Open & FAIR Data

INFX 551

Spring 2017

Information School
UNIVERSITY *of* WASHINGTON

- Discussion of readings
    - Integrated conceptual lecture (e.g. - I’ll be the annoying one that makes extended comments during discussion)
- Git - quick introduction
- Github - even quicker introduction
- Data Pitch
- Groups + Discussion

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.”

ands

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. This is the summary of the full Open Definition which the Open Knowledge Foundation created in 2005 to provide both a succinct explanation and a detailed definition of open data.”

OPEN KNOWLEDGE

Kitchen

Conceptualizing Data

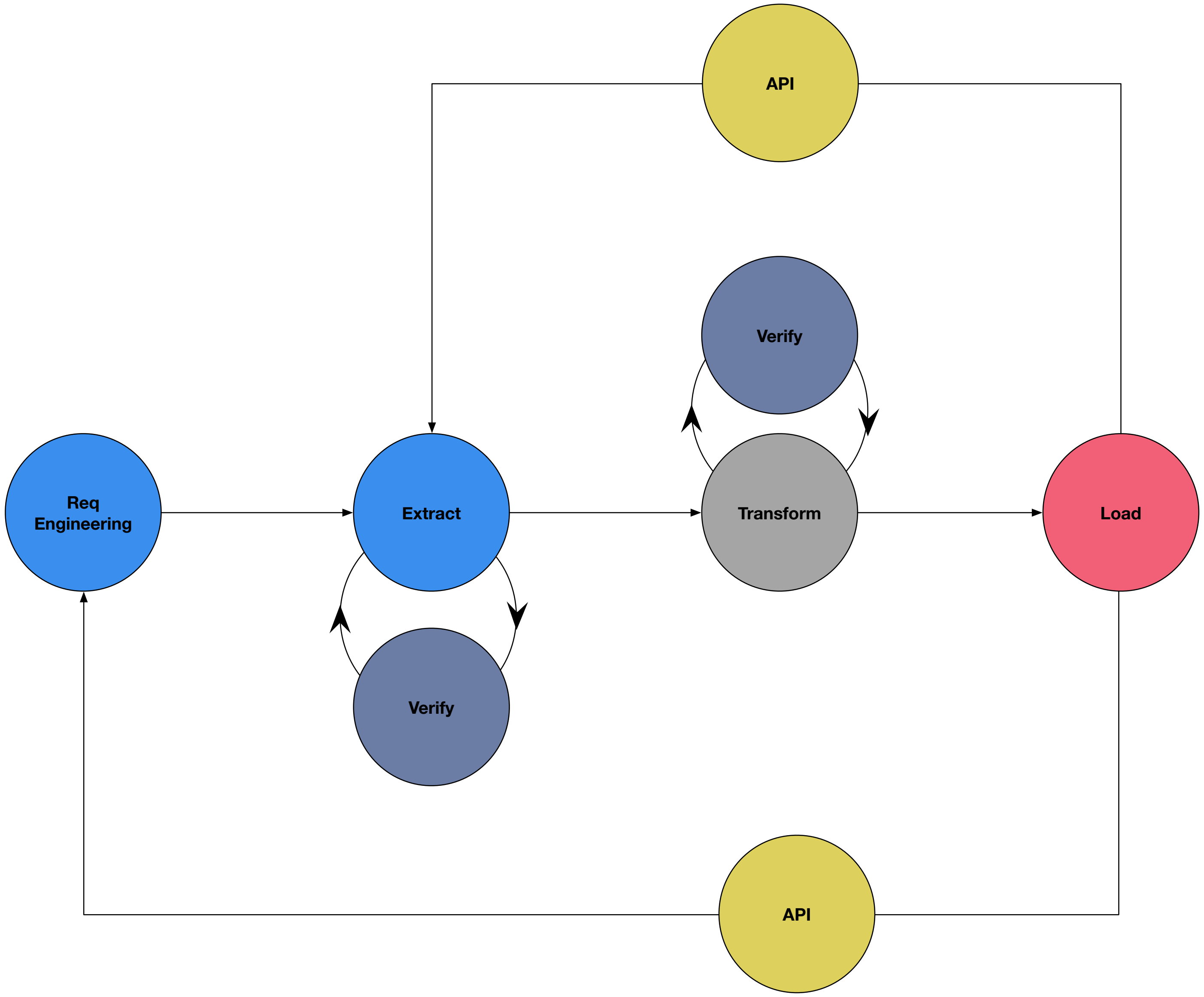
Epistemology

vs.

