Data Management
&
Data Management Plans

INFX 551
Winter 2018

Course Outline

Week 3
(final week in this module)

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

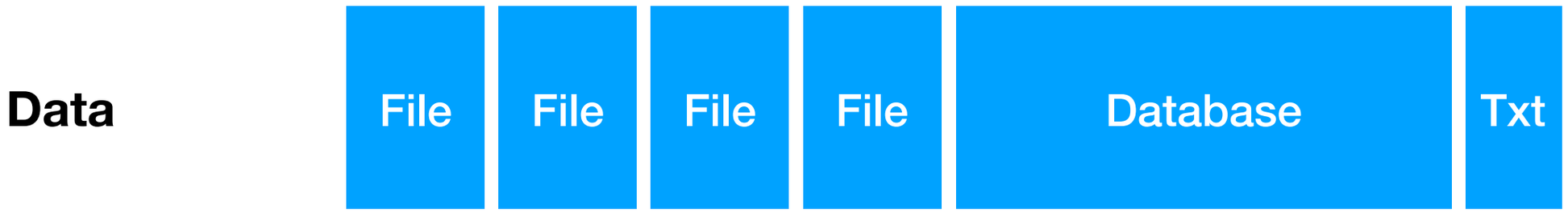
Topics

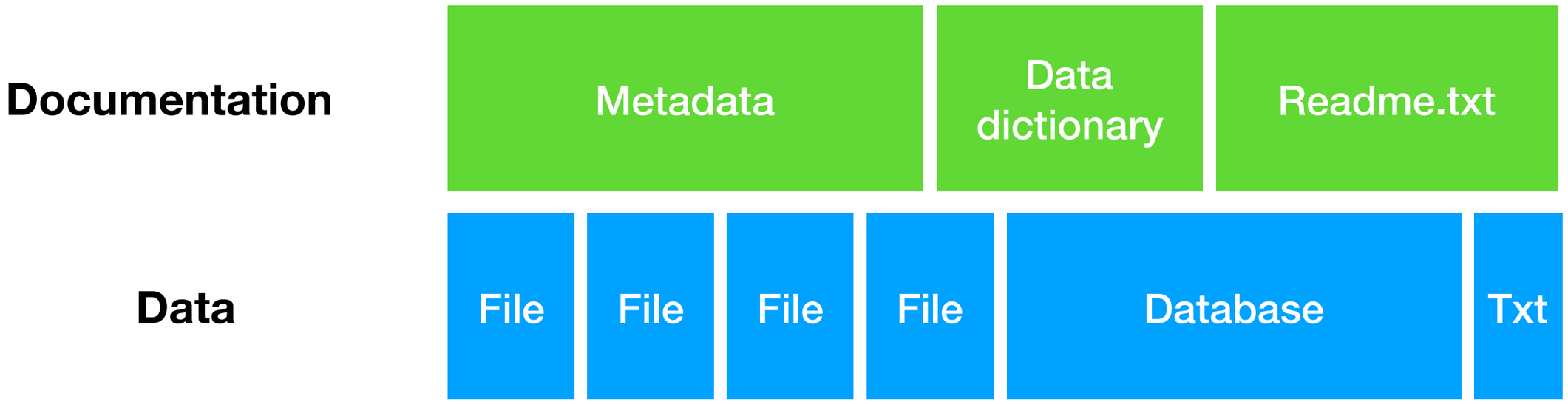
- Data Management
- Data Management Plans

