

# Measuring the Quality of an Integrated Schema

**Fabien Duchateau, Zohra Bellahsene**



ER 2010, Vancouver

# Outline

- 1 Introduction
  - Context
  - Motivations
  - Contributions
- 2 Quality Metrics
  - Overview
  - Completeness
  - Minimality
  - Structurality
  - Schema Proximity
- 3 Experiments
- 4 Conclusion

# Introduction

Schema integration is a central task for data integration

- discovering correspondences/mappings between input schemas
- merging input schemas into an integrated schema based on discovered mappings
- using this integrated schema as a uniform interface for querying

Mappings quality is computed with popular metrics (precision, recall, F-measure) but we lack metrics for evaluating the quality of an integrated schema

# Running Example

Two libraries decide to fusion their catalogs of media

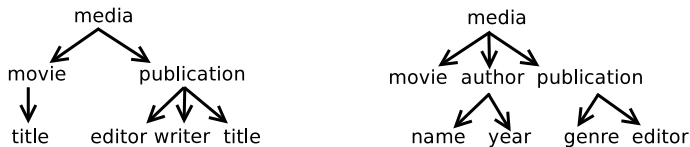


Figure: Schemas used by the two libraries to query their catalog

# Running Example

Two libraries decide to fusion their catalogs of media

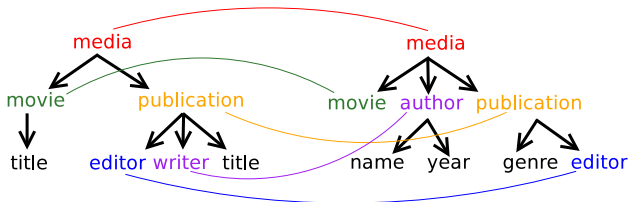


Figure: Mappings between the two library schemas

# Running Example

Two libraries decide to fusion their catalogs of media

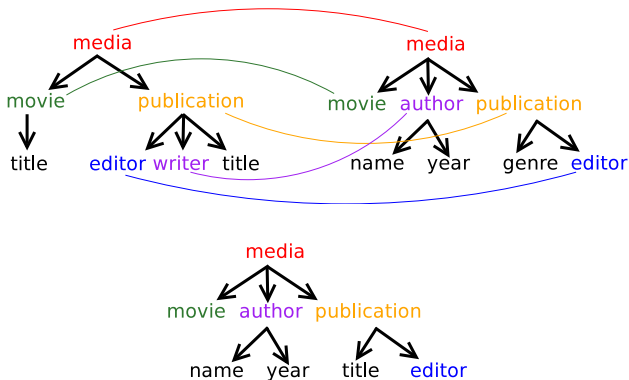


Figure: A possible integrated schema

# Running Example

Two libraries decide to fusion their catalogs of media

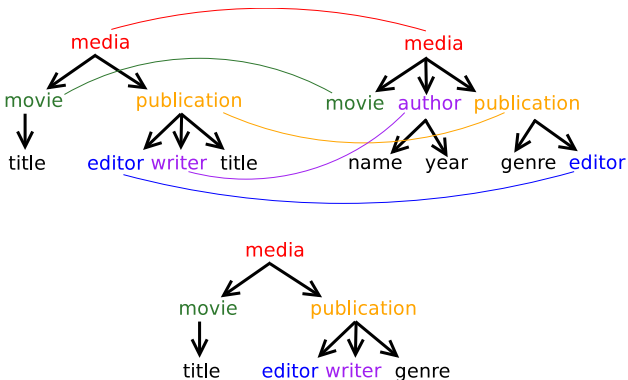


Figure: Another possible integrated schema

# Motivations

Integrated schemas strongly depends on the application domain and user needs.

Why evaluating integrated schemas ?

- improve query execution
- if several integrated schemas have been generated, metrics could help users to select the most suitable one
- estimate the cost of a full integration process
- when manual evaluation is not possible (e.g., in dynamic or large scale scenarios)



# Contributions

In this context, we propose to evaluate the quality of integrated schemas by:

- extending two metrics (**completeness** and **minimality**)
- providing a new metric dealing with structure (**structurality**)
- computing the similarity between two schemas (**schema proximity**)
- analyzing results of these metrics applied to schema matching tools

- 1 Introduction
  - Context
  - Motivations
  - Contributions
- 2 Quality Metrics
  - Overview
  - Completeness
  - Minimality
  - Structurality
  - Schema Proximity
- 3 Experiments
- 4 Conclusion

