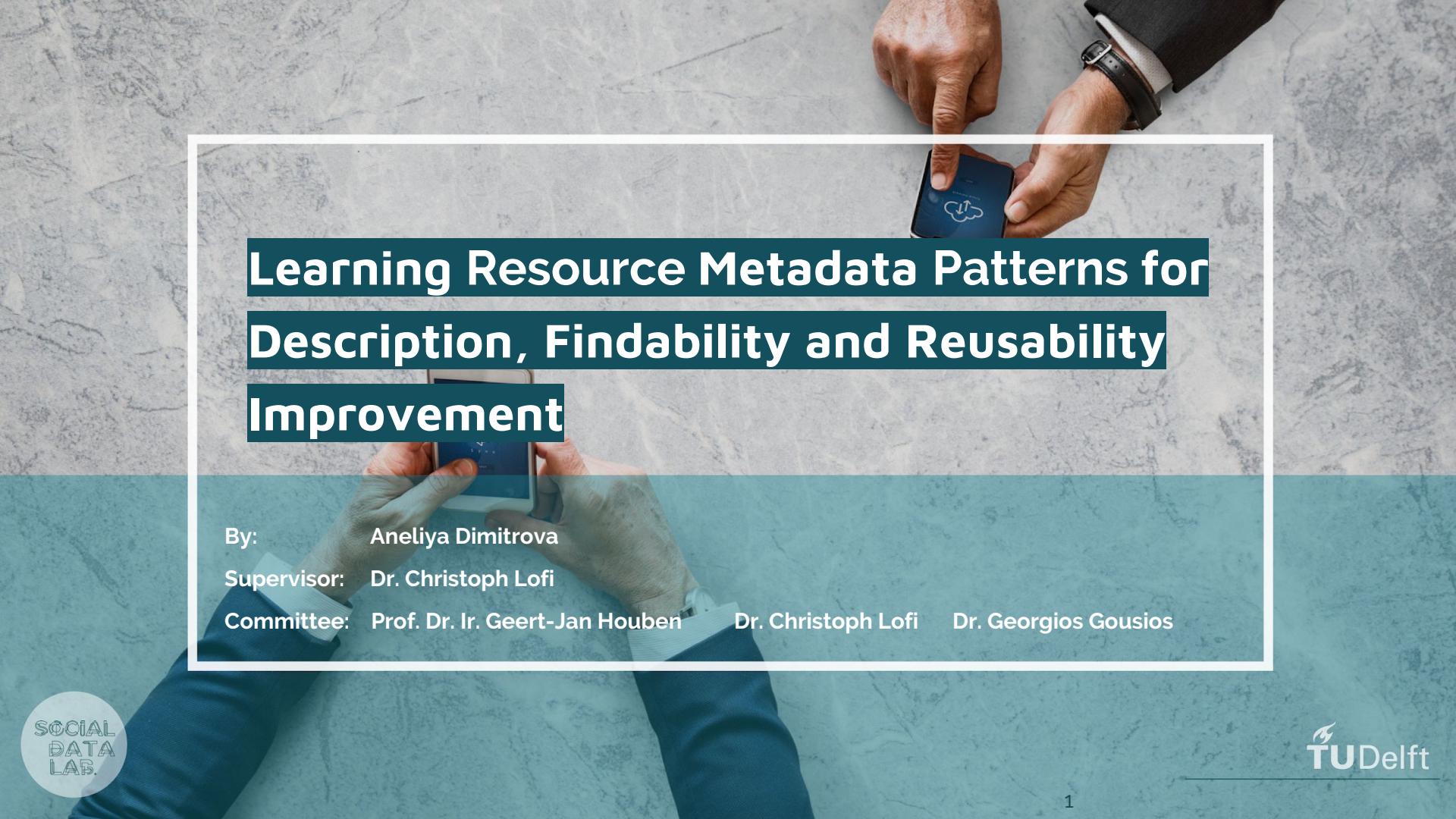


Learning Resource Metadata Patterns for Description, Findability and Reusability Improvement

By: Aneliya Dimitrova

Supervisor: Dr. Christoph Lofi

Committee: Prof. Dr. Ir. Geert-Jan Houben Dr. Christoph Lofi Dr. Georgios Gousios

SOCIAL DATA LAB.

TUDelft

Consequences of the rise of popularity of online education

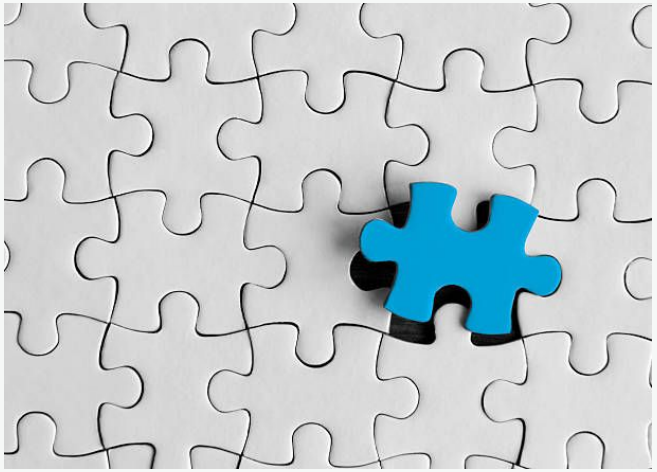
Consequences of the rise of popularity of online education

- A lot of educational resources are being created
- Hard to organize them

Consequences of the rise of popularity of online education

- A lot of educational resources are being created
- Hard to organize them
- Findability issues for authors and learners
  - Hard to find the right content among hundreds of options

How to solve this?
With metadata



Current issues with metadata

- **Completeness**: number of fields utilized
- **Accuracy**: correctness of the metadata values
- **Consistency:** equal data written differently
  (e.g. affiliation: “TU Delft”, affiliation: “Delft University of Technology; affiliation: “University of Delft” etc.)
- Difference between metadata schema **standards** applied across the different LMS systems

However:
Obtaining high-quality metadata is hard and potentially expensive

How to solve this issue?

Provide easily understandable workflows to practitioners to support them in collecting complete and high quality metadata.

Contributions

1. Extensive **analysis** on state-of-the-art for metadata generation
   a. Identify **Gaps** between literature and practice
   b. Generalize into a metadata type **taxonomy**

2. **Design patterns** for metadata extraction

3. **Proof of concept**

TU Delft

State-of-the-art
Metadata standards

Metadata **Core schema standards**

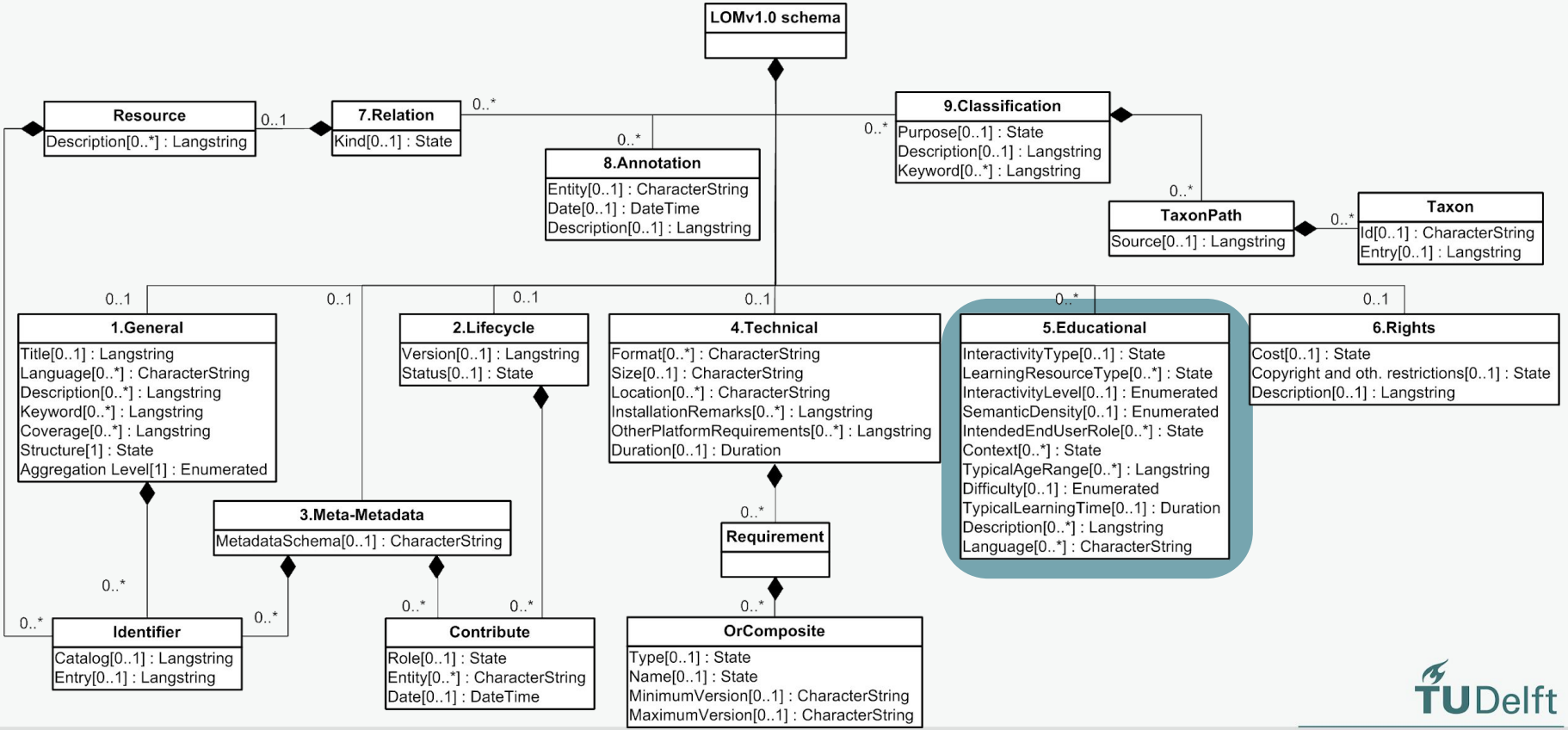
- **LOM** (a common work between IEEE and IMS Global Learning Consortium)
  - 76 elements
- **Dublin Core (DC)**
  - 15 elements
- IMS LD
- ARIADNE
- Sharable Content Object Reference Model (SCORM)