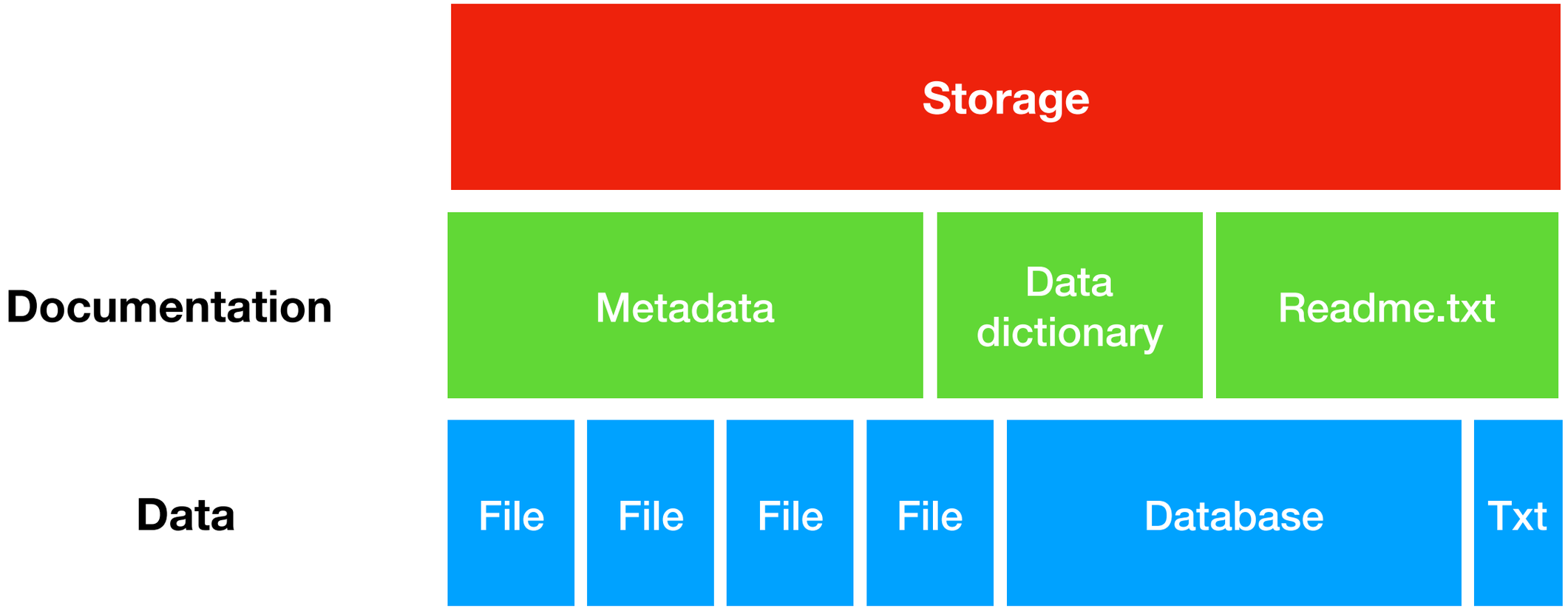
Data are various types of information objects playing a role of evidence

Data curation is the active and ongoing **management** of data throughout a lifecycle of use, including its reuse in unanticipated contexts

