a number of metrics, such as grant funding, hiring, acceptance at scholarly journals, and productivity, and it might be tempting to think that gender inequity will soon be a problem of the past. However, a large-scale analysis based on over eight million papers across the natural sciences, social sciences, and humanities reveals a number of understated and persistent ways in which gender inequities remain. For instance, even where raw publication counts seem to be equal between genders, close inspection reveals that, in certain fields, men predominate in the prestigious first and last author positions. Moreover, women are significantly underrepresented as authors of single-authored papers. Academics should be aware of the subtle ways that gender disparities can occur in scholarly authorship.

Citation: West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT (2013) The Role of Gender in Scholarly Authorship. PLoS ONE 8(7): e66212. doi:10.1371/journal.pone.0066212

Editor: Lilach Hadany, Tel Aviv University, Israel

Received January 25, 2013; **Accepted** May 7, 2013; **Published** July 22, 2013

Funding: This work was supported in part by NSF grant SBE-0915005 to CTB, NSF Graduate Research Fellowship grant DGE-1147470 to MMK, and a generous gift from JSTOR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jevinw@u.washington.edu

Introduction

Gender inequities and gender biases persist in higher education. After decades of high female enrollment in most PhD fields, women represent one-quarter of full professors and earn on average 80% of the salary of men in comparable positions [1]. A recent report [2] surveyed 1000 faculty across six science and engineering disciplines and found men publish significantly more in chemistry and mathematics, while women publish more in electrical engineering (there were no significant differences found in biology, civil engineering, and physics). A recent experiment tested the role of gender in hiring by asking 127 science faculty to evaluate potential lab manager applications and found faculty gave identical applications higher scores if the applicant had a male name [3]. Another recent analysis of commissioned articles in two prestigious journals published in 2010 and 2011 showed that women scientists are underrepresented; for instance, women wrote just 3.8% of earth and environmental sciences articles for Nature News & Views, although they represent 20% of the scientists in this discipline [4]. With the use of alphabetical authorship listings declining over time [5], and given the complexity of evaluating intellectual contributions [6] in increasingly collaborative efforts, understanding patterns of authorship order becomes increasingly important.

Here we use the JSTOR corpus—a body of academic papers from a range of scholarly disciplines spanning five centuries—to examine trends in the gender composition of academic authorship through time. We pay particular attention to authorship order, given that first and sometimes last author publications are at least as important as raw publication counts for hiring, promotion, and tenure, particularly in scientific fields [7]. Studies of authorship in the medical literature reveal, for instance, that women have been historically underrepresented in the prestige positions of first and last author, and that while discrepancies have recently declined in the first author position, women remain underrepresented as last authors [8,9,10,11]. To view authorship patterns in their disciplinarity context, we use a network-based community detection approach to categorize hierarchically each paper in our study corpus. This yields a hierarchical classification of all papers in our study and allows us to study and compare patterns of gender representation in individual fields of any size and scale.