Application Profiles

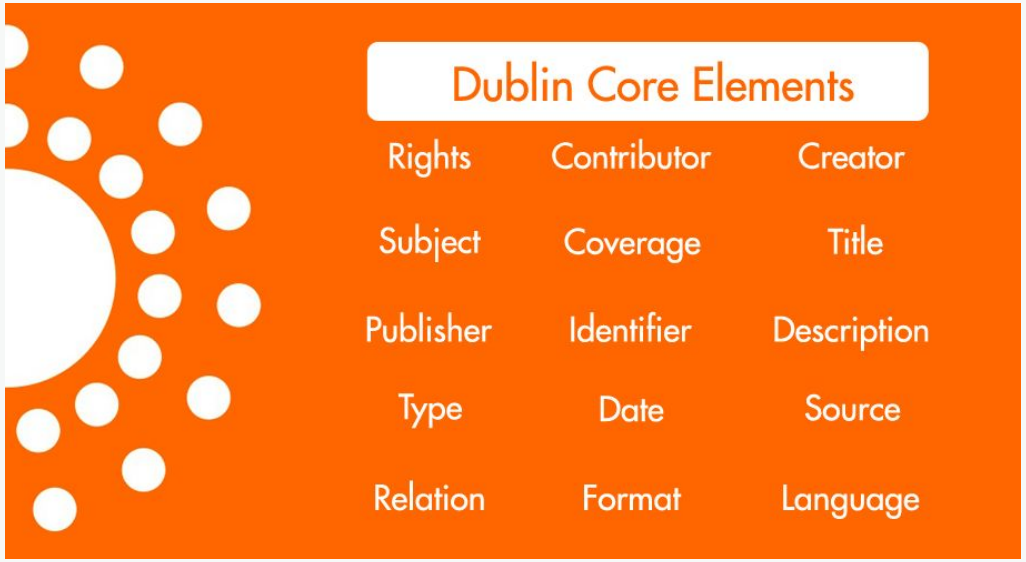
(Combination of metadata elements from different schemas and combined into a new one)

- DC-ED
- UK Learning Object Metadata Core
- Many LOM-based application profiles

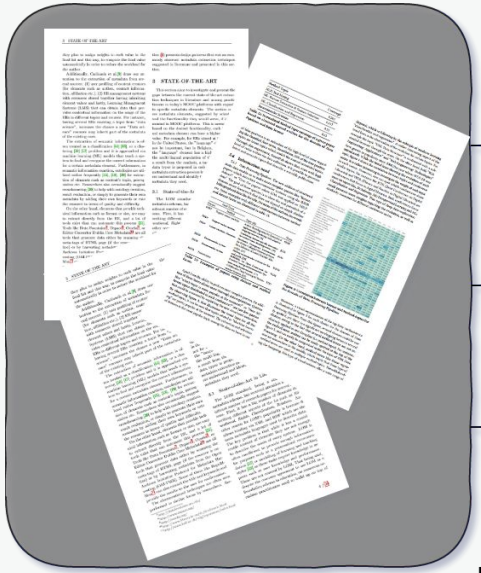
State-of-the-art
Metadata standards - LOM



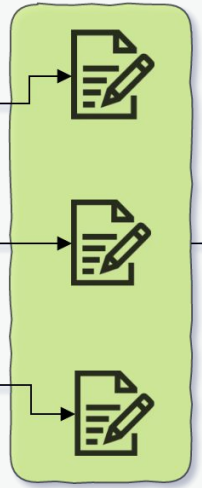
State-of-the-art
Metadata standards - Dublin Core



Towards metadata type taxonomy

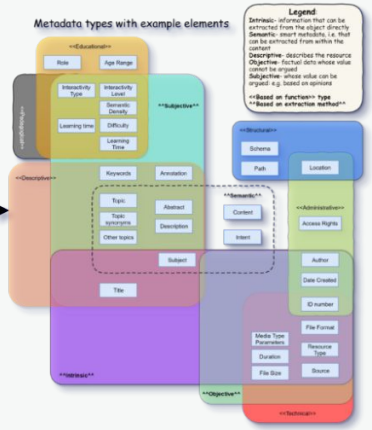


Extensive literature research



Extracting the metadata elements and how they fill the values

Clustered by function and extraction method



Metadata taxonomy

State-of-the-art
Metadata types

By functionality:

Descriptive Metadata: describes the content of the educational resource (ER)

Educational: educational aspects of the ER

Technical: technical parameters of the ER entity

Structural Metadata: for navigation

Administrative Metadata: for content management

By way of extracting:

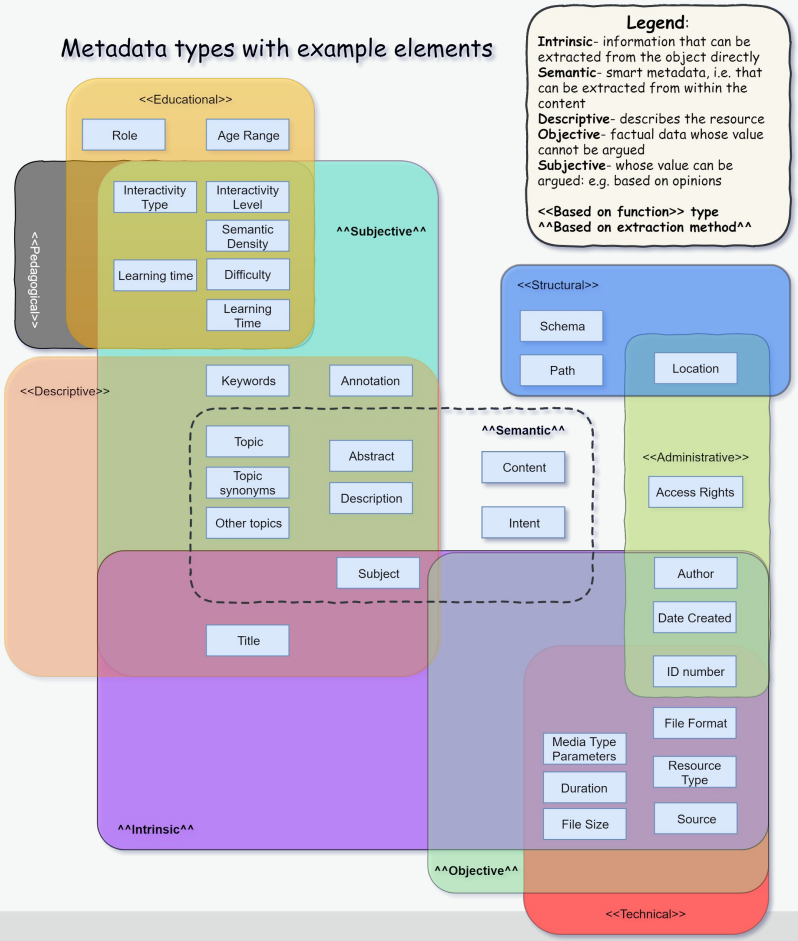
Intrinsic: can be extracted directly from the ER (e.g. technical and some descriptive elements)