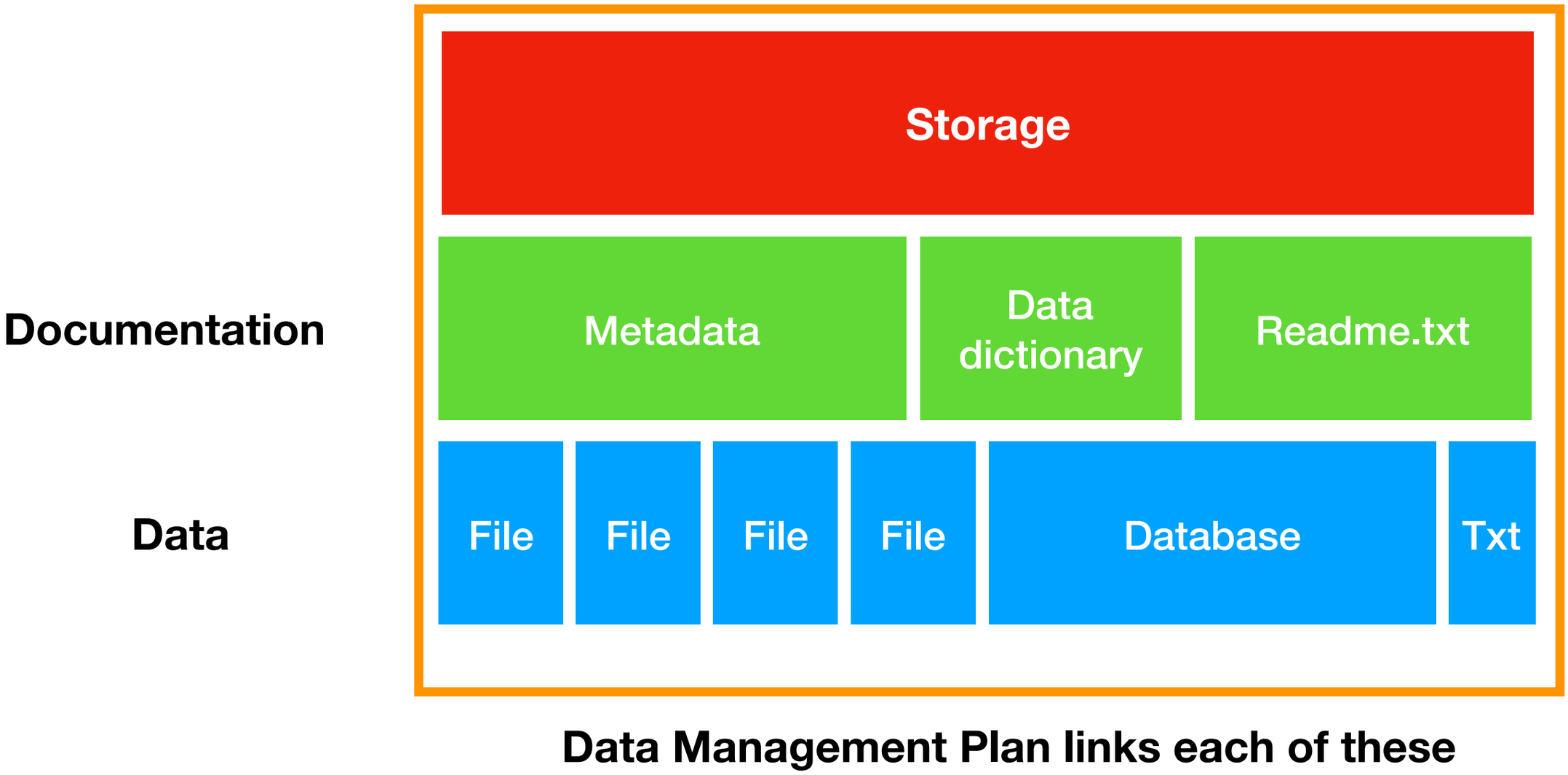
A data repository is both a digital archive for **storing and preserving data**, as well as an institution that establishes policies, and governance for **data management**.


Data
File
File
File
File
Database
Txt


Documentation
Metadata
Data dictionary
Readme.txt
Data
File
File
File
File
Database
Txt



A repository is both a physical archive for **storing data**, as well as an *institution* that establishes policies, and governance for **data management**.



A repository is both a physical archive for **storing data**, as well as an *institution* that establishes policies, and governance for **data management**.

Managing Data (general)

- **Types:** What types of data are being created, collected, and how should they be preserved

- **Roles:** Who might use the data, and how might those needs change over time?

- **Documentation:** What types of metadata standards are appropriate? What kinds of other documentation might also be useful?

- **Normalization:** Should data be normalized before being stored?

Data Management plan

(simply formalizes data management)

A **data management plan (DMP)** is a written document that describes the data you expect to acquire or generate during the course of a project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

DMPs are created by data producers with consultation of data curators…

Ideally before data collection begins…

This applies almost exclusively to Research Data that result from Grant Funded research (~90% of research in the USA).

Requirements

- National Institutes of Health (2007)
- National Science Foundation (2011)
- Federal Government Agencies (2011)
- National Endowment of the Humanities (NEH); NEA; Sloan Foundation; Bill & Melinda Gates Foundation, etc. (2013)

Benefits

- Improves data quality and decreases total time needed to complete clinical trials (Krishnankutty et al 2012)
- Increases usability of data from federally funded research projects (Henson, 2014)
- Provides higher quality metadata (Philson et al 2014)
- Reduces time to reuse and build upon previous studies (Hemminger et al, 2016)

Managing Data (specific to funding)

- **Description:** List the format, average size, volume, and/or estimated number of data files produced. Tools and methods used for collection. Who is involved in management at each stage of data collection? How will data storage costs be covered?

- **Standards:** What standards will be used to format the data? (XML, ASCII, CSV, MySQL, netCDF) Is the format open?

- **Metadata Standards:** What standards will be used to describe the data? Which files will metadata be created for? How, and who will create this documentation?

- **Data Access:** What controls will be set on who can access the data, when, and will they be charged anything to do so?

- **Intellectual Property and Reuse:** Licenses, copyright, etc.

Demo this week will be for the DMP tool.