Research Lifecycles: Metadata

TDO Chapter 5

Why metadata?

- allowing others to independently understand and re-use your work / reproducibility

- standards are evolving and will become requirements for a data scientist to both generate and produce metadata.

Example from from TDO Chapter 5

Mt St Helen's pictures:

- the type of camera, lens, shutter speed, light sensitivity, aperture, and other settings,

- information about the geographic and temporal circumstances surrounding the image's creation: the date, time and location on Earth where the photograph is taken,

- biographical information about the photographer to help viewers relate to the photographer and better understand the photograph's context,

- licenses and copyright information to associate with the picture—who owns it and how it can be used.

“Descriptor”

- a term “assigned to a resource to designate its properties, characteristics, or meaning, or its relationships with other resources.”

- can also be called “keywords,”“index terms,” attributes, attribute values, elements, “data elements,” “data values,” “the vocabulary,” “variables,” “features,” “properties,” “measurements,” “labels,” or “tags.”

“Metadata”

“Metadata is … [a] structured description for information resources of any kind. Metadata is more useful when supported by a metadata schema that defines the elements in the structured description.”

“The resource descriptions themselves serve to enable discovery, reuse, access control, and the invocation of other resources needed for people or computational agents to effectively interact with the primary ones described by the metadata.”

Historical Note

“In 1986, the Standard Generalized Markup Language (SGML) formalized the Document Type Definition (DTD) as a metadata form for describing the structure and content elements in hierarchical and hypertextual document models. SGML was superseded in 1997 by eXtensible Markup Language (XML), whose purpose was structured and computer-processable web content.”

“Resource Description Framework”

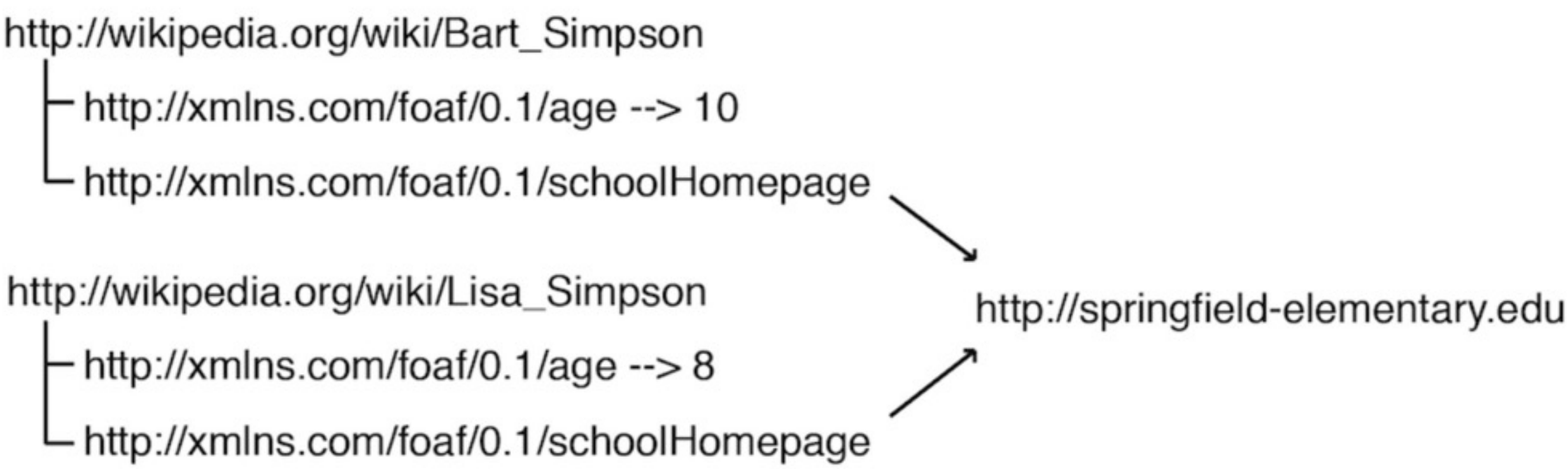
- “The Resource Description Framework (RDF) is a standard model for making computer-processable statements about web resources; it is the foundation for the vision of the Semantic Web.”

- “Use URIs to identify not only things “on” the web, like web pages, but also things “off” the web like people or countries. For example, we might use the URI http://springfield-elementary.edu/ to refer to Springfield Elementary itself, and not just the school’s web page.

- “RDF models all descriptions as sets of “triples,” where each triple consists of the resource being described (identified by a URI), a property, and a value.”