## Overview (1/2)

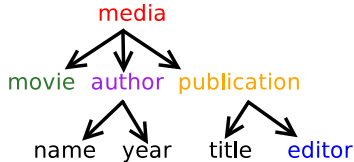
A few metrics defined between two schemas [dCMBS07]. In our context, we have a reference integrated schema

The reference integrated schema can be:

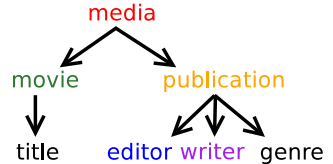
- provided by an expert
- a global repository / common vocabulary
- one of the input schemas

Evaluating the quality of an integrated schema produced by a tool against the reference integrated schema means that we assess how similar the tool schema is w.r.t. the reference schema

## Overview (2/2)



**Schema Generated by a Tool**



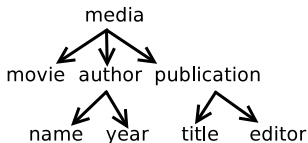
**Reference Schema**

## Completeness (1/2)

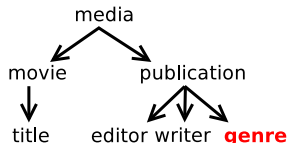
**Completeness** checks that all elements in the reference integrated schema are covered by the tool integrated schema

Completeness is in the range  $[0, 1]$ , with a 1 value meaning that the tool integrated schema include all elements present in the reference integrated schema

## Completeness (2/2)



**Schema Generated by a Tool**



**Reference Schema**

$$\text{comp}(S_{\text{tool}}, S_{\text{ref}}) = 0.86$$

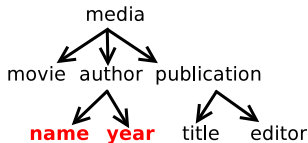
The tool integrated schema lacks one element (*genre*) according to the reference integrated schema

## Minimality (1/2)

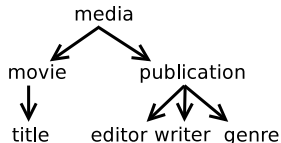
**Minimality** checks that no redundant or extra element appears in the integrated schema

Minimality is in the range  $[0, 1]$ , with a 1 value meaning that the tool integrated schema does not include extra-elements related to reference integrated schema

## Minimality (2/2)



**Schema Generated by a Tool**



**Reference Schema**

$$\min(S_{tool}, S_{ref}) = 0.71$$

The tool integrated schema has two extra-elements (*name* and *year*) w.r.t. the reference integrated schema.



## Structurality (1/4)

Structurality denotes “*the qualities of the structure an object possesses*”

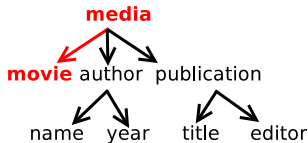
Why structure is important for integrated schemas ?

- relationships of schemas may not have semantics, thus their implicit structure encompasses this semantic
- users are accustomed to query a specific schema, thus they might prefer an integrated schema with a similar structure

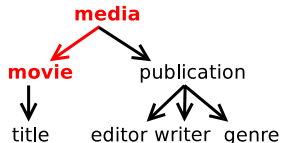
Intuition: an element in both integrated schemas shares a maximum number of common ancestors, and no extra ancestor have been added in the tool integrated schema.

# Structurality (2/4)

## Element movie:



**Schema Generated by a Tool**

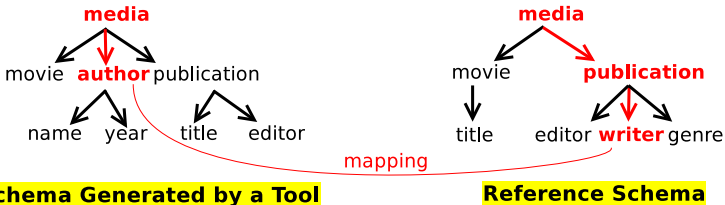


**Reference Schema**

$Ancestors_{tool} = \{media\}$  and  $Ancestors_{ref} = \{media\}$   
 $structElem(movie) = 1$

# Structurality (2/4)

## Element writer:



$$Ancestors_{tool} = \{media\} \text{ and } Ancestors_{ref} = \{media, publication\}$$

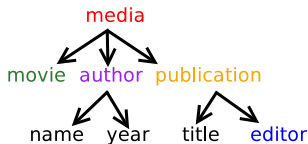
$$structElem(writer) = \frac{1}{2}$$

## Structurality (3/4)

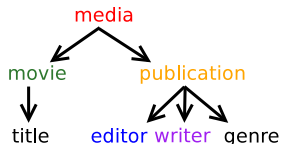
**Structurality** is the sum of all element structuralities (except for the root element) divided by this number of elements