Objective: factual data

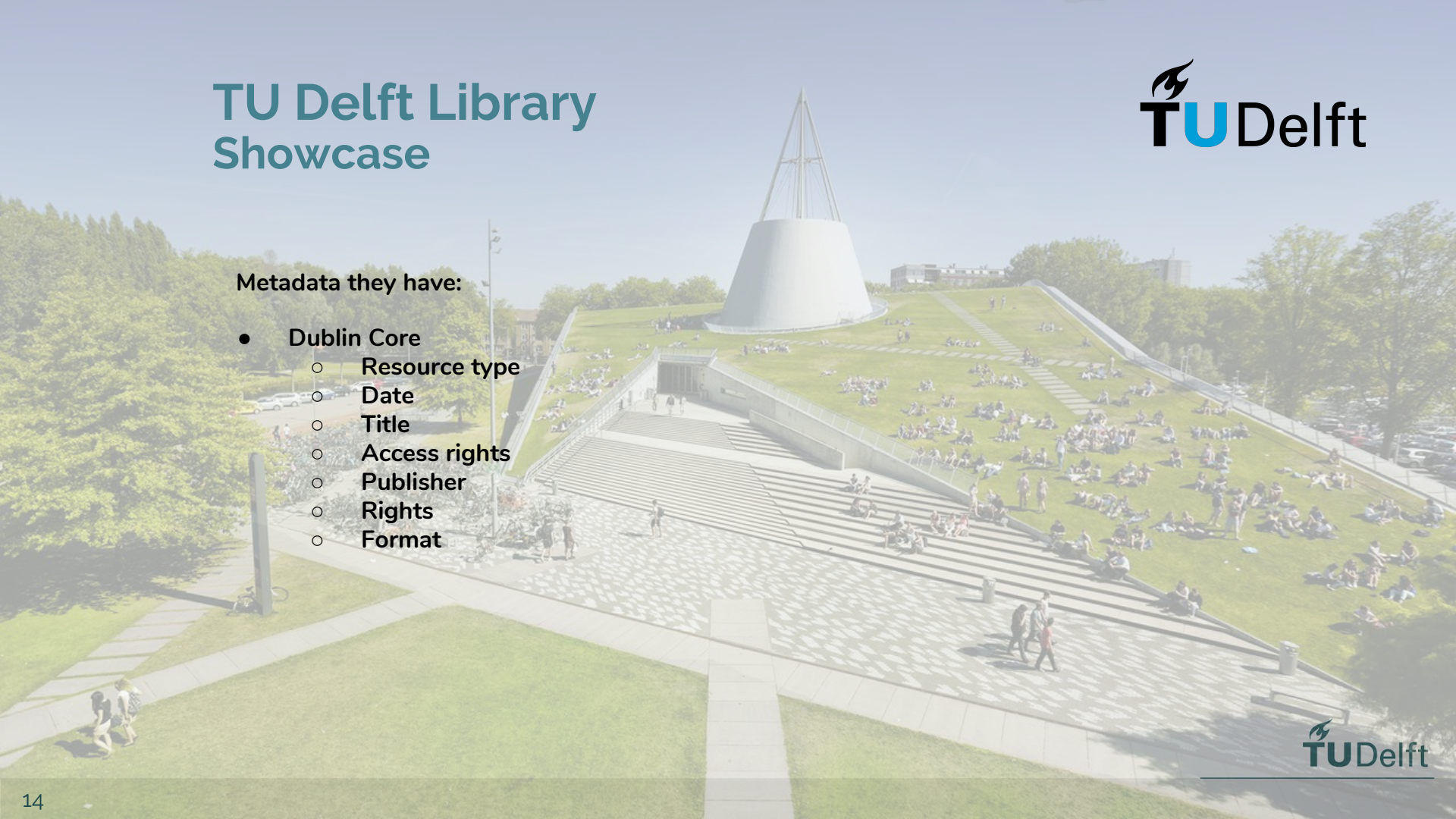
Subjective: based on opinion

Semantic: relying on content analysis

Metadata type taxonomy



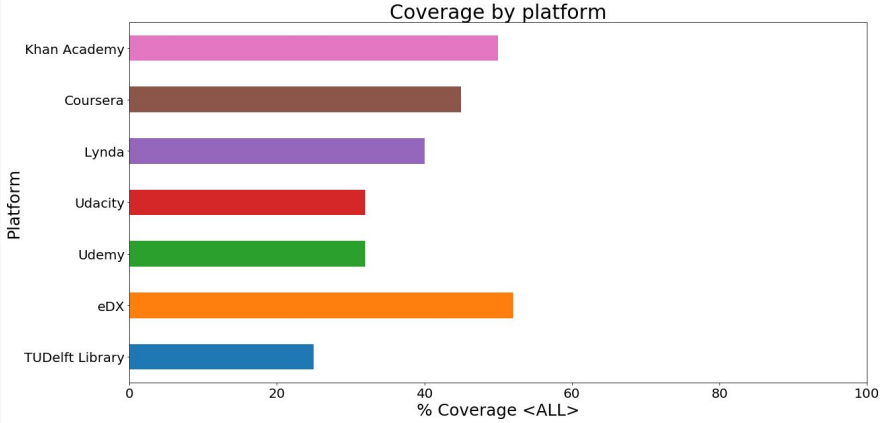
TU Delft Library
Showcase

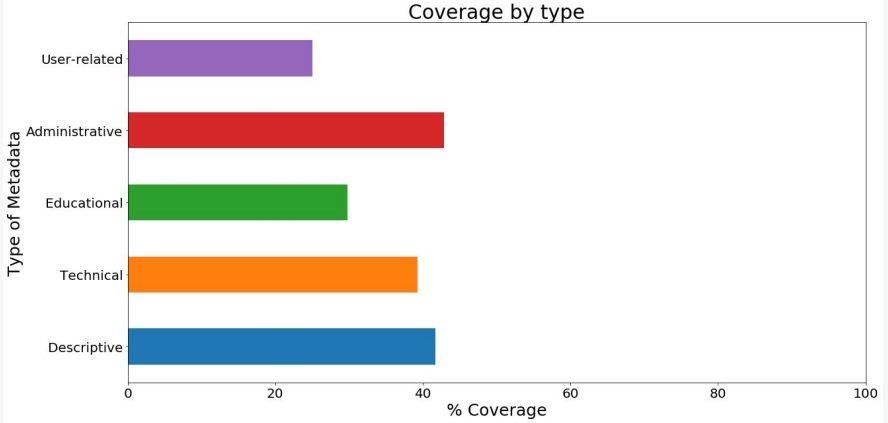


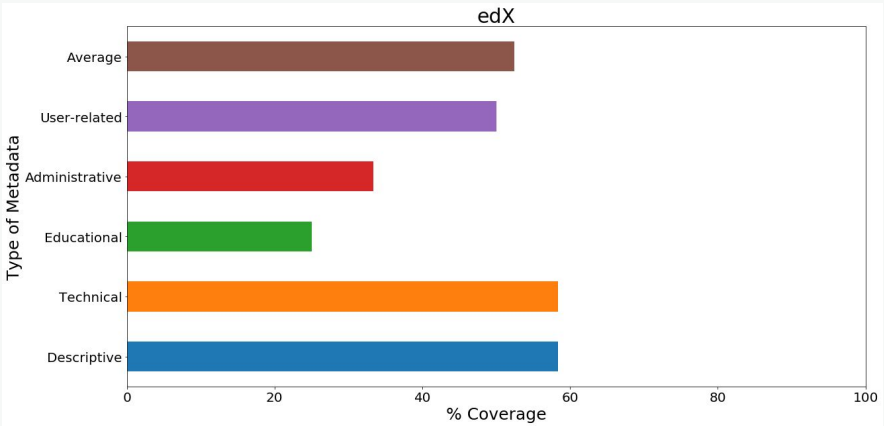
Metadata they have:

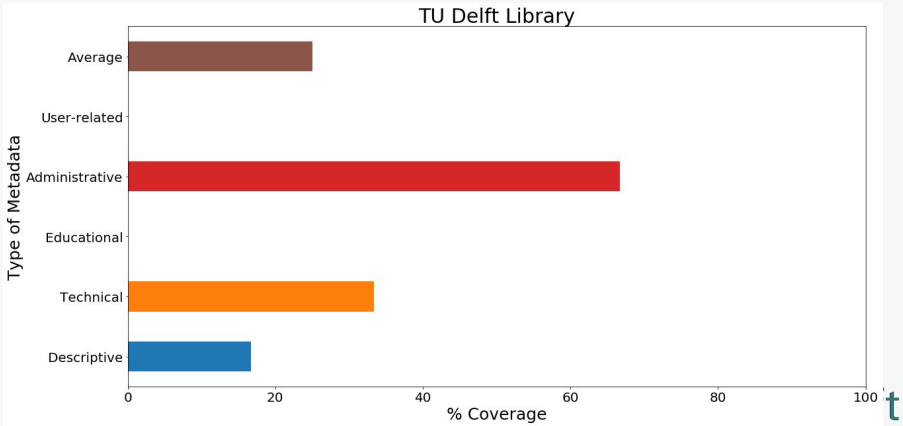
- **Dublin Core**
  - **Resource type**
  - **Date**
  - **Title**
  - **Access rights**
  - **Publisher**
  - **Rights**
  - **Format**

State-of-the-art
MOOC Platforms









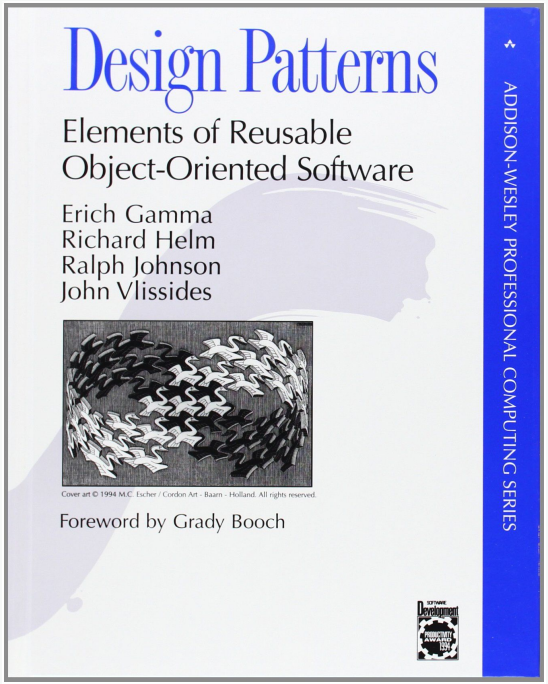
Extracting metadata

Size or **Format** vs. **Semantic Density**

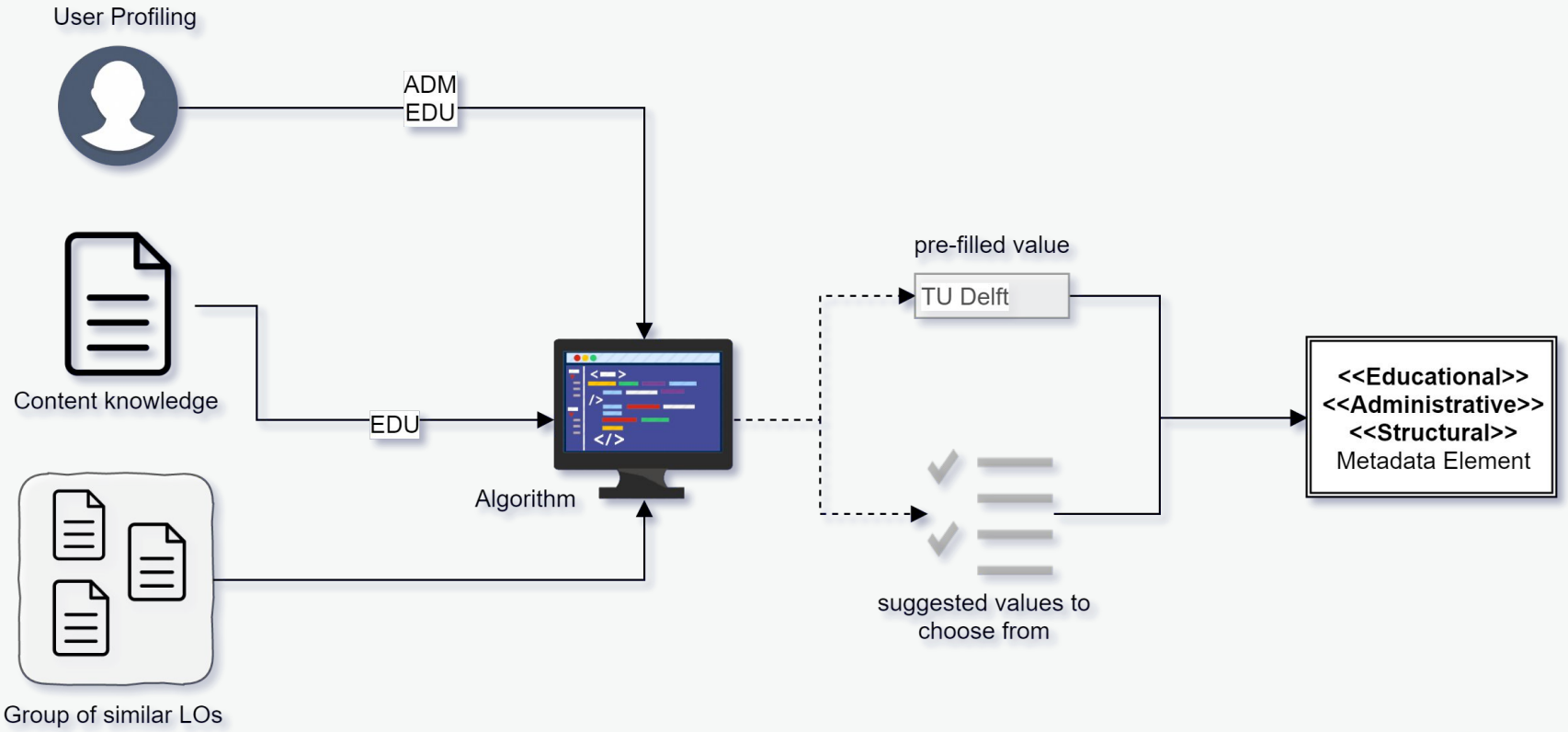
Design patterns

Seen in architecture and Software Engineering

- Provide solutions to commonly occurring problems
- Abstract enough, e.g. ensure wider range of applicability
- Based on previous experience of practitioners and / or researchers



MI6

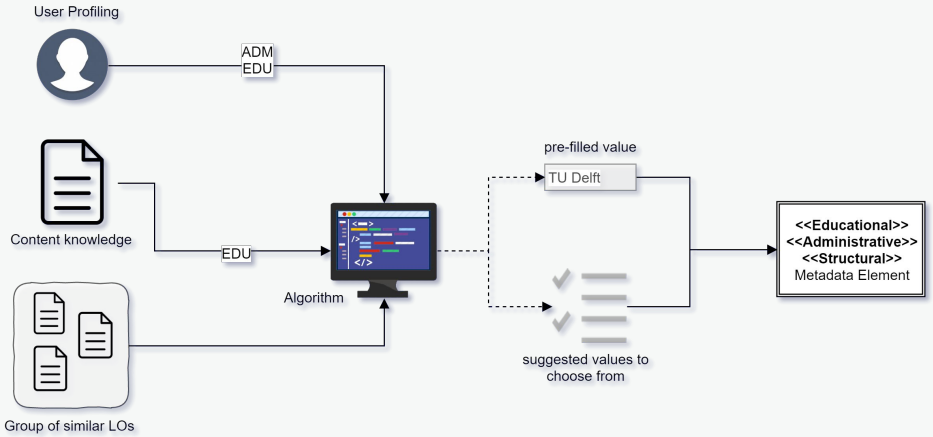

User Profiling
ADM
EDU
Content knowledge
EDU
Algorithm
pre-filled value
TU Delft
suggested values to
choose from
<<Educational>>
<<Administrative>>
<<Structural>>
Metadata Element
Group of similar LOs

MI6

Keep track of user and content

Requirements:

- Information about the author and / or content must be available beforehand
- Expert to decide:
  - how to collect the data
  - to implement a script or an algorithm



Types applicable to:

- Administrative
- Structural
- Educational

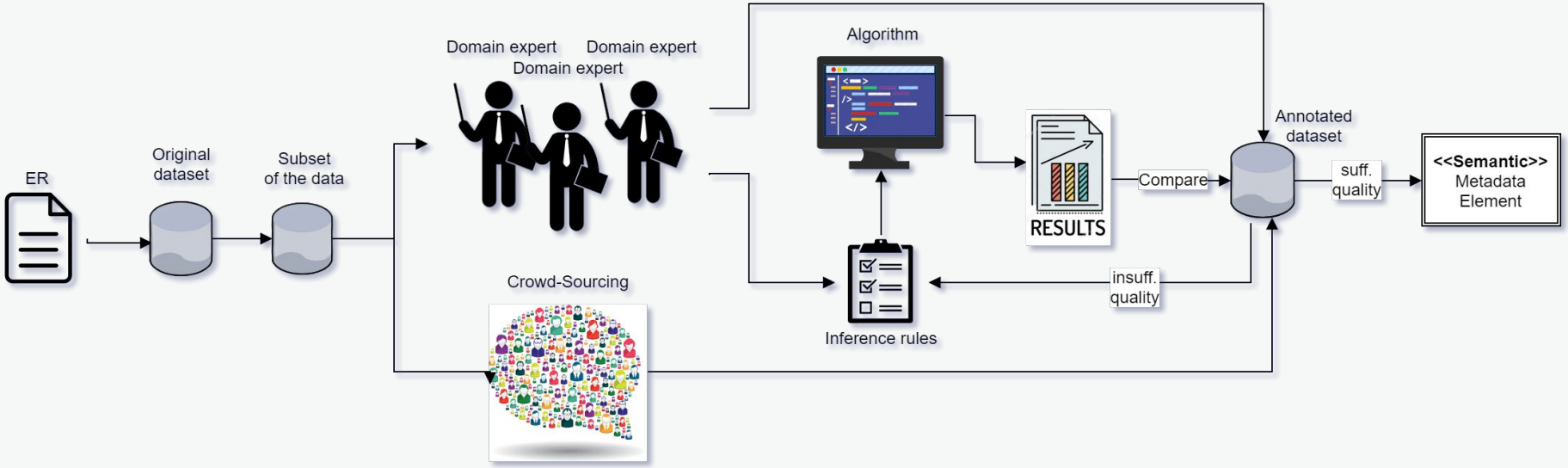
Pros:

+ Saves time and input every time
+ Generally a cheap solution

Cons:

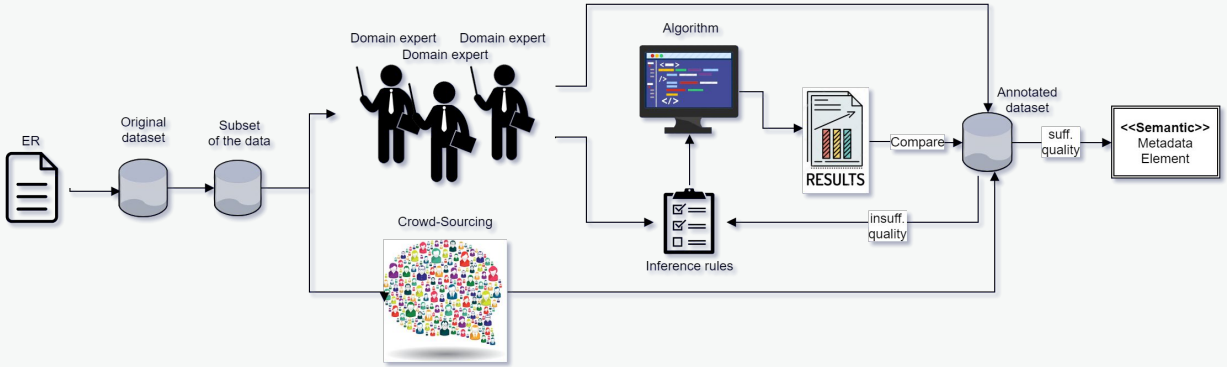
- Potentially wrong information
- Outdated Information

SMRT D@TA


ER
Original dataset
Subset of the data
Domain expert
Domain expert
Domain expert
Algorithm
Crowd-Sourcing
Inference rules
RESULTS
Compare
Annotated dataset
suff. quality
insuff. quality
<<Semantic>>
Metadata Element

SMRT D@TA

Smart extraction of data without specific knowledge



Requirements:

- Analytical thinking
- A developer
- Annotated dataset (optionally)

Types applicable to: Semantic

- Difficulty
- Didactic Intent
- Semantic Density

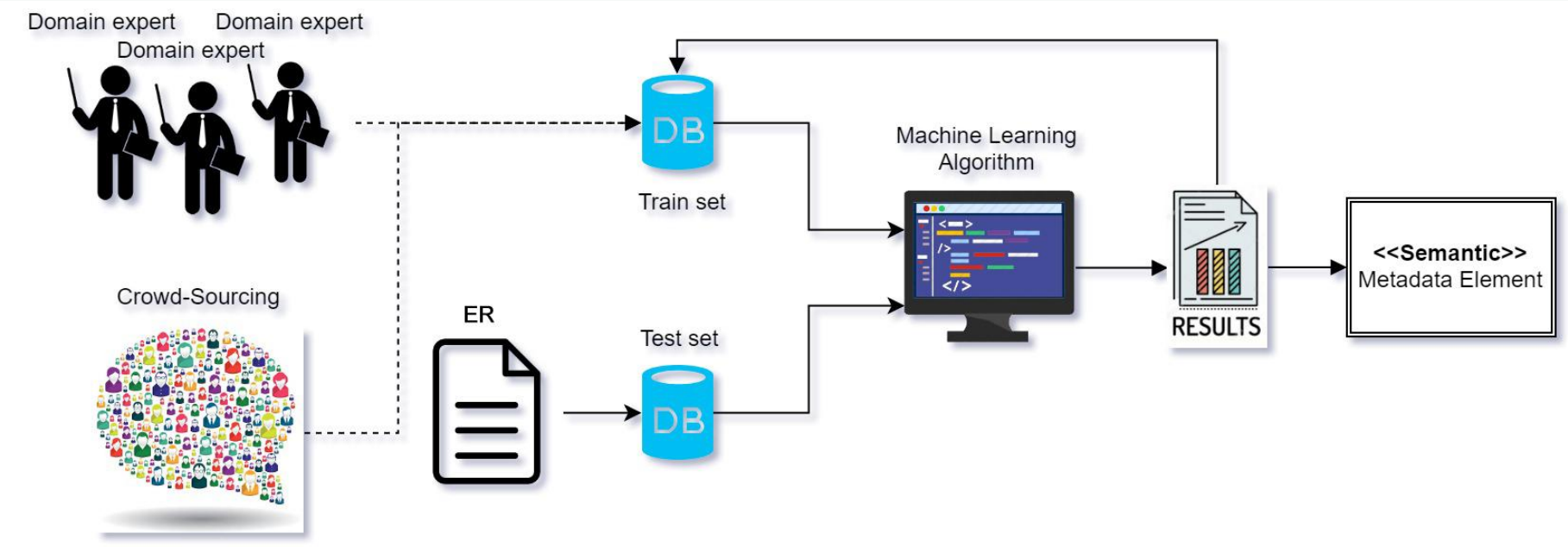
Pros:

+ No complex algorithms
+ No technical knowledge is required

Cons:

- Annotating a big dataset can be costly and time-consuming
- Rules are hard to apply to data of different types

SMRT LRN

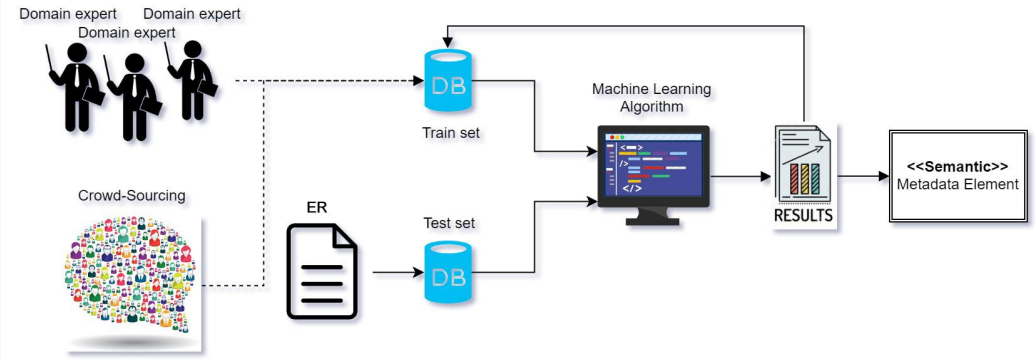

Domain expert
Domain expert
Domain expert
Crowd-Sourcing
ER
DB
Train set
Test set
DB
Machine Learning
Algorithm
RESULTS
<<Semantic>>
Metadata Element

SMRT LRN

Apply ML for smart extraction of metadata

Requirements:

- Skilled expert
- A developer
- Mid-to-Big sized dataset



Types applicable to:

- Semantic
  - Difficulty
  - Didactic Intent
  - Semantic Density

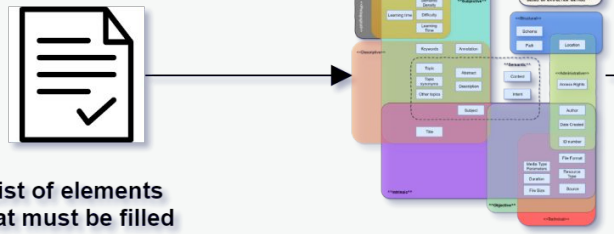
Pros:

+ Overall a low-budget solution
+ Could be applied to different types of ER content data

Cons:

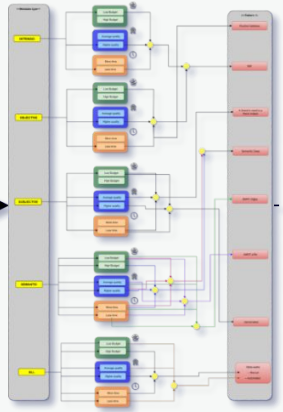
- Obtaining the train dataset can be a timely and costly task