Methods

The JSTOR corpus

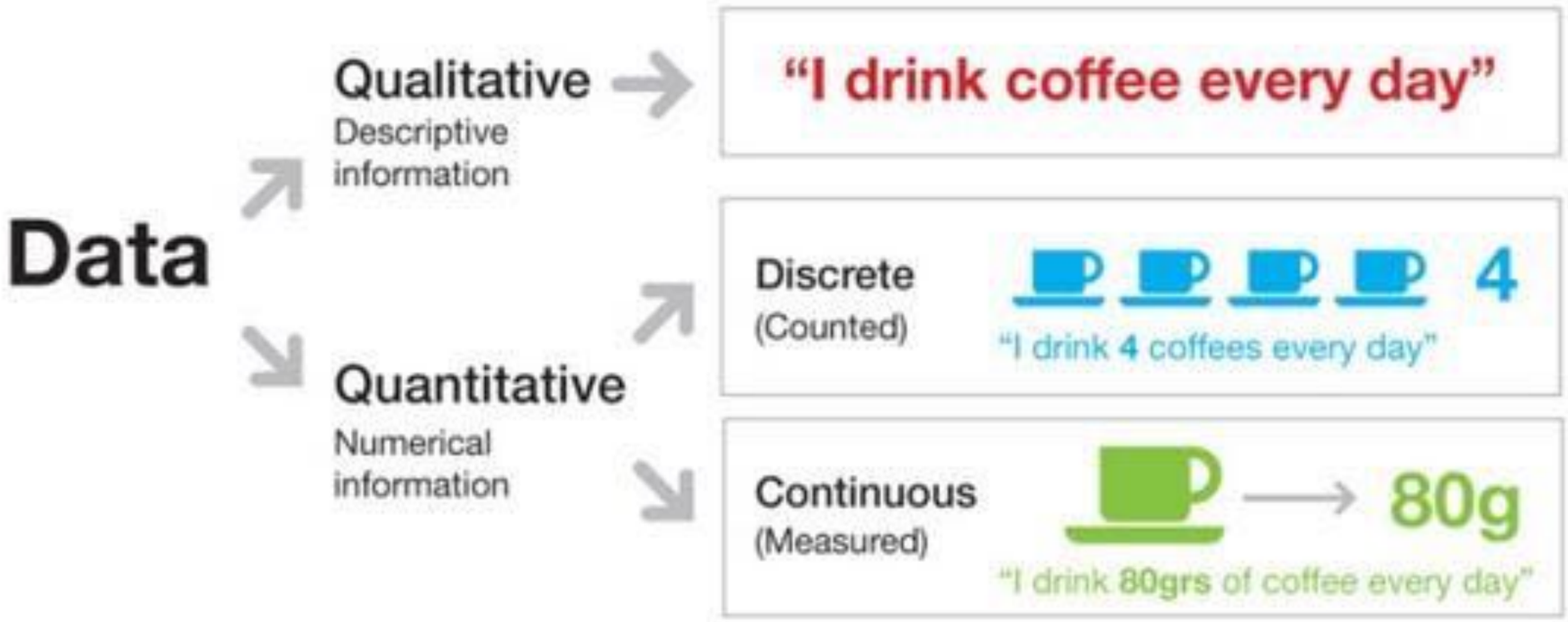
The JSTOR corpus (http://www.jstor.org) is a digital archiver of published scholarly research that spans the sciences and humanities from 1545 to the present day. At the time of this analysis, the JSTOR corpus comprised 8.3 million documents ranging from 1545 until early 2011, including 4.2 million research articles. Approximately 1.8 million of these documents (91% of which are research articles) cite or are cited by other documents in the JSTOR corpus and thus are amenable to network analysis. We call this group the "JSTOR network dataset". Moreover 94% of these 1.8 million articles are part of a single giant component of the citation network, such that any of these articles can be reached from any other by following citation trails forwards and backwards. We restrict our analysis to the JSTOR network dataset because this is the portion of the JSTOR corpus that we can hierarchically categorize using citation information. For a list of the main fields available in JSTOR dataset, see Table 1. The gender composition of the identified authors in the network dataset (21.9% female) is close to that of the identified authors in the entire corpus (20.8% percent).

GBIF
Global Biodiversity Information Facility

Roles

```
<text xmlns="http://www.tei-c.org/ns/1.0" xml:id="d1">
   <body xml:id="d2">
      <div1 type="book" xml:id="d3">
         <head>Songs of Innocence</head>
         <pb n="4"/>
         <div2 type="poem" xml:id="d4">
            <head>Introduction</head>
            <lg type="stanza">
               <l>Piping down the valleys wild, </l>
               <l>Piping songs of pleasant glee, </l>
               <l>On a cloud I saw a child, </l>
               <l>And he laughing said to me: </l>
            </lg>
            <lg type="stanza">
               <l>"Pipe a song about a Lamb!" </l>
               <l>So I piped with merry chear. </l>
               <l>"Piper, pipe that song again;" </l>
               <l>So I piped, he wept to hear. </l>
            </lg>
            <lg type="stanza">
               <l>"Drop thy pipe, thy happy pipe; </l>
               <l>Sing thy songs of happy chear:" </l>
               <l>So I sung the same again, </l>
               <l>While he wept with joy to hear. </l>
            </lg>
            <lg type="stanza">
               <l>"Piper, sit thee down and write </l>
               <l>In a book, that all may read." </l>
               <l>So he vanis'd from my sight, </l>
               <l>And I pluck'd a hollow reed, </l>
            </lg>
            <lg type="stanza">
               <l>And I made a rural pen, </l>
               <l>And I stain'd the water clear, </l>
               <l>And I wrote my happy songs </l>
               <l>Every child may joy to hear. </l>
            </lg>
```

Roles


Data
Qualitative
Descriptive information
"I drink coffee every day"
Quantitative
Numerical information
Discrete
(Counted)
4
"I drink 4 coffees every day"
Continuous
(Measured)
80g
"I drink 80grs of coffee every day"

For this course, let's assume that….