Ontology

?

TL;DR - “Data do not exist independently of the ideas, instruments, practices, contexts and knowledges used to generate, process and analyze them.”



Category 1

- Captured
- Exhaust (trace)
  - Transient
- Derived

Category 2

- Primary
- Secondary
- Tertiary

Databases and Data Infrastructures

"databases and data infrastructures do not simply support research, they fundamentally change the practices and organisation of research – the questions asked, how they are asked, how they are answered, how the answers are deployed, who is conducting the research and how they operate as researchers."

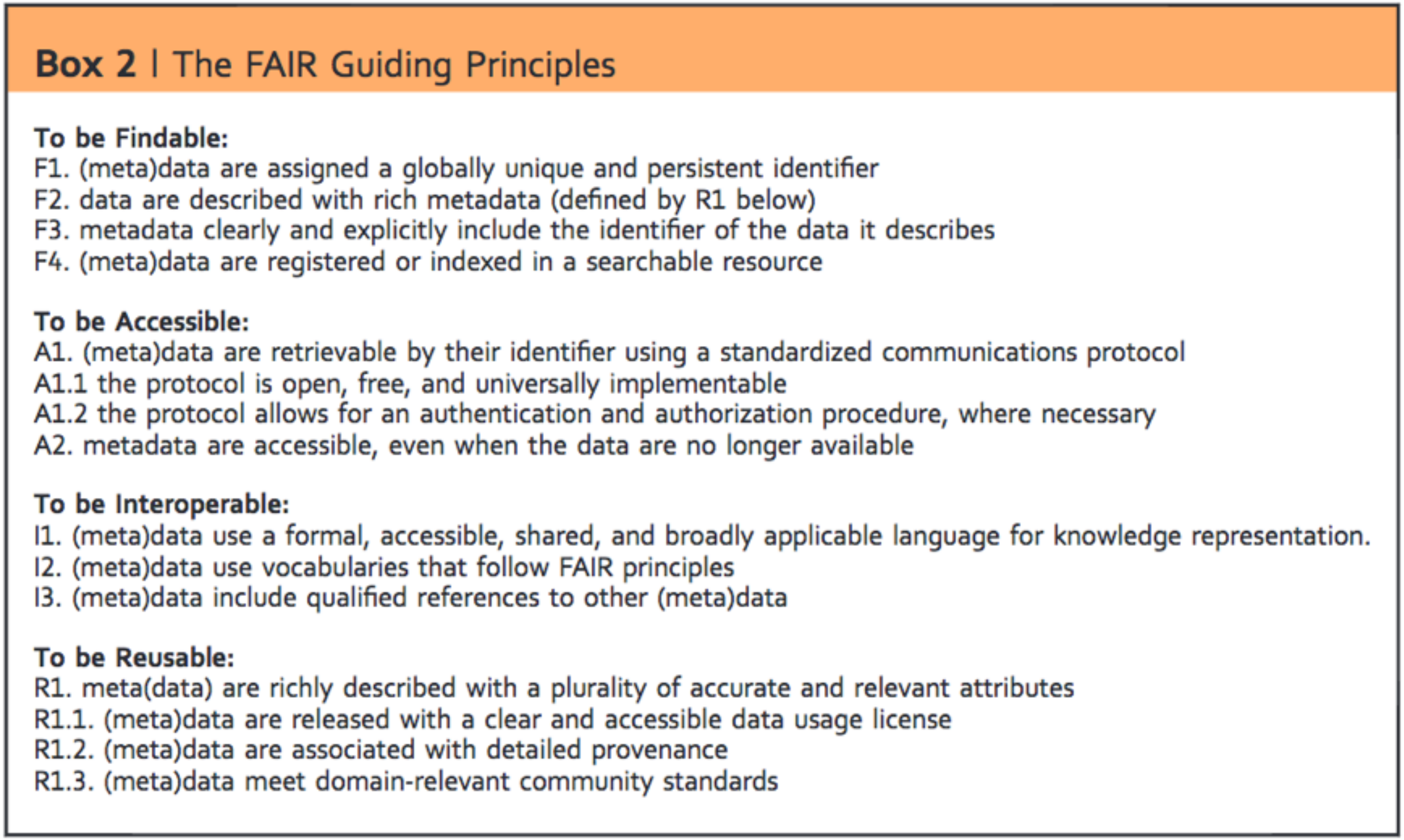
- Kitchin 2013

Table 1.3 The apparatus and elements of a data assemblage

Apparatus	Elements
Systems of thought	Modes of thinking, philosophies, theories, models, ideologies, rationalities, etc.
Forms of knowledge	Research texts, manuals, magazines, websites, experience, word of mouth, chat forums, etc.
Finance	Business models, investment, venture capital, grants, philanthropy, profit, etc.
Political economy	Policy, tax regimes, public and political opinion, ethical considerations, etc.
Governmentalities and legalities	Data standards, file formats, system requirements, protocols, regulations, laws, licensing, intellectual property regimes, etc.
Materialities and infrastructures	Paper/pens, computers, digital devices, sensors, scanners, databases, networks, servers, etc.
Practices	Techniques, ways of doing, learned behaviours, scientific conventions, etc.
Organisations and institutions	Archives, corporations, consultants, manufacturers, retailers, government agencies, universities, conferences, clubs and societies, committees and boards, communities of practice, etc.
Subjectivities and communities	Of data producers, curators, managers, analysts, scientists, politicians, users, citizens, etc.
Places	Labs, offices, field sites, data centres, server farms, business parks, etc., and their agglomerations.