Support users to collect metadata
Process

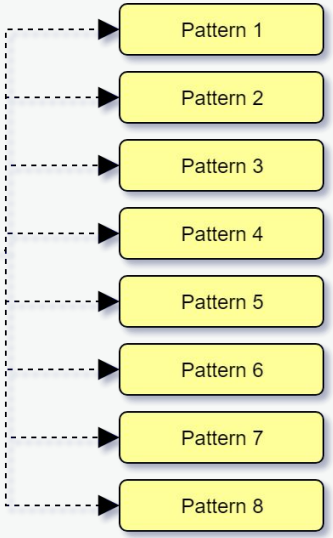


List of elements that must be filled

Find elements and their types according to the type taxonomy

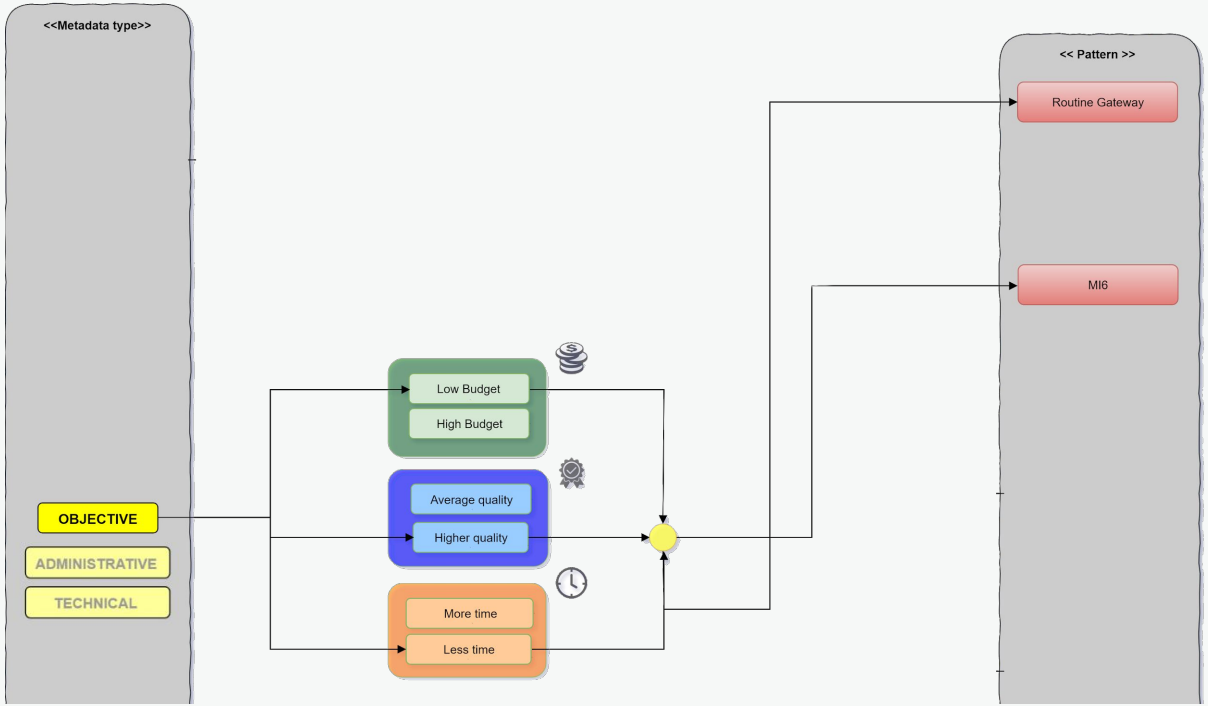


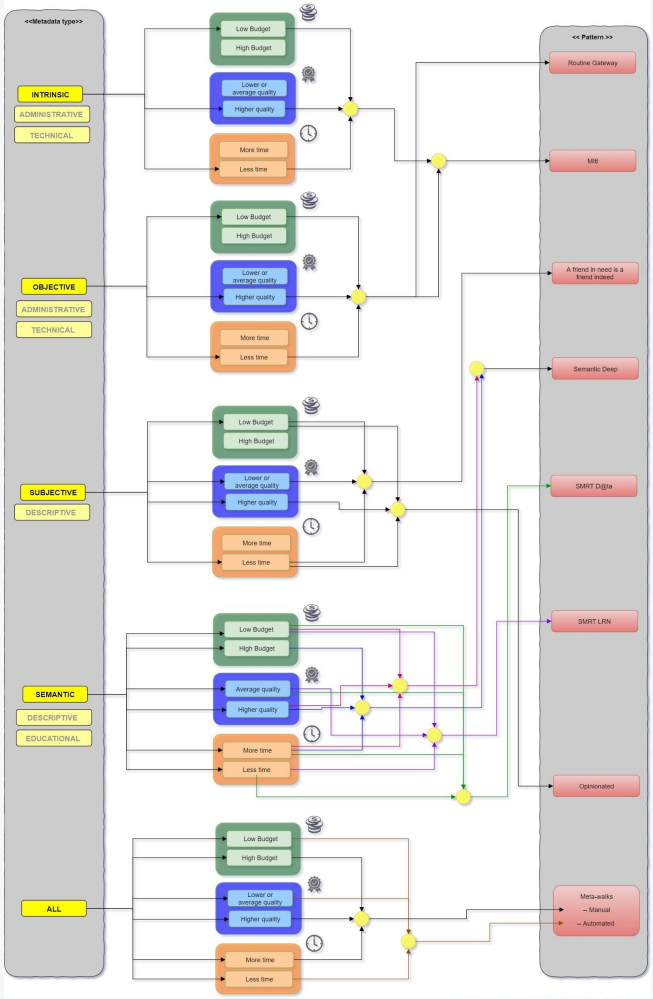
Evaluate requirements and based on the type of metadata find the most suitable pattern



Adopt the pattern accordingly

Decision tree





TU Delft Library
Showcase

Project: Create a federated search tool for OERs
Task: Search into multiple Educational resource repositories
Task: Unify results, rerank them and provide a list of 10 final results

Features:

- Provide resource type (audio, video, pdf, data file etc.)
- Easy modification of a resource, e.g. remove a watermark
- Easy jump to certain part of the resource, e.g. video sequencing by topic
- Incorporate into this their own resources (publications, videos etc)
- Provide suggestions to the user of what other materials are used by users of the current material
- **Split videos by teacher's intention (example, explanation, overview etc.)**

Metadata they **have**:

- Dublin Core
  - Resource type
  - Date
  - Title
  - Access rights

Metadata they **need**:

- Quality
- DOI
- Start/end time
- Topic
- **Didactic Intent**
- User-related data
- Statistical data about the resources

26

Proof of Concept

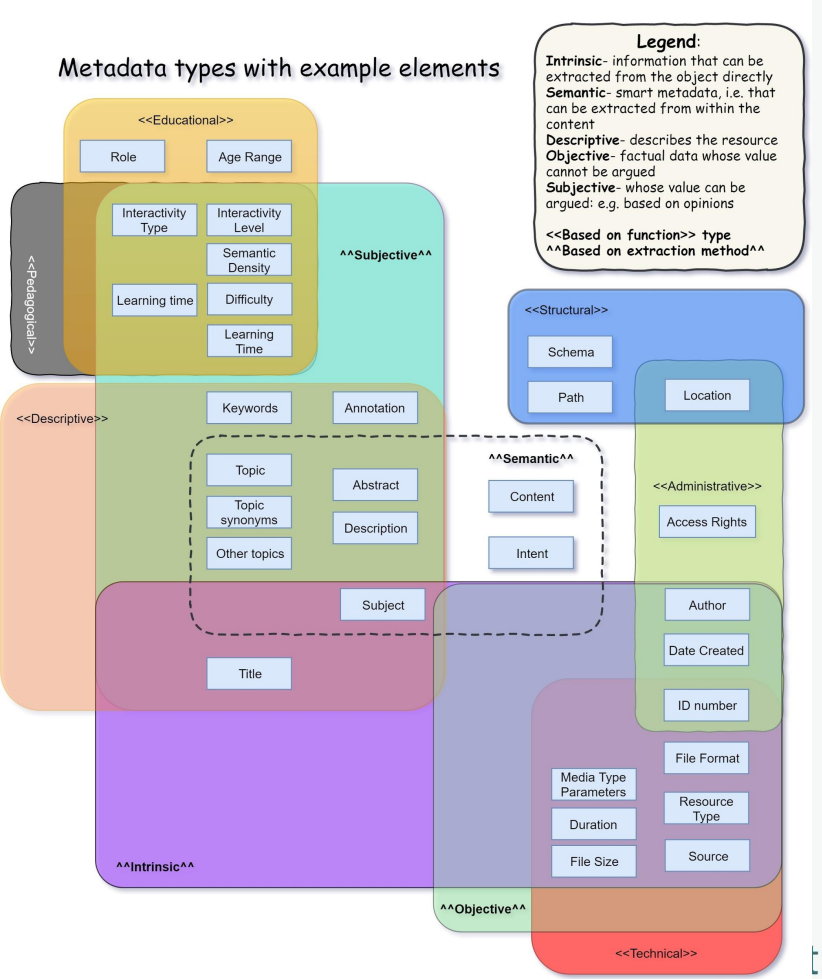
Extract semantic metadata “**Didactic intent**” as per the TU Delft Library showcase

Didactic intent: The intent of the teacher at a certain moment of the lecture

- **(CD) Concept description**: explanation of a concept
- **(CM) Concept mention**: only mention a concept prior to explaining it
  - useful for jumping to a certain moment e.g. in a video
- **(EX) Example**: giving an example in the process of explanation
- **(AP) Application**: giving practical advice or relevant literature for a certain concept
- **(SM) Summary**: giving summary of a concept or overview of the lecture

Proof of Concept

Understand what is the metadata type “Didactic intent”.

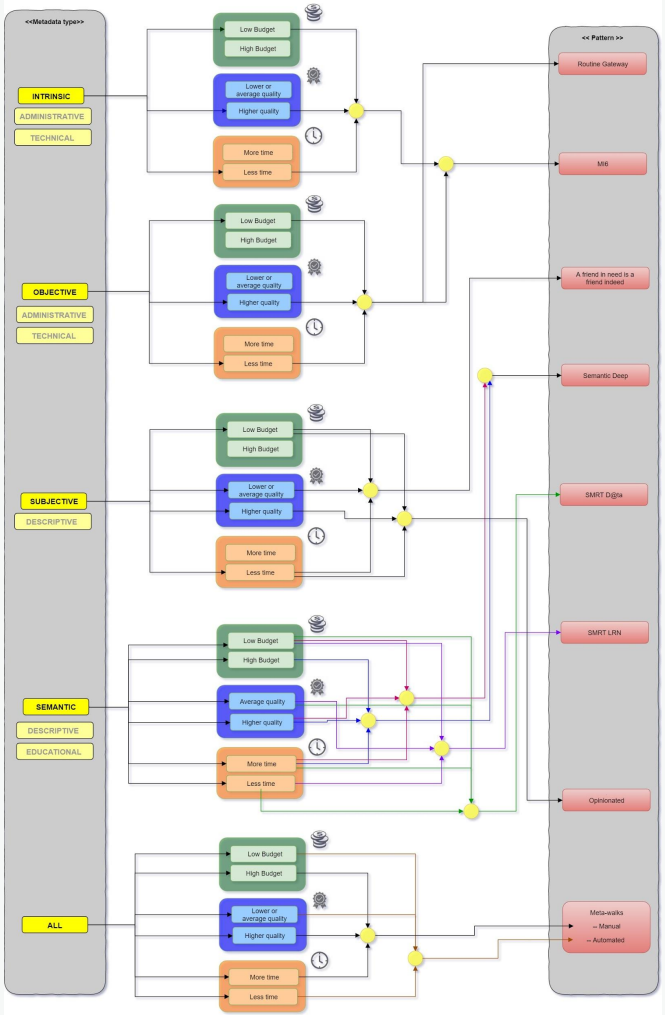


Proof of Concept

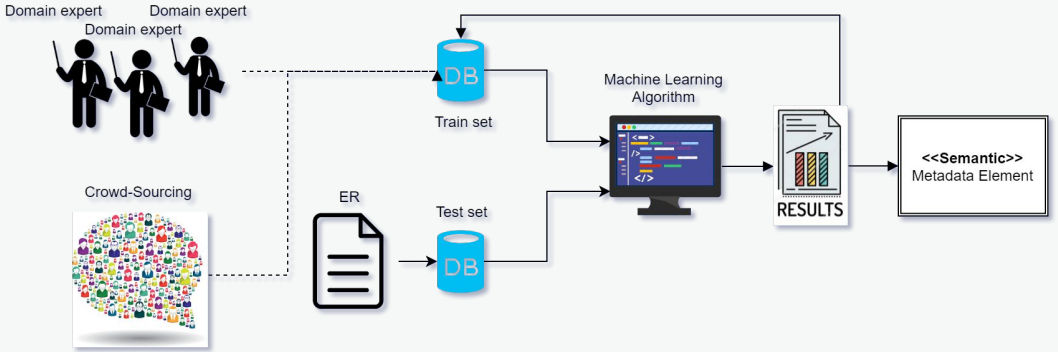
- Low budget
- Average Quality
- Willing to spend more time on the task, but ideally less is better