Data are various **types** of information objects playing the **role** of evidence.

Evidence of…
Infection (Public Health)
Patterns of Consumer Behavior (Business)
Weather (Atmosphere)

Unstructured vs **Structured**

Unstructured

Early Lead

Boston Marathon 2017: Geoffrey Kirui tops American Galen Rupp and Edna Kiplagat cruises

By Cindy Boren April 17, 2017

Geoffrey Kirui, of Kenya, crosses the finish line. (Elise Amendola/Associated Press)

Geoffrey Kirui pulled away from Galen Rupp over the last few miles, winning the men's division of the 121st Boston Marathon on Monday and ending Rupp's bid to become the first American winner since 2014. In the women's division, Edna Kiplagat made the most of her first appearance in the Boston Marathon, separating from the pack at the 18-mile mark and cruising to a victory.

Kirui, a 24-year-old runner from Kenya, won in 2:09:37 on a warm day in Boston. Rupp, the Olympic bronze medalist in the event, finished second in his first big-city American marathon in 2:09:58 and Suguru Osako of Japan was third in 2:10:28. In his three marathons, Rupp has finished first (in the Olympic trials), third and second. "It lived up to and exceeded all my expectations," Rupp told NBC. "[The crowd] really lifted me those last three miles."

Structured

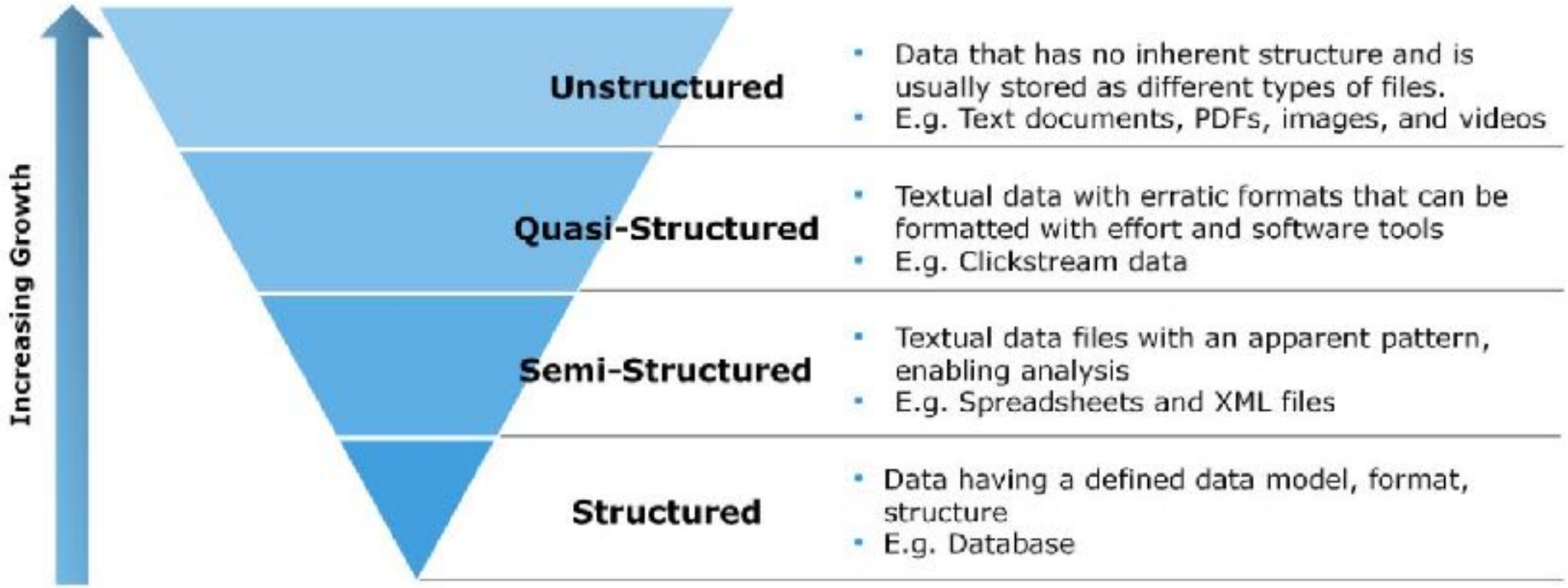
Name	Time	Nationality	Place
Geoffrey Kirui	2:09:37	Kenya	1
Galen Rupp	2:09:58	USA	2
Suguru Osako	2:10:28	Japan	3

Unstructured

Structured

?