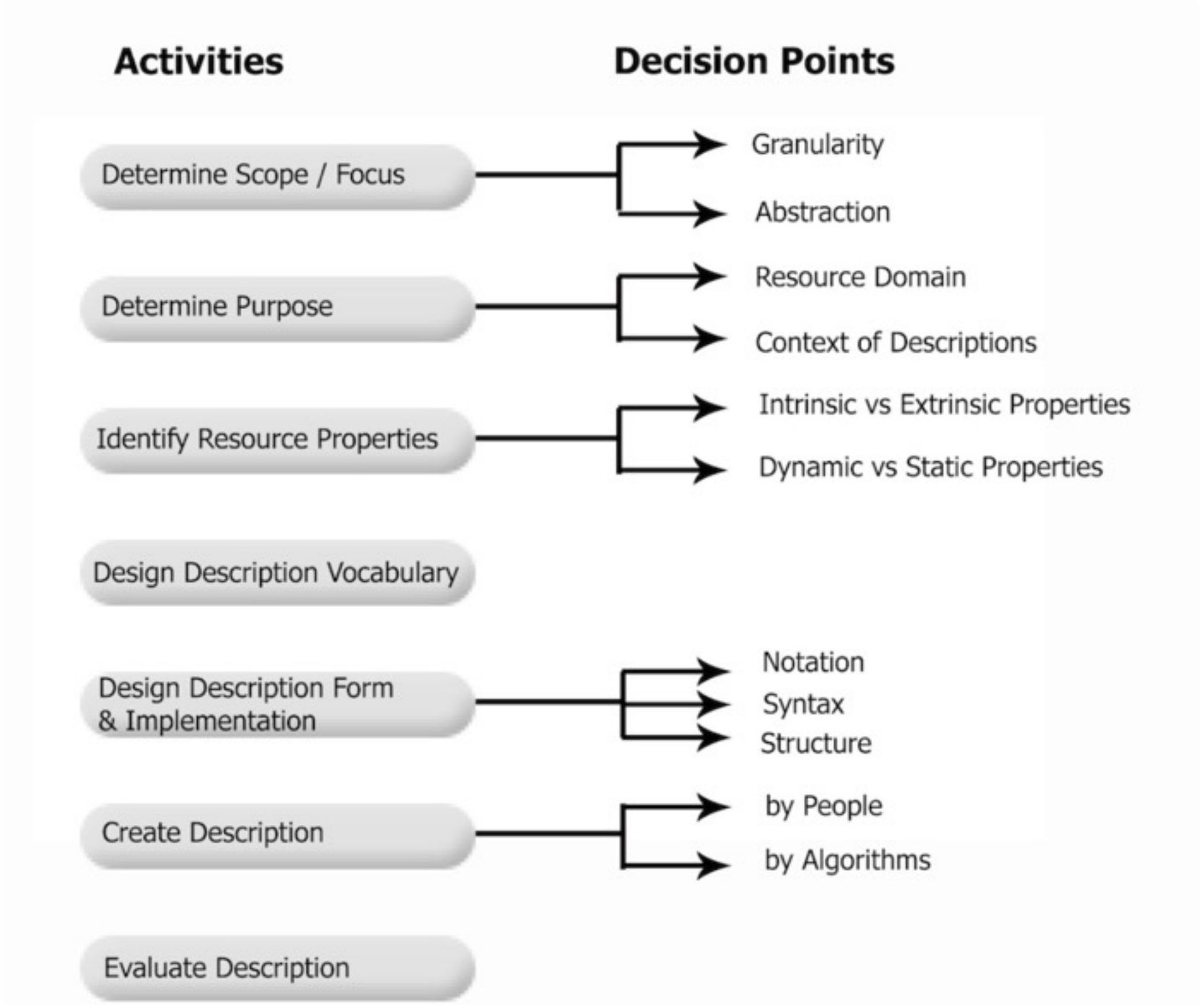
“Resource Description Framework”

Figure 5.1. RDF Triples Arranged as a Graph.



Two RDF triples can be connected to form a graph when they have a resource, property, or value in common. In this example RDF triples that make a statement about the home page of the elementary school attended by Bart Simpson and Lisa Simpson can be connected because they have the same value, namely the URI for Springfield Elementary.

Figure 5.3. The Process of Describing Resources.



The process of describing resources consists of seven steps: Determining the scope and focus, determining the purpose, identifying resource properties, designing the description vocabulary, designing the description form and implementation, creating the descriptions, and evaluating the descriptions.

Principles for Choosing Descriptions

- **User Convenience**: Choose description terms with the user in mind; these are likely to be terms in common usage among the target audience.

- **Representation**: Use descriptions that reflect how the resources describe themselves; assume that self-descriptions are accurate.

- **Sufficiency and Necessity**: Descriptions should have enough information to serve their purposes and not contain information that is not necessary for some purpose; this might imply excluding some aspects of self-descriptions that are insignificant.

- **Standardization**: Standardize descriptions to the extent practical, but also use aliasing to al low for commonly used terms.

- **Integration**: Prefer the same properties and terms for all types of resources.