=>

- SMRT LRN
- SMRT D@TA



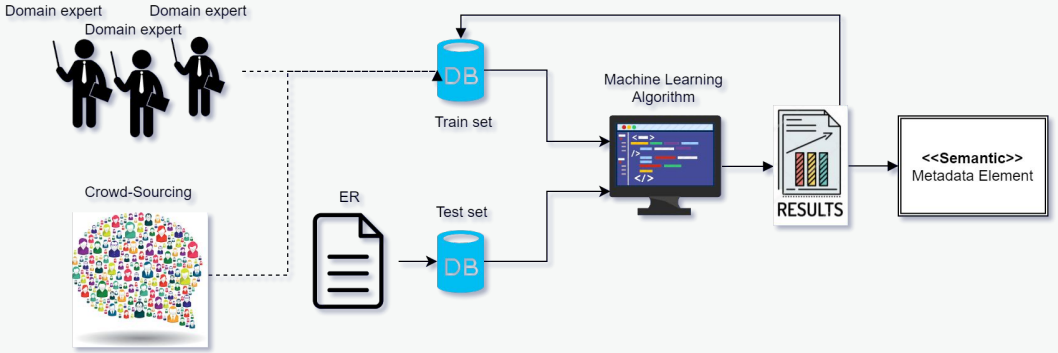
Proof of Concept <SMRT LRN>



Requires:

- Textual data
- (domain) expert(s) or crowdsourcing to label the dataset
- Domain expert to select and implement an algorithm
    - Extraction approach as a classification problem

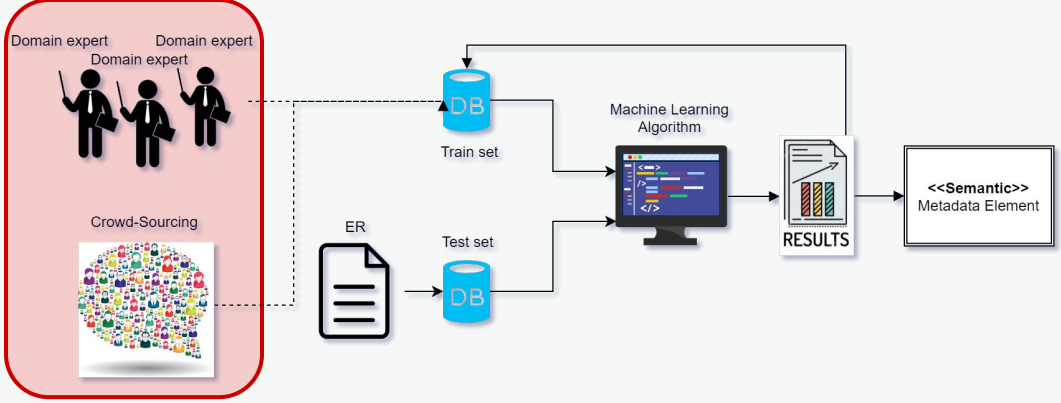
Proof of Concept <SMRT LRN>



Dataset

- MOOC scripts from 11 different courses, such as Data Management, IoT, Mobile Robots, Calculus, Physics, Spacecraft dynamic etc.
  - All extracted from videos, e.g. spoken language text
  - Assumption: the domain doesn't matter due to the nature of our classes - non-topic related
- Total of 556 files, or a total of *38482 sentences*
  - We took a subset of it to create a ground truth: 1/4th of the sentences
    - around *9000 sentences*

Proof of Concept <SMRT LRN>



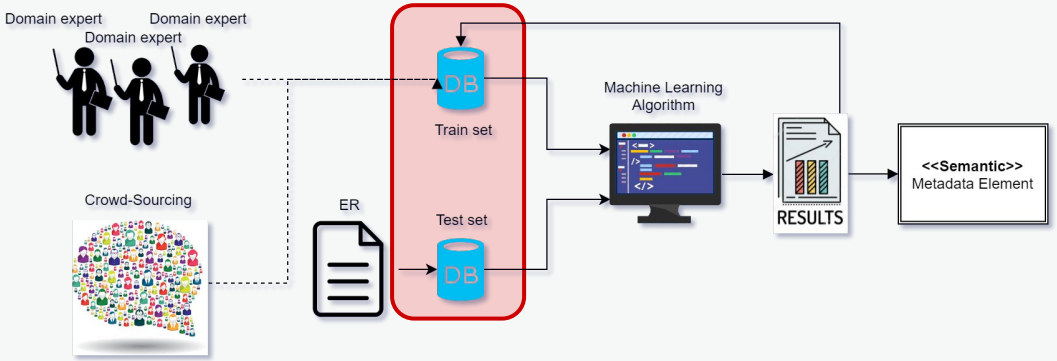
Data annotation:

- Performed by human annotators over all ~9000 sentences
- One out of 5 classes is assigned to each sentence

Data pre-processing:

- Stop-word removal
- Punctuation removal
- Functions & other unimportant things removed
- Tokenization
- Stemming

Proof of Concept <SMRT LRN>



Algorithm Implementation

- Naive Bayes classifier

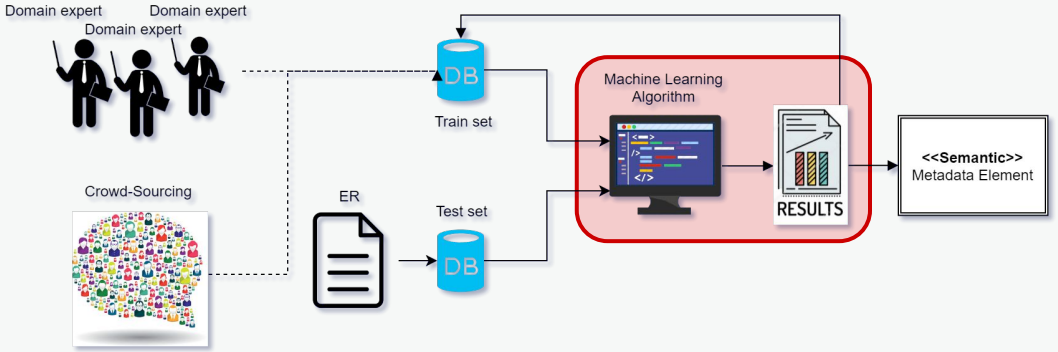
- Train vs test set split 80:20 %
- Random selection
- Running 10 times

Data further cut down to 250 instances per class and model is trained again.

	Sentence	
Label	count	unique
AP	250	250
CD	250	248
CM	250	250
EX	250	250
SM	250	250

Naive Bayes

Proof of Concept <SMRT LRN>



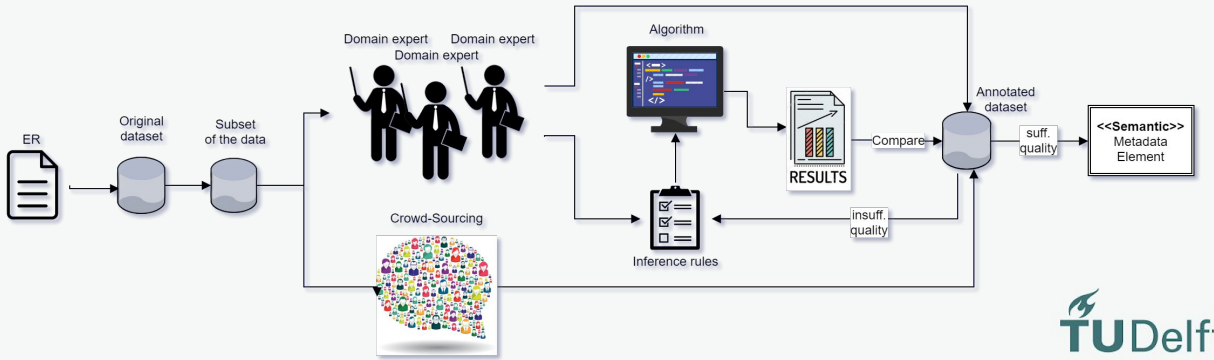
Results:

Class	Precision	Recall	F1	Accuracy
AVG	0.85	0.84	0.84	0.94

Proof of Concept
<SMRT D@TA>

Requires:

- Textual data ✔
- (domain) expert(s) to analyse the content
- (domain) expert(s) to outline rules for extraction
- Developer to implement an algorithm
- Crowdsources or manually annotated dataset to evaluate results against it ✔



Proof of Concept
<SMRT D@TA>

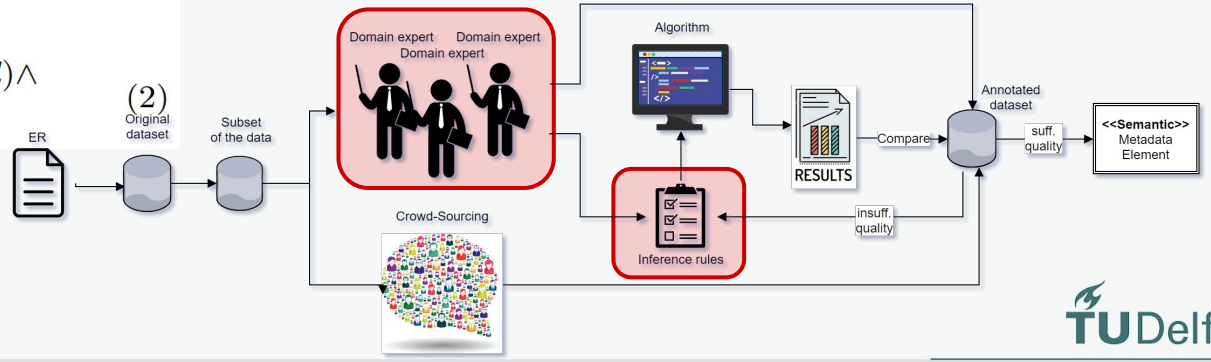
Analysis:

- For grammar rules that could put a sentence into specific class
- Term co-occurrence

Defining inference rules, combining both analysis

$$C \leftarrow d, (T_1 \in d) \wedge \neg(TN_1 \in d \vee TN_2 \in d \vee \ldots TN_n \in d) \quad (1)$$