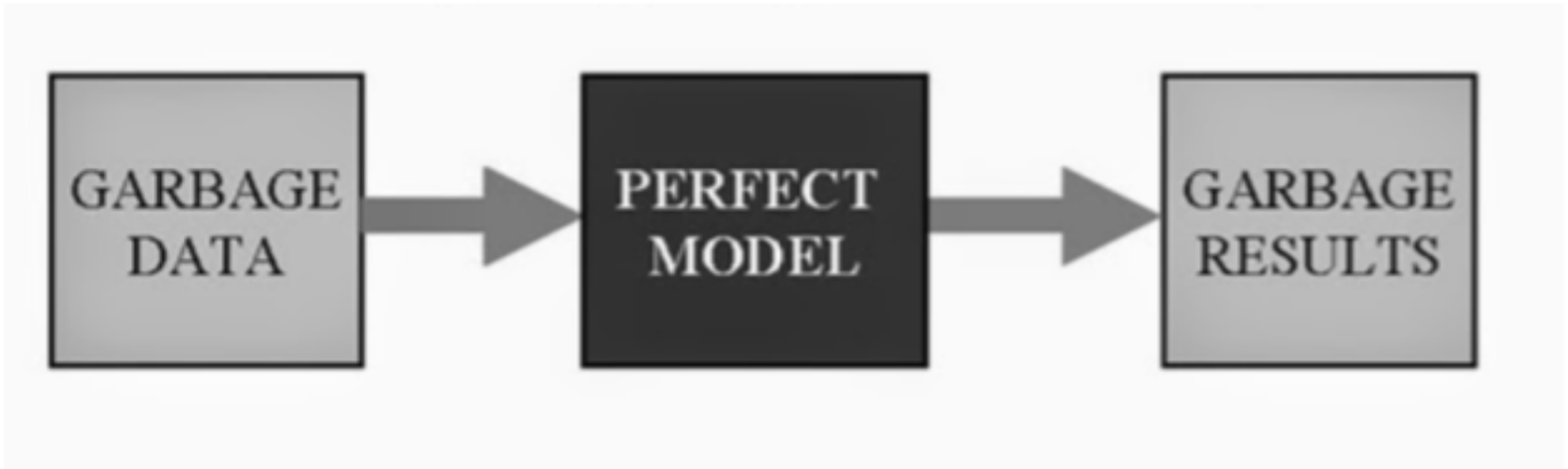
```
<text xmlns="http://www.tei-c.org/ns/1.0" xml:id="d1">
   <body xml:id="d2">
      <div1 type="book" xml:id="d3">
         <head>Songs of Innocence</head>
         <pb n="4"/>
         <div2 type="poem" xml:id="d4">
            <head>Introduction</head>
            <lg type="stanza">
               <l>Piping down the valleys wild, </l>
               <l>Piping songs of pleasant glee, </l>
               <l>On a cloud I saw a child, </l>
               <l>And he laughing said to me: </l>
            </lg>
            <lg type="stanza">
               <l>"Pipe a song about a Lamb!" </l>
               <l>So I piped with merry chear. </l>
               <l>"Piper, pipe that song again;" </l>
               <l>So I piped, he wept to hear. </l>
            </lg>
            <lg type="stanza">
               <l>"Drop thy pipe, thy happy pipe; </l>
               <l>Sing thy songs of happy chear:" </l>
               <l>So I sung the same again, </l>
               <l>While he wept with joy to hear. </l>
            </lg>
            <lg type="stanza">
               <l>"Piper, sit thee down and write </l>
               <l>In a book, that all may read." </l>
               <l>So he vanis'd from my sight, </l>
               <l>And I pluck'd a hollow reed, </l>
            </lg>
            <lg type="stanza">
               <l>And I made a rural pen, </l>
               <l>And I stain'd the water clear, </l>
               <l>And I wrote my happy songs </l>
               <l>Every child may joy to hear. </l>
            </lg>
```


