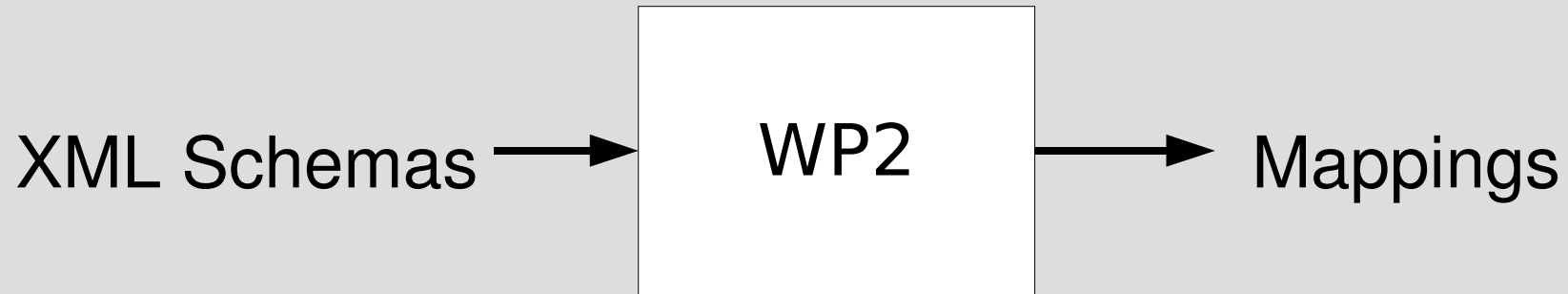


FORUM: WP2

Découverte de Mappings entre
Schémas

WP2 Inputs / Outputs



- Mapping = path.Att1 , path.Att2, relationship
- Relationship
 - equivalence (égalité)
 - hyponymie (inclusion)
 - hyperonymie

BMatch: Inputs / Outputs

XML Schemas



Bmatch:

- terminological and context-based approach
- indexing structure

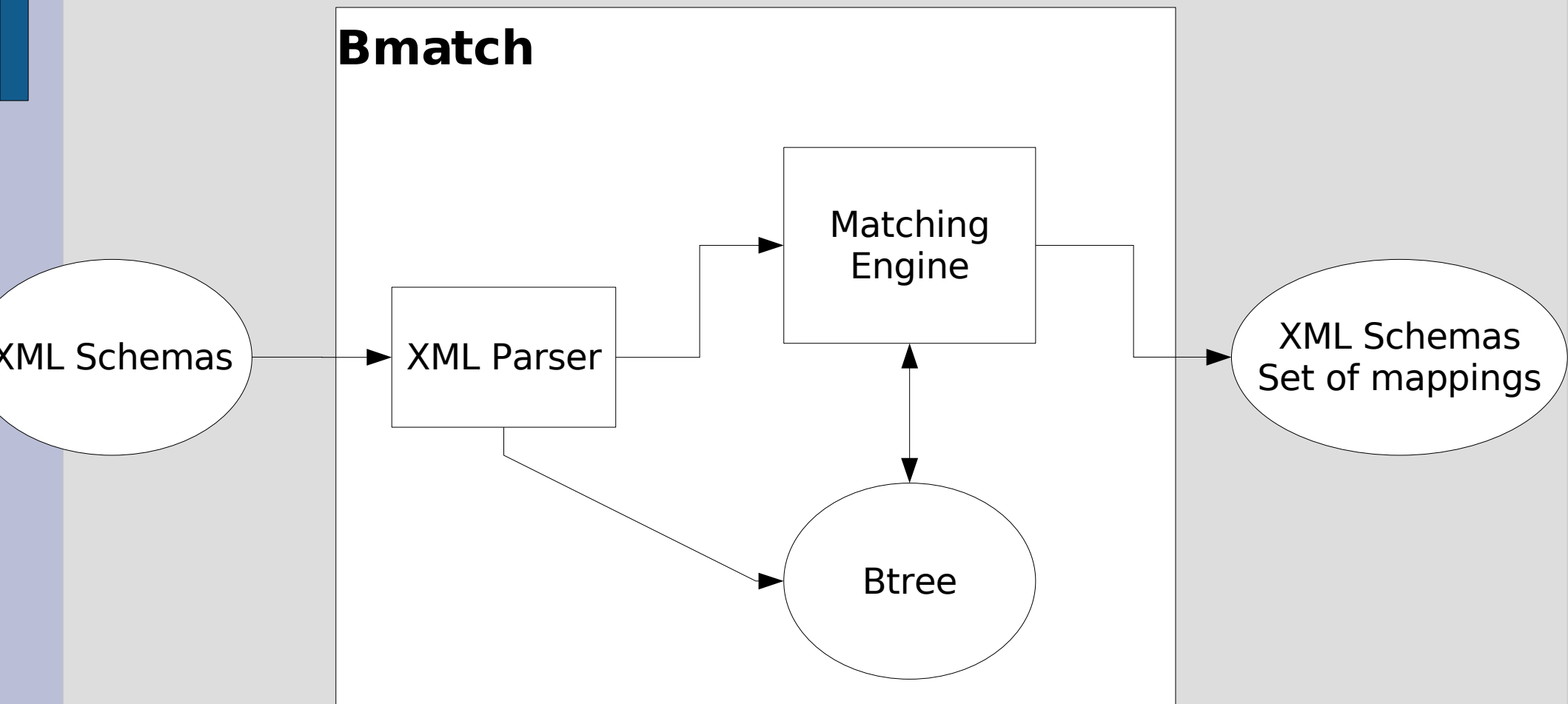


XML Integrated schema or/and set of mappings

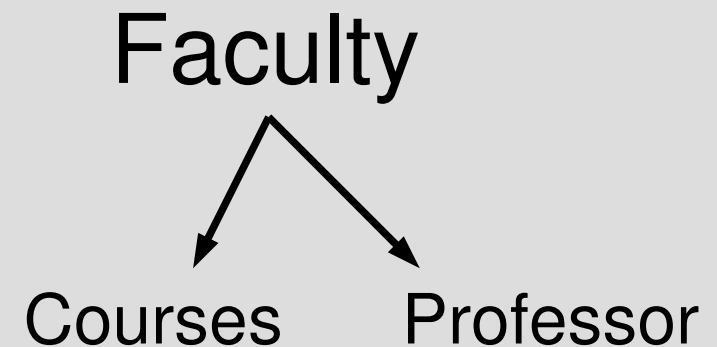
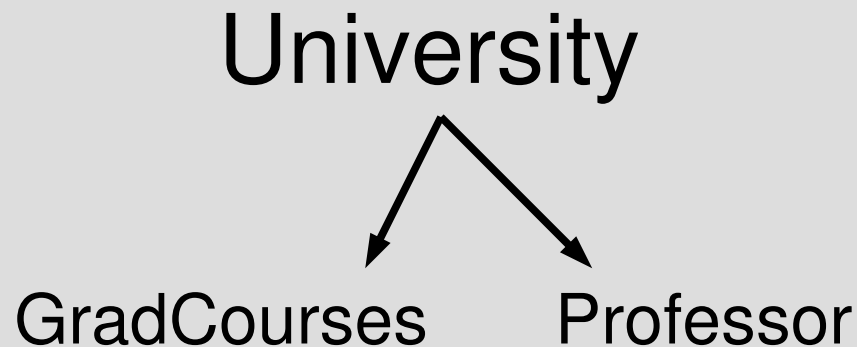
Features

- Semantic aspect (RCIS07)
 - terminological (Levenhstein and 3grams)
 - structural (context based using cosine measure)
- Performance aspect (SMDB-ICDE 07)
 - indexing structure (B-tree)

Architecture



Terminological Example with BMatch

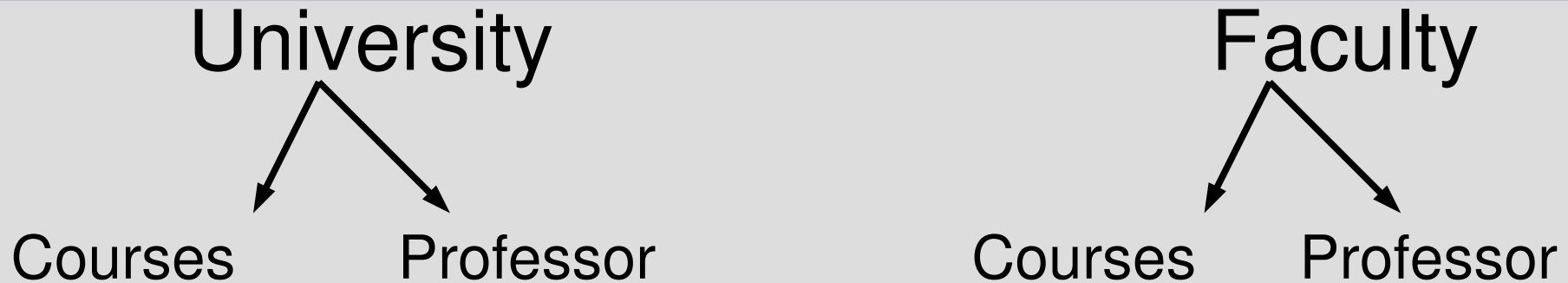


$3\text{grams}(\text{GradCourses}, \text{Courses}) = 0.2$

$\text{Lev}(\text{GradCourses}, \text{Courses}) = 0.42$

$\text{StringMatching}(\text{GradCourses}, \text{Courses}) = 0.31$

Context Example with BMatch



$\text{StringMatching}(\text{Faculty}, \text{University}) = 0.002$

$\text{Context}(\text{University}) = \{\text{University}, \text{Courses}, \text{Professor}\}$

$\text{Context}(\text{Faculty}) = \{\text{Faculty}, \text{Courses}, \text{Professor}\}$

$\text{CosineMeasure}(\text{Context}(\text{University}), \text{Context}(\text{Faculty})) = 0.37$

Semantic Aspect

- Context of node n
 - represents the most important neighbour nodes of n
 - each of them is assigned a weight depending on the relationship with n

$$\omega_1(n_c, n_i) = 1 + \frac{K}{\Delta d + |\text{lev}(n_c) - \text{lev}(n_a)| + |\text{lev}(n_i) - \text{lev}(n_a)|}$$

- String Matching is the average between
 - Levenhstein distance
 - n-grams

A 2-steps Matching Algorithm

- Discovering linguistic similarities :
 - String Matching between 2 node labels.
 - if above a given threshold, replacement of one of the label by the other.
- Cosine Measure using context :
 - due to replacements, the contexts of two nodes can be very similar

Performance Aspect

- B-tree indexing structure
- Based on the fact that most similar labels share a common token
- Tokens are indexed, with references to all labels containing the indexing token

Ongoing Work

- Use of DHT instead of B-tree
- LIRIS work: a combination between BMatch similarity values and Lyes' approach.