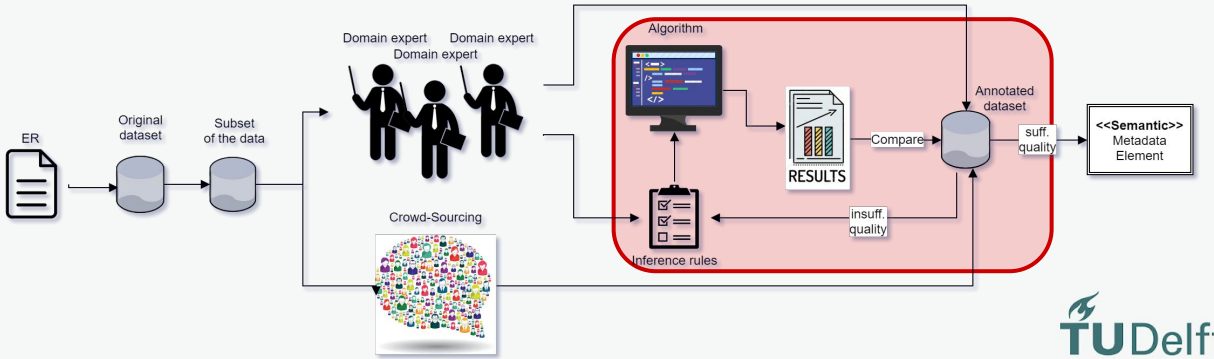
$$C \leftarrow d, (T_1 \in d \vee \cdots \vee T_n \in d) \wedge (GT_1 \in d \vee GT_2 \in d) \quad (2)$$



Proof of Concept
<SMRT D@TA>

Implement and run
Obtain results by comparing extracted with labelled classes:

Class	Precision	Recall	F1	Accuracy
AVG	0.42	0.43	0.38	0.86



Conclusions

- Clearly visible gap
- With the help of the taxonomy, decision tree and patterns, gap can be reduced
- Patterns:
  - are easy to follow and apply
  - don’t limit the specific decisions taken by the user
  - provide a strong starting point for the user
- Combining the patterns, we can extract metadata for the most useful types
- Improve metadata extraction => improve metadata completeness
- Metadata completeness improves ER findability

Future work

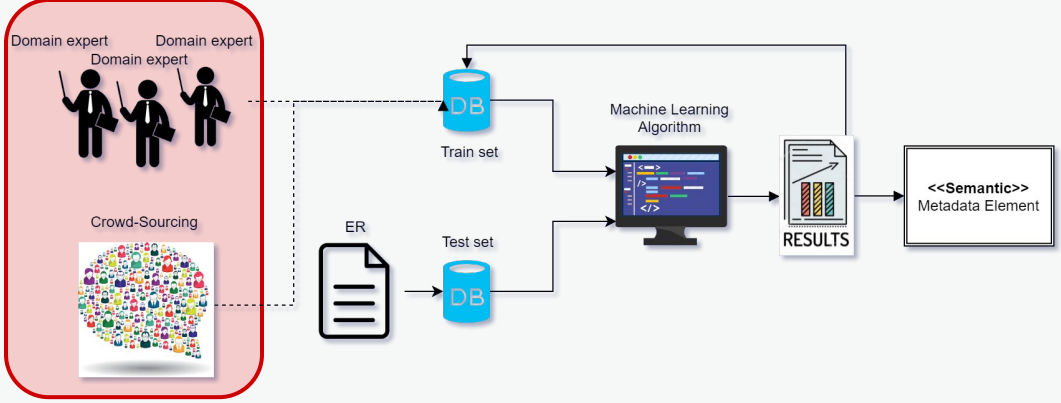
- The type taxonomy can be expanded to include more metadata types

- Perform a more extensive study among practitioners about pattern applicability


Thank you
for your
attention!
Web
Information
Systems
TU Delft
Delft
University of
Technology

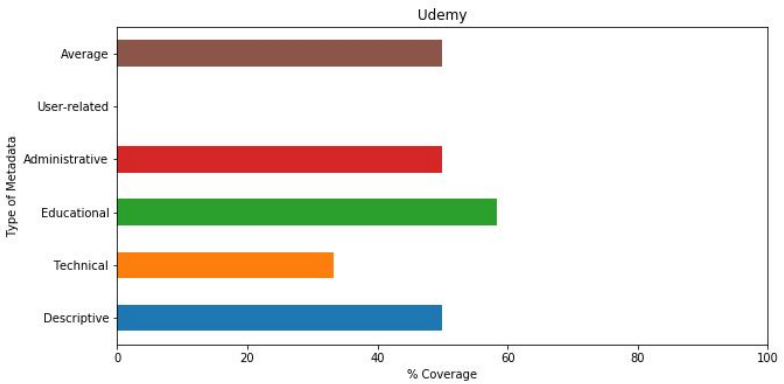
Appendix

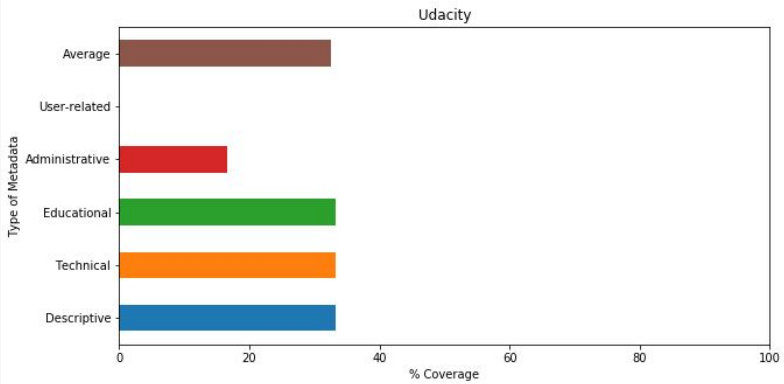
Proof of Concept <SMRT LRN>

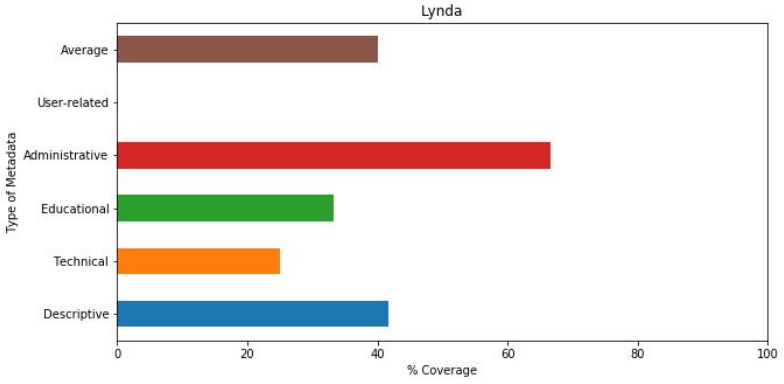


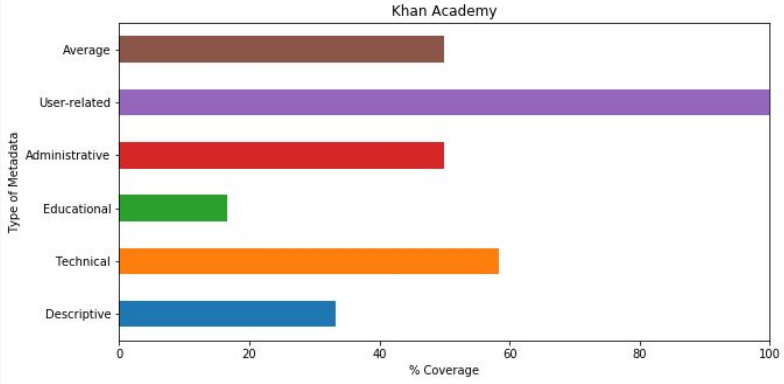
(Labels) Classes are:

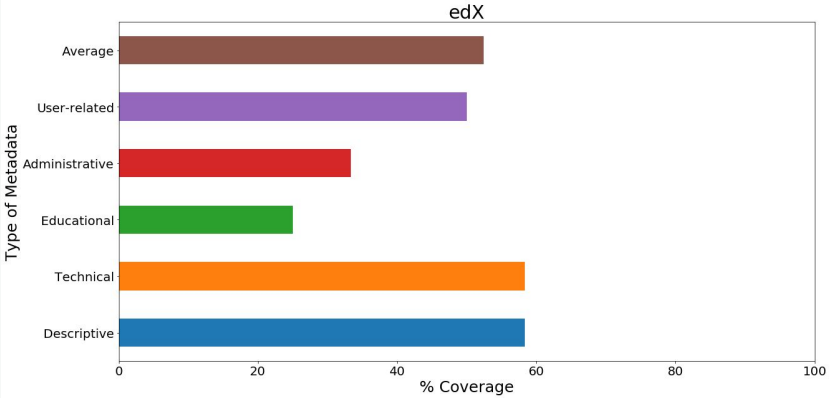
- **(CM) Concept mention: Explanation of the Main concept(s) of the learning object (LO)**
  - CM | this sort of object is called a differential
- **(CD) Concept description: Concept mentioned and explained or not explained right after.**
  - CD | now,the reason that we care about the robot's frame of reference,is because....
- **(EX) Example: Concept example**
  - EX | for example,let's work out the derivative of this product,the product of number and number
- **(AP) Application: Practical advice for the concept**
  - AP | but by following data management best practices throughout the research life cycle...
- **(SM) Summary / Overview: Summary of the previous lecture, of the next lecture or of the concept of the current lecture**
  - SM | last time,we derived the equations of motion of a rigid body
  - SM | let's summarize the situation

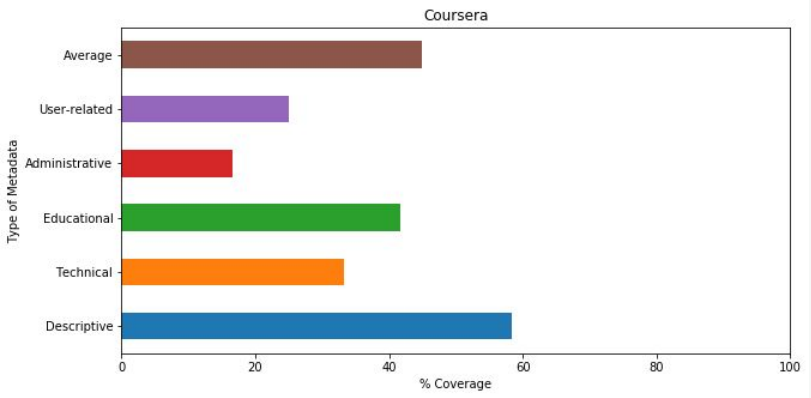

Udemy
Average
User-related
Administrative
Educational
Technical
Descriptive
Type of Metadata
0
20
40
60
80
100
% Coverage

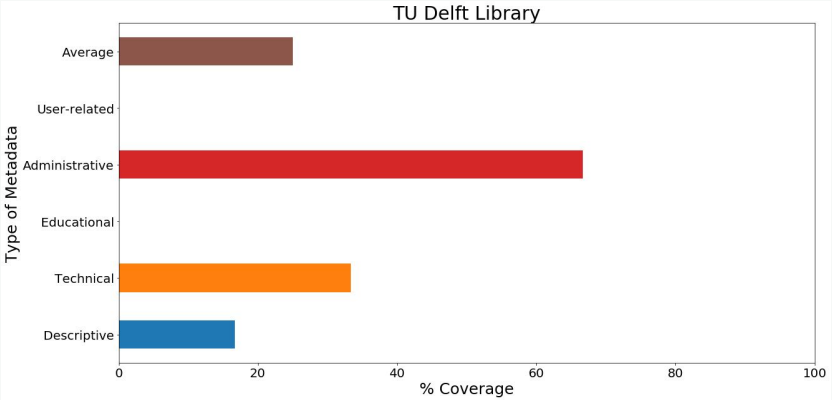

Udacity
Average
User-related
Administrative
Educational
Technical
Descriptive
Type of Metadata
0
20
40
60
80
100
% Coverage