The root element is excluded because of its strong weight (the whole structurality is already rewarded or penalized since the root appears or not in all element structuralities)

# Structurality (4/4)



**Schema Generated by a Tool**



**Reference Schema**

$$\text{struct}(S_{\text{tool}}, S_{\text{ref}}) = 0.625$$

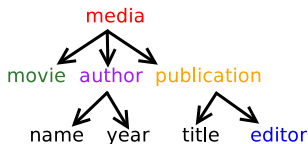
Half of the elements are correctly placed, one is missing (*genre*) and two are misplaced (*author/writer* and *title*)

## Schema Proximity (1/2)

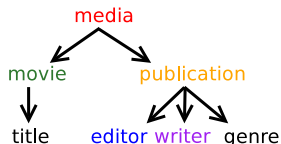
**Schema proximity** computes the similarity between two integrated schemas

It is a weighted average of completeness, minimality and structurality

## Schema Proximity (2/2)



**Schema Generated by a Tool**



**Reference Schema**

$$\text{prox}(S_{\text{tool}}, S_{\text{ref}}) = \frac{0.86 + 0.71 + 0.625}{3} = 0.73$$

The tool integrated schema is 73% similar to the reference integrated schema

- 1 Introduction
  - Context
  - Motivations
  - Contributions
- 2 Quality Metrics
  - Overview
  - Completeness
  - Minimality
  - Structurality
  - Schema Proximity
- 3 Experiments
- 4 Conclusion



# Experiment Protocol

**Schema matching tools:** COMA++ [ADMR05] and Rondo (Similarity Flooding) [MGMR02]

**Datasets:** domain experts have generated a reference integrated schema (and a reference set of mappings)

**Evaluation:** We run the matching tools to discover mappings. These mappings are not (in)validated. Then, the tools use these mappings to produce an integrated schema.

## Experiments Report (1/2)

Completeness is slightly affected by the mapping quality

If a correct mapping is missed, all elements of this mapping are added into the integrated schema => no impact for completeness

Minimality is strongly correlated to recall

A correct mapping not discovered by the tool => redundancies in the integrated schema

Structurality strongly depends on the matching tool's algorithm

Incorrect mappings discovered by the tool => misplaced elements in the integrated schema

## Experiments Report (2/2)

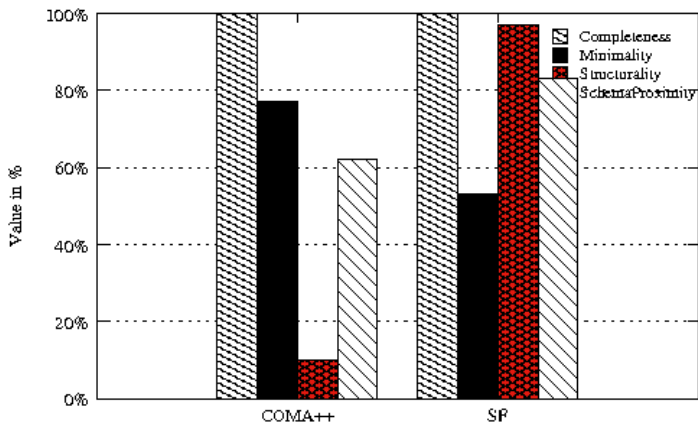


Figure: Dataset about currency web services: high completeness, average minimality and variable structurality

# Conclusion




We have presented:

- new metrics for evaluating integrated schemas, including their structure
- completeness is slightly affected by mapping quality contrary to minimality
- structurality depends on the matching tool algorithm and on mapping quality

Future work:

- extend the metrics to ontologies (w.r.t relationships such as *is-a* or *instance-of*)
- measure the impact of structurality for query execution

# Thank you !

-  David Aumueller, Hong Hai Do, Sabine Massmann, and Erhard Rahm.  
Schema and ontology matching with COMA++.  
In *ACM SIGMOD*, pages 906–908, 2005.
-  Maria da Conceição Moraes Batista and Ana Carolina Salgado.  
Information quality measurement in data integration schemas.  
In *QDB*, pages 61–72, 2007.
-  Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm.  
Similarity flooding: A versatile graph matching algorithm and its application to schema matching.  
In *ICDE*, pages 117–128, 2002.