

## Slide 1



A typical organization produces or has access to a wide variety of data sources, each including a totally different set of data types, formats, coding systems and semantics.

In order to cope with the challenge of being able to map, identify and be familiar with our organizational data, we must employ tools to describe it in a way so that is reasonably doable to collect, integrate and eventually analyse at any point of time.

Metadata is a primary tool to do this.

### What is metadata?



Source: IMDB.



What is metadata?

Simply put, metadata is data which describes data.

Let's have a look at this movie, for example, as it is maintained in the films database IMDB.

The film itself is our core data. But in order to find it using search, or just to be able to get a first impression, we need to have some features describing it.

### What is metadata?

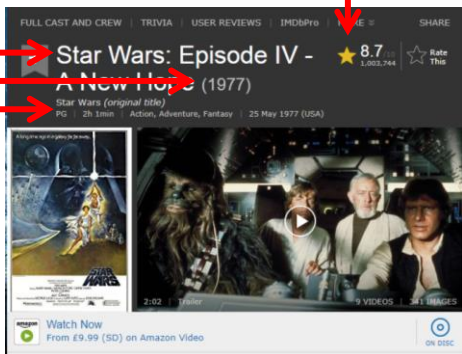


Source: IMDB.



What do we know about this film? Its title, for instance.


What is metadata?



The screenshot shows the IMDb page for 'Star Wars: Episode IV - A New Hope (1977)'. Red arrows point to the following metadata fields:


- Star Wars: Episode IV - A New Hope (1977)
- Star Wars (original title)
- PG - 2h 1min - Action, Adventure, Fantasy - 25 May 1977 (USA)
- 8.7 (1,002,744)
- Rate This

Source: IMDB.



Its release year.  
Its parental guidance classification.  
Its rating.


What is metadata?



The screenshot shows the IMDb page for 'Star Wars: Episode IV - A New Hope (1977)'. Red arrows point to the following metadata elements:

- Star Wars: Episode IV - A New Hope (1977)
- Star Wars (original title)
- PG - 28.1min - Action, Adventure, Fantasy
- 25 May 1977 (USA)
- 8.7 (1,002,744)
- Watch Now From £9.99 (SD) on Amazon Video
- ON DISC

Source: IMDB.



Its length.

Its type.

Its release location, etc.

These are all Metadata: data that describes data.

Metadata is required for discovery and search of resources (such as films, books, customer records or orders), for resource management, for access control and for long term preservation.

For these reasons, metadata must be machine-readable, descriptive and consistent.

**What is metadata?**

---

*"Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information."*

-- National Information Standards Organization  
<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>

Metadata provides information enabling to make sense of **data** (e.g. documents, images, datasets), **concepts** (e.g. classification schemes) and **real-world entities** (e.g. people, organisations, places, paintings, products).

 OPEN DATASUPPORT

Slide 6



Here are two good definitions from the National Information Standards Organization and from Open Data Support.

“Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve, use or manage an information resource.”


**Who uses metadata?**

---

***Metadata in Action: Amazon and its Affiliates***

Amazon.com has a worldwide online retail presence covering many different categories of goods. Metadata drives many parts of the company's operations. The metadata starts with the publishers of books or providers of other types of goods, as part of their own inventory systems. These suppliers send this metadata to Amazon, which integrates it with similar kinds of information from thousands of other providers to build its own website and sell products to users. Amazon collects metadata on sales and further uses it to provide customers with recommendations and optimize its relationships with suppliers. Amazon also makes the metadata about the products it brokers available to affiliate sites that build their own services on top of it, increasing sales through Amazon and driving business to the original supplier.

Source: Riley, J., 2017



Who uses metadata? Well, every organization that holds an amount of data that is challenging to manage.

Not just Amazon about their products. Companies must store metadata about their customers, employees and the events they manage; libraries and museums about their books or other assets; governments about their citizens, their allies, and their enemies—this is all metadata.

Social media platforms like Facebook store likes, shares and friendships. Facebook then uses it to track, analyse, suggest and promote. Pinterest users create boards that categorize and describe items and then use this metadata to search and recommend. So does Instagram, Twitter and others (Riley 2017).