Marketplace	For data, its derivatives (e.g., text, tables, graphs, maps), analysts, analytic software, interpretations, etc.

Wilkinson et al 2016

F.A.I.R data principles



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Choose one data curation perspective (science, humanities, government open data, or industry):

- **Gray**, J., Szalay, A. S., Thakar, A. R., Stoughton, C., & van denBerg, J. (2002). Online Scientific Data Curation, Publication, and Archiving. Microsoft Research Technical Report. Redmond, WA: Microsoft Research.

- **Curry**, E., Freitas, A., & O'Riáin, S. (2010). The role of community-driven data curation for enterprises. In D. Wood (Ed.), Linking Enterprise Data (pp. 25-47). New York, NY: Springer.

- **Flanders**, Julia, & Muñoz, Trevor. An introduction to humanities data curation. DH Curation Guide.

- **Sahuguet**, A., Krauss, J., Palacios, L., & Sangokoya, D. (2014). Open Civic Data: Of the People, For the People, By the People. IEEE Data Eng. Bull., 37(4), 15-26.

“Data is not in and of itself a **kind** of evidence but a multifaced object which can be mobilized as evidence in support of an argument.” (Owens, 2011)

- Data are information artifacts that are malleable - much like the ‘objects’ of traditional humanistic interpretation.
- Put another way, data can be read in ways that are similar to texts
- When computed upon - data can be used to produce novel readings (e.g. close vs. distant reading)

"…data is open to a range of hermeneutic tactics for interpretation. In much the same way that encoding a text is an interpretive act, so are creating, manipulating, transferring, exploring, and otherwise making use of data sets. Therefore, data is an artifact or a text that can hold the same potential evidentiary value as any other kind of artifact. "

–Owens, 2011

GIT lecture.

Data Pitch

1. Topic
2. User Community
3. Potential Data
4. Potential User Community

Groups

By next week

Name your group.

Start an ‘organization’ on Github.

Fork the ‘INFX 551’ repository.

Post a description of your group that is similar to my data pitch.