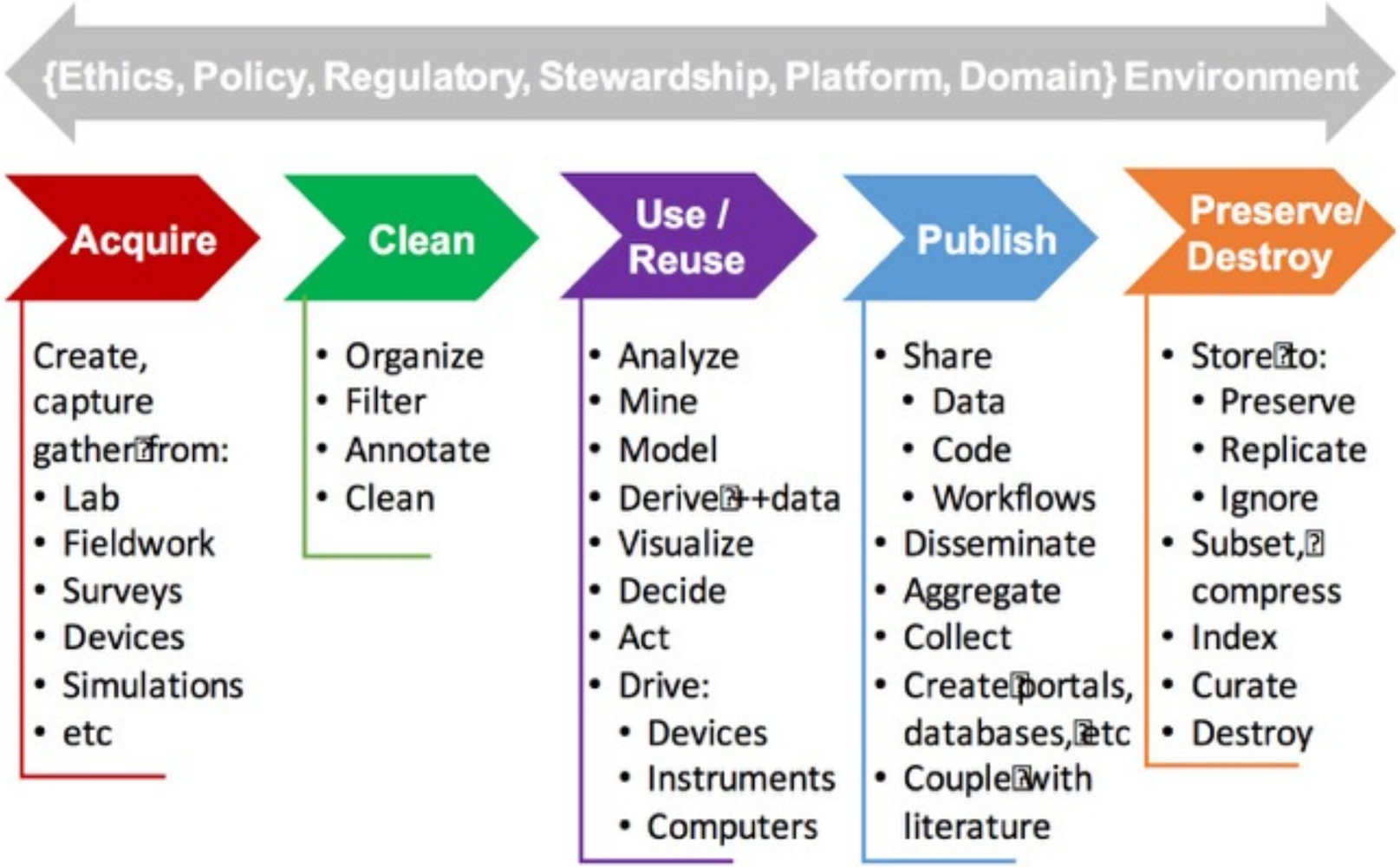


FIGURE 1: The Data Life Cycle and Surrounding Data Ecosystem

The *Data Life Cycle* is critical to understanding the opportunities and challenges of making the most of digital data. **Figure 1** shows a simplified cartoon with essential components of the data life cycle. Data is *acquired* from some source (measured, observed, generated), *cleaned* and edited to remove the outliers inevitable in real-world measurement scenarios and render it suitable for subsequent analysis; *used* (or reused) via some analysis leading to insight, action, or decision; *published* or disseminated in some way so the community at large is made aware of the data and its outcome(s); *preserved* (or not) so that others can revisit and reuse this data now or in the future. Surrounding this overall pipeline is a broader *environment* of concerns: *stewardship* to maximize the quality of the data and promote effective use, *ethics* issues that touch on proper or improper actions with these data; *policy* and *regulatory* constraints that impose legal limitations on these data; *platform* and infrastructure issues that affect technically how we can work with data; and *domain* and disciplinary needs specific to the application communities that create, operate, and use the data from these pipelines.