### Why use metadata?

---



<http://shiyali.blogspot.co.uk/2013/06/books-in-big-house.html>

Organizations reach a point of storing an amount of data that is challenging to manage very early in their life span.

At this point, it is impossible to be familiar and locate relevant data for analysis.



Why use metadata?

---



Saved from:  
[journal.jennileemarigomen.com](http://journal.jennileemarigomen.com)



Saved from:  
<http://shiyali.blogspot.co.uk/2013/06/books-in-big-house.html>



Saved from:  
<https://vicbooks.wordpress.com/2011/05/03/cool-bookshelves-part-i/>



Most likely, a typical organization looks like this: where data originates from more than one source.

Perhaps the same book sits in two different piles. Perhaps it even looks different.

### Why use metadata?

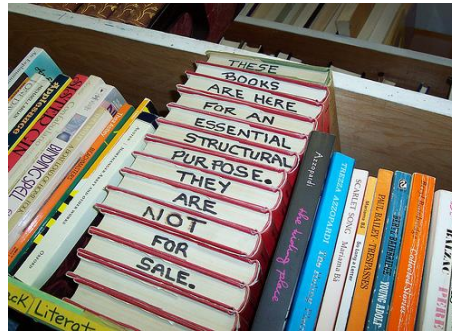


<https://wutheringwhimsy.wordpress.com/>



If we maintain some information about each item, customer, book or purchase, we can then make it simpler to find, sort, and locate what we need.

### How to manage metadata?

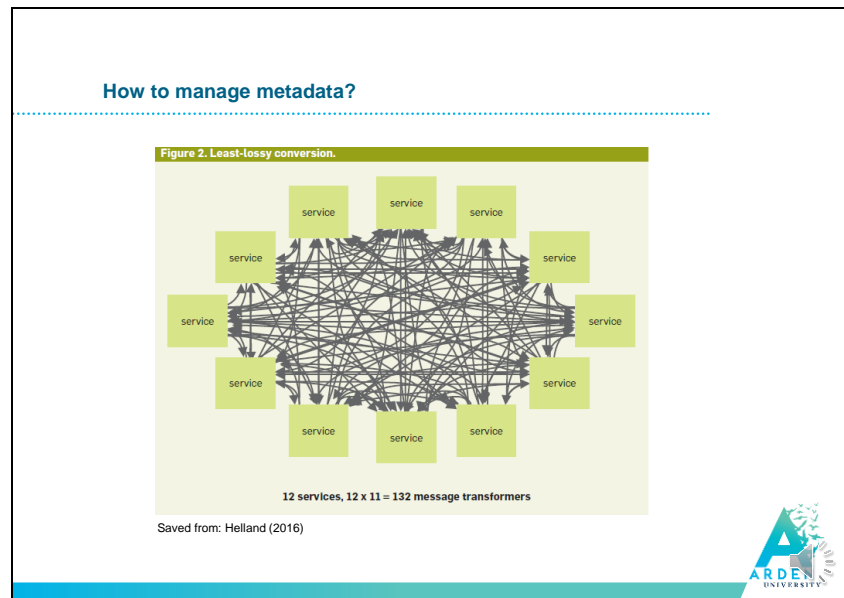


Saved from: <https://vicbooks.wordpress.com/2011/05/03/cool-bookshelves-part-i/>



### How to manage Metadata?

For an organization to know which variables of data are available from which systems and how to retrieve it, we need to manage some central infrastructure to support the description of all data sets and sources.