Lynda
Average
User-related
Administrative
Educational
Technical
Descriptive
Type of Metadata
0
20
40
60
80
100
% Coverage


Khan Academy
Average
User-related
Administrative
Educational
Technical
Descriptive
Type of Metadata
0
20
40
60
80
100
% Coverage


edX
Average
User-related
Administrative
Educational
Technical
Descriptive
Type of Metadata
0
20
40
60
80
100
% Coverage


Coursera
Average
User-related
Administrative
Educational
Technical
Descriptive
Type of Metadata
0
20
40
60
80
100
% Coverage

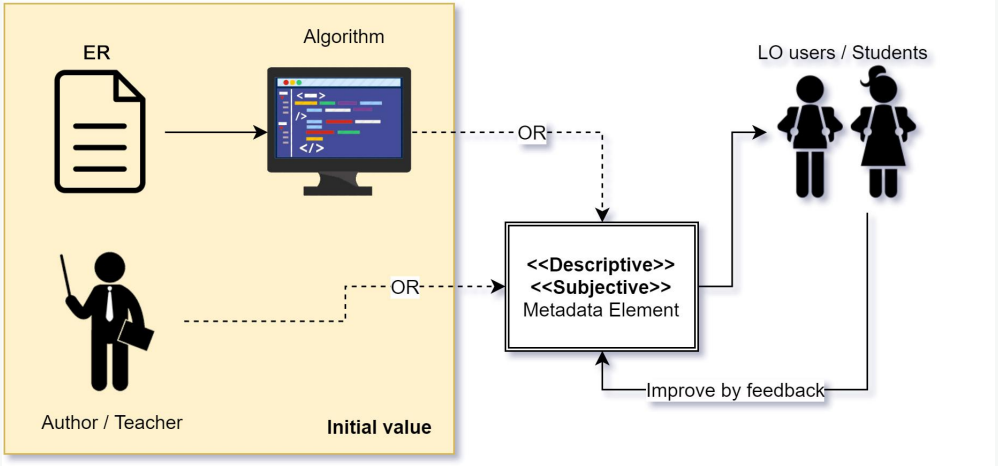

TU Delft Library
Average
User-related
Administrative
Educational
Technical
Descriptive
Type of Metadata
0
20
40
60
80
100
% Coverage

A friend in need

Improve objectivity of the ER

Requirements:

- Initial value supplied
- Expert to implement feedback system
- Learners must be given opportunity to give feedback



Types applicable to:

- Subjective
- Descriptive
- Educational

Pros:

+ easy to implement
+ cheap
+ keeps the ER up to date and unbiased

Cons:

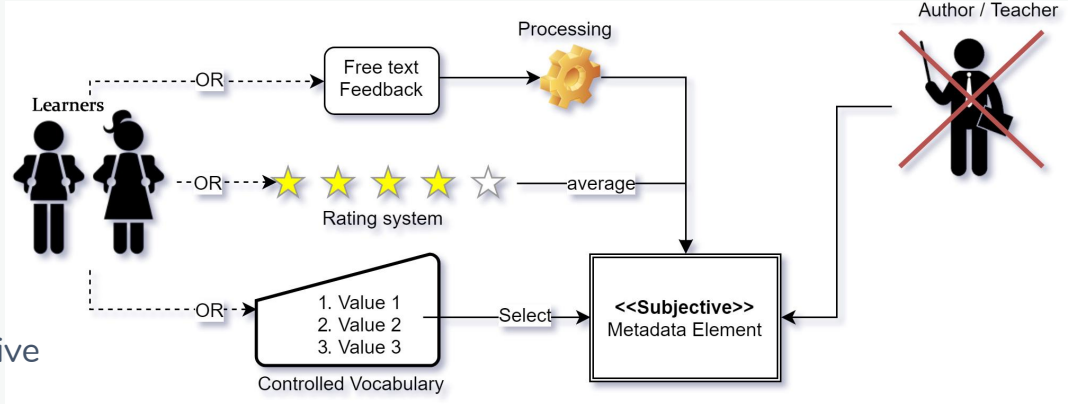
- could be time consuming for teacher

Opinionated

Improve objectivity of the ER

Requirements:

- Expert to implement feedback system
- Learners must be given opportunity to give opinion
  - free text
  - rating system
  - controlled vocabulary



Types applicable to:

- Subjective
- Descriptive
- Educational

Pros:

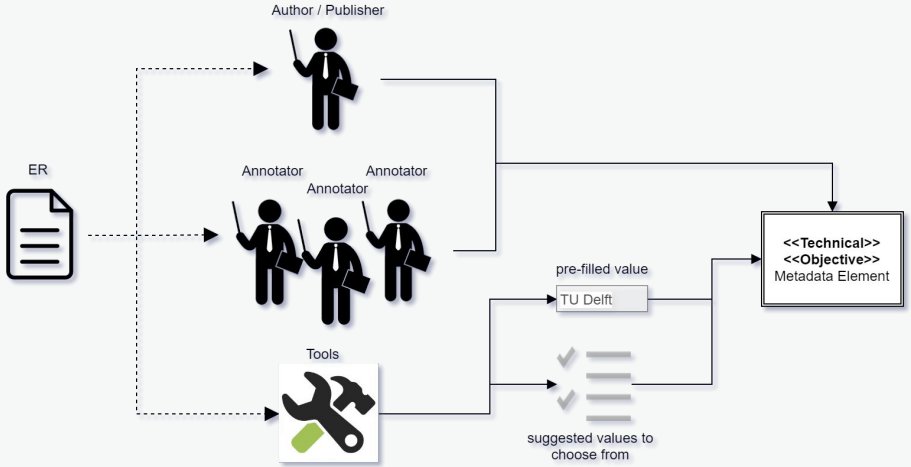
+ easy to implement
+ cheap
+ keeps the ER up to date and unbiased
+ No initial value needed

Routine Gateway

Quick and simple extraction

Requirements:

- Directly extractable from the resource entity
- Data value is objective
- Expert to implement a solution



Types applicable to:

- Technical
- Objective

Pros:

+ very easy to implement (algorithmically)
+ scalable

Cons:

- manual input is tedious, error-prone or demotivating
- annotators group grows in number with the growth of ERs

Semantic Deep

Make more sense of metadata with complex hierarchical relationships of data

Requirements:

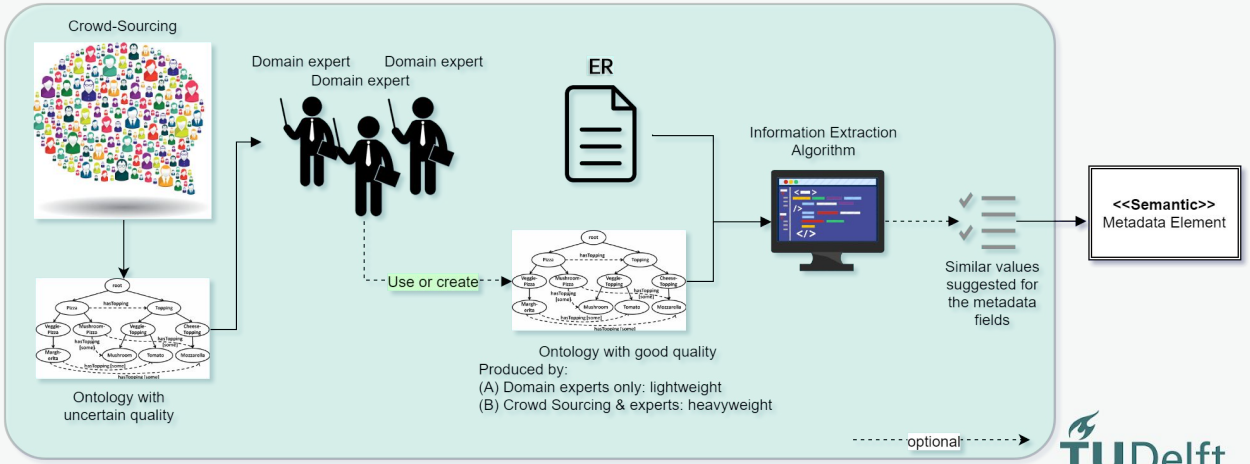
- Suitable ontology
- Ontology and domain expert(s)
- Willing to spend more time on the task
- Potentially higher budget

Types applicable to: Semantic

- Topic
- Prerequisite Knowledge

Pros:
+ can significantly improve metadata values

Cons:
- suitable ontology often lacks
- time-consuming to create an ontology

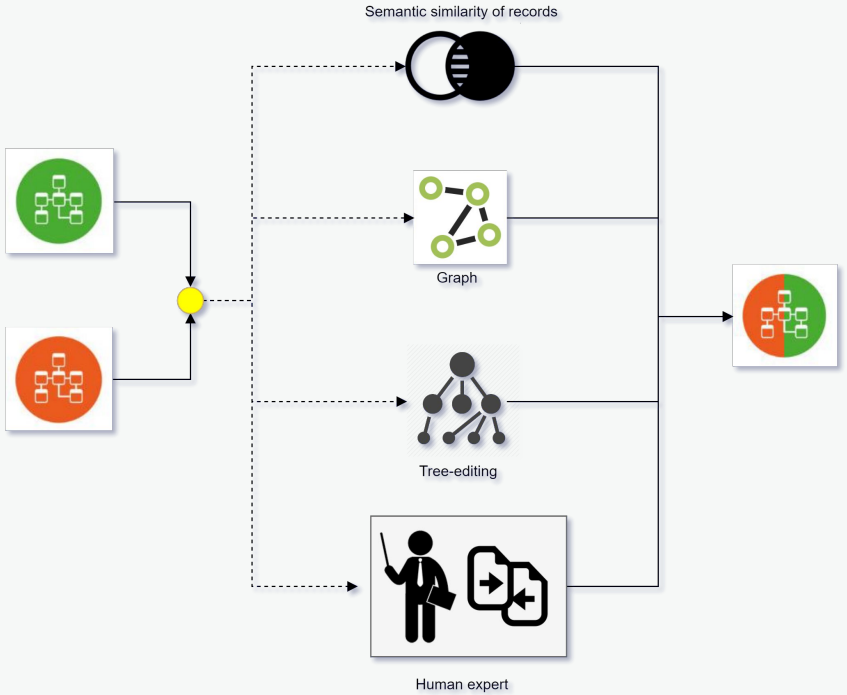


Meta-walk

Map metadata values from different schemas

Requirements:

- XML structure of the metadata
- Knowledge and experience with metadata
- Expert for metadata



Pros:

+ no need to extract everything from scratch
+ one-time implementation

Cons:

- could introduce high time-complexity with bigger schemas
- manual approach could be error-prone