XBenchMatch