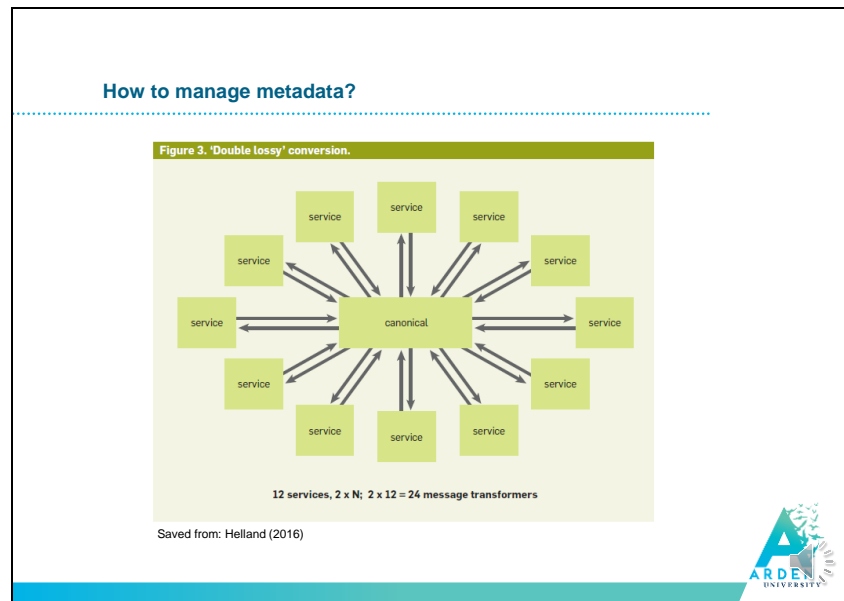
When applications or services are independently developed, it is only natural that they use different naming, concepts and representations to store the same information.

Since the information about the same resource might reside in two or more data sources, we will need to be able to translate it. For example, if we have data about customer acquisition in one system, data about their continuous orders from another system, and logs of their actions from a website's monitoring analytics, we might have to be able to triangulate this data.

However, their name or details might not be identical in all the systems and variables about them might be called by different names, stored using different types or formats and accessed differently.

Each translation can be very lossy. By the time the translation occurs, a loss of knowledge has occurred.

Merely maintaining a translation between each pair of sources might result in a pile of knots, which is very hard to maintain and understand.



One of the solutions to metadata management is to capture a canonical representation, where for each service we just maintain a translation to/from the canonical representation.

Using canonical metadata reduces the number of translations that need to be maintained.

Such canonical representations are sometimes called Data Standards.

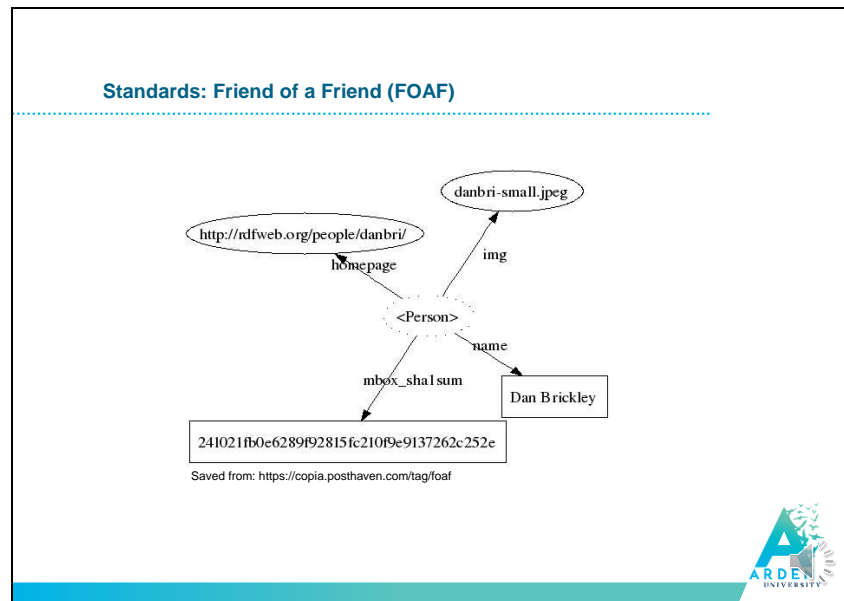
### Standards

---

- <http://schema.org/>
  - a collaboration between Yahoo, Bing and Google to make search have more canonical semantics
  - includes formats for metadata about common entities, such as person, place, event, organization, etc.
- Dublin Core: <http://dublincore.org/>
  - metadata for published material
- FOAF
  - metadata for people and organizations
- Vertical specific standards (e.g., for healthcare data)
- etc.