Increasing Growth
Unstructured
Data that has no inherent structure and is usually stored as different types of files.
E.g. Text documents, PDFs, images, and videos
Quasi-Structured
Textual data with erratic formats that can be formatted with effort and software tools
E.g. Clickstream data
Semi-Structured
Textual data files with an apparent pattern, enabling analysis
E.g. Spreadsheets and XML files
Structured
Data having a defined data model, format, structure
E.g. Database

Curation

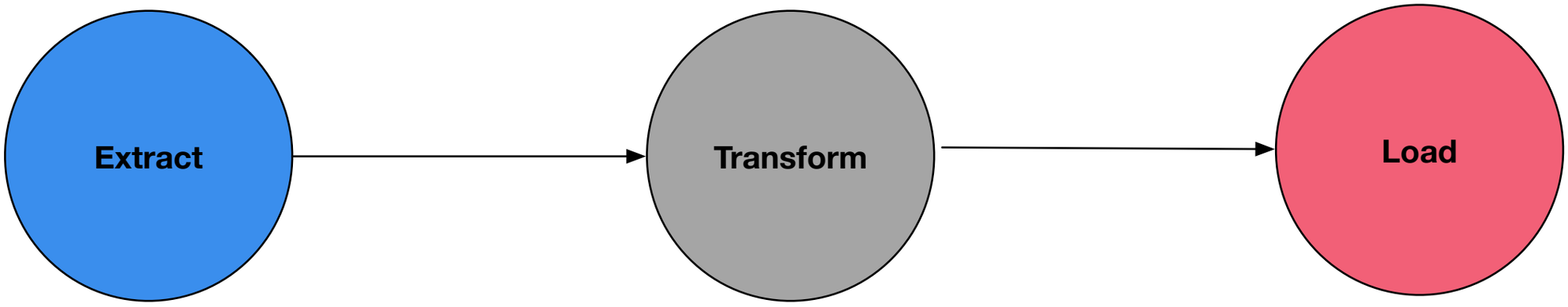
Data Curation

Old computer science saying "Garbage in = Garbage out"

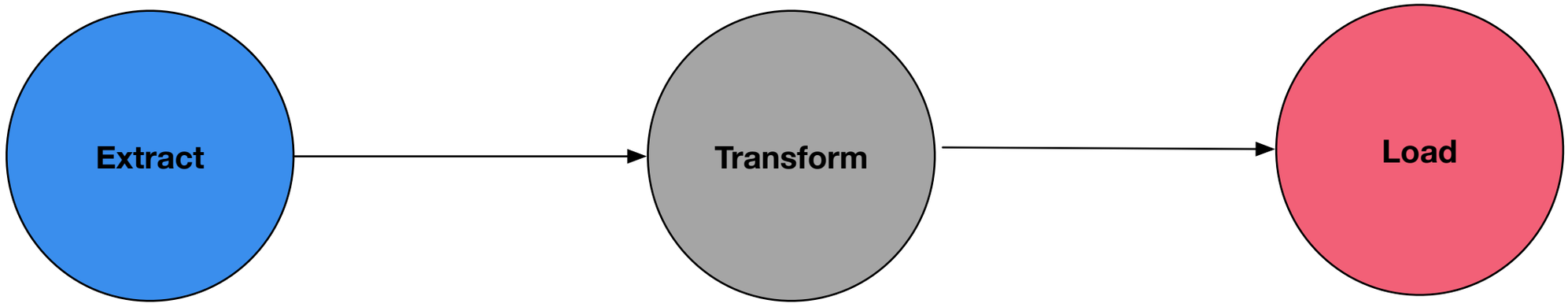


Data curation says "Quality in = Quality out"