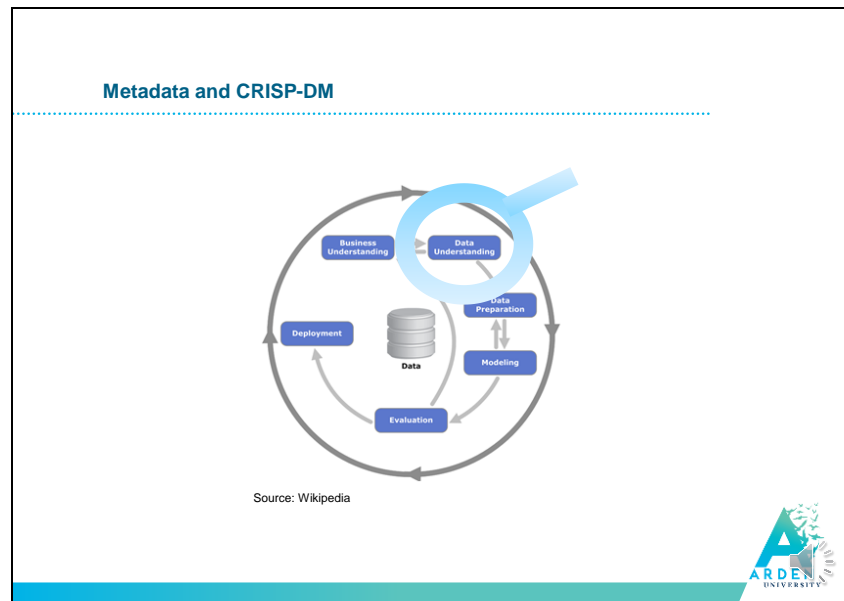


Just as we use industry standards such as TCP, IP, XML or JSON to make it easier to communicate in a machine-readable format, there have been developed many industry standards to canonically describe metadata.



For example: the FOAF (Friend of a Friend standard) provides descriptive metadata about resources such as person, organization and project, along with a list of properties describing how people and organizations interact with each other, with social media and more.

An organization can use such industry standards to represent its metadata canonically, or if the standard does not represent its data truthfully enough, to extend it or optionally even develop its own canonical representation.

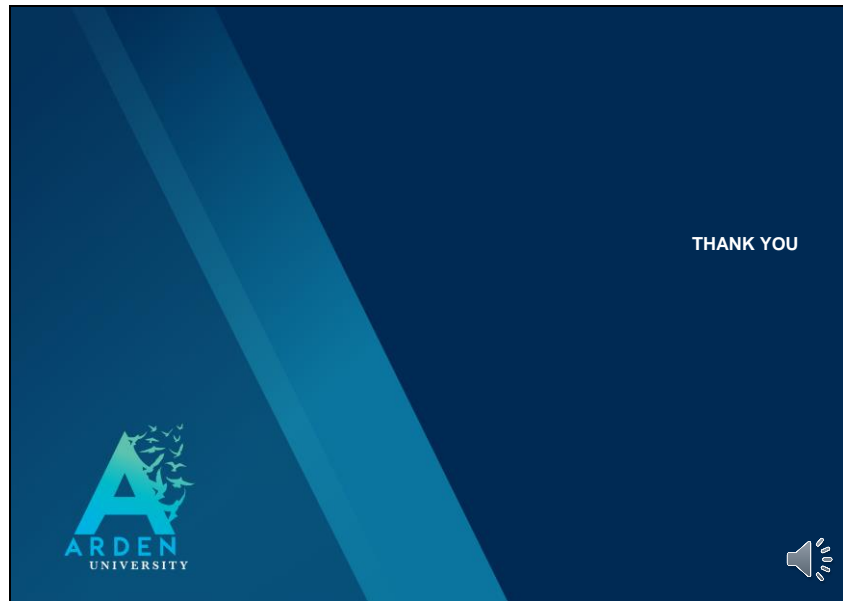


Back to CRISP-DM: recording metadata across the whole cycle can help us better communicate between different phases and teams and make sure the collection, processing, analysis and deployment are done right.

If we fail to do so, we are very likely to introduce mismatches between the apparent meaning of fields and the meaning stated in field names or definition.



Slide 17



Thank you.