ETL: Simplest form of Curation Workflow

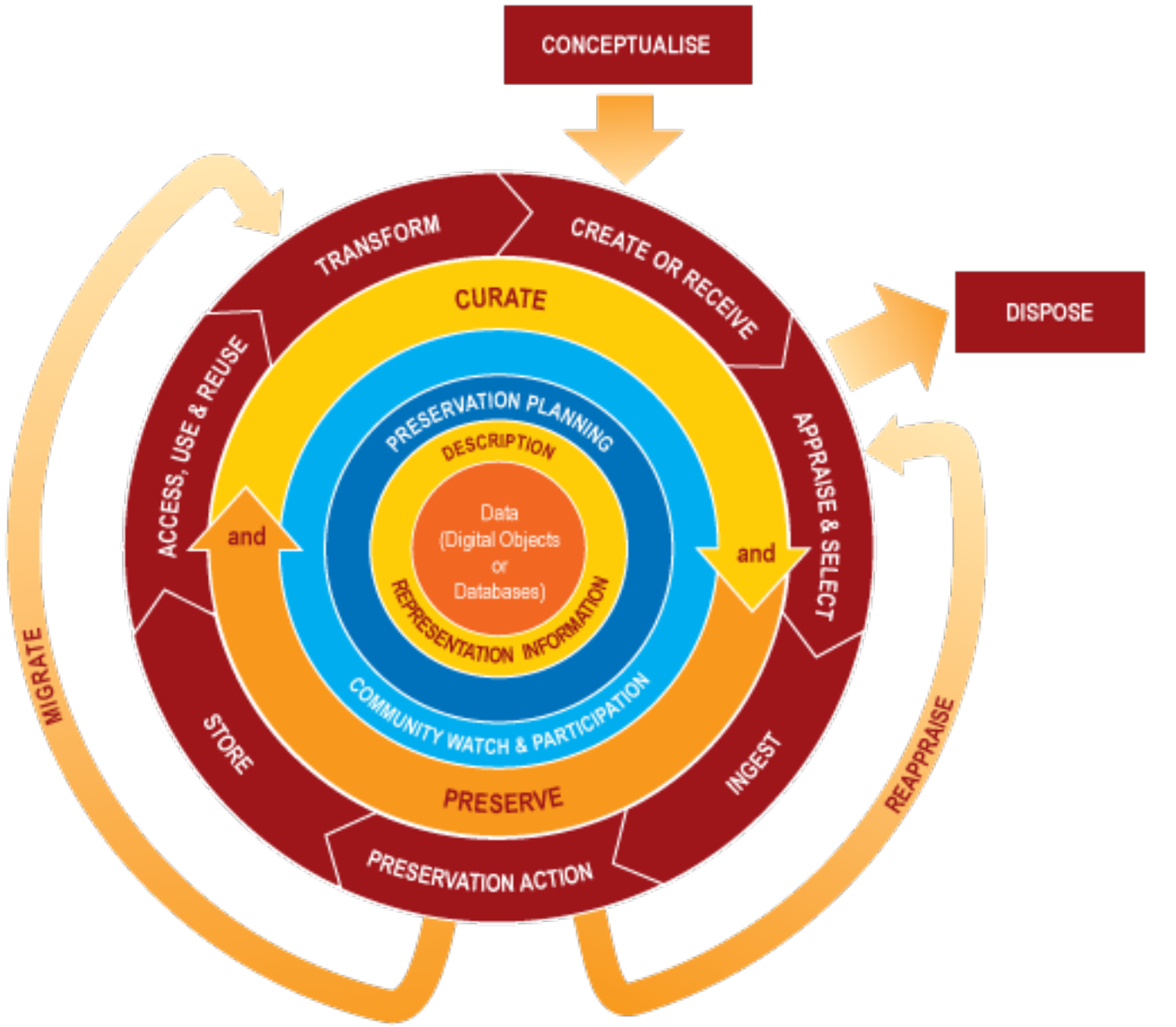


ETL: Simplest form of Curation Workflow

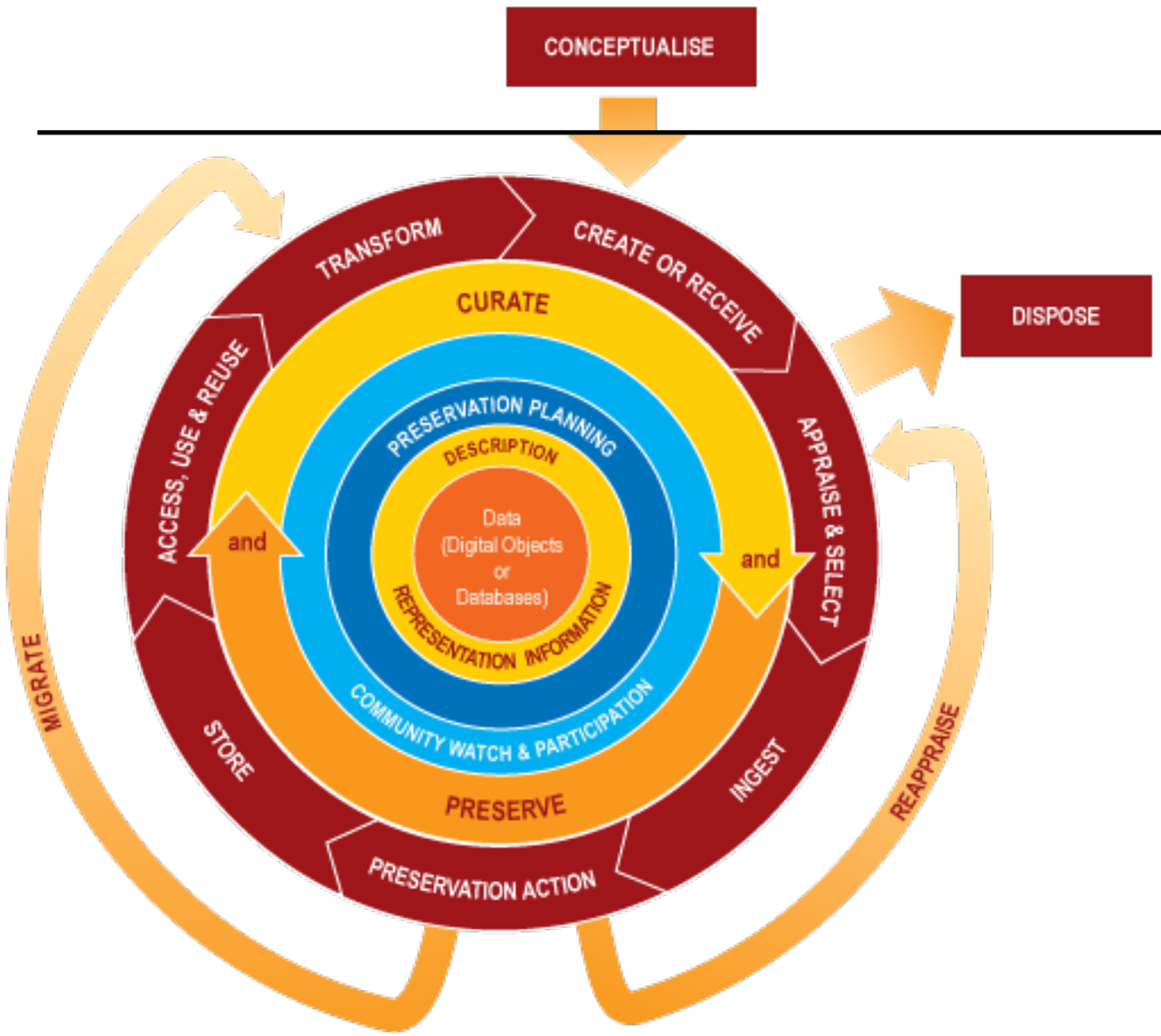



```
<text xmlns="http://www.tei-c.org/ns/1.0" xml:id="d1">
  <body xml:id="d2">
    <div1 type="book" xml:id="d3">
      <head>Songs of Innocence</head>
      <pb n="4"/>
      <div2 type="poem" xml:id="d4">
        <head>Introduction</head>
        <lg type="stanza">
          <l>Piping down the valleys wild, </l>
          <l>Piping songs of pleasant glee, </l>
          <l>On a cloud I saw a child, </l>
          <l>And he laughing said to me: </l>
        </lg>
        <lg type="stanza">
          <l>"Pipe a song about a Lamb!" </l>
          <l>So I piped with merry chear. </l>
          <l>"Piper, pipe that song again;" </l>
          <l>So I piped, he wept to hear. </l>
        </lg>
        <lg type="stanza">
          <l>"Drop thy pipe, thy happy pipe; </l>
          <l>Sing thy songs of happy chear:" </l>
          <l>So I sung the same again, </l>
          <l>While he wept with joy to hear. </l>
        </lg>
        <lg type="stanza">
          <l>"Piper, sit thee down and write </l>
          <l>In a book, that all may read." </l>
          <l>So he vanish'd from my sight, </l>
          <l>And I pluck'd a hollow reed, </l>
        </lg>
        <lg type="stanza">
          <l>And I made a rural pen, </l>
          <l>And I stain'd the water clear, </l>
          <l>And I wrote my happy songs </l>
          <l>Every child may joy to hear. </l>
        </lg>
```

Data Curation Lifecycle
(more complex workflow)



Data Curation Lifecycle
(more complex workflow)



Upstream

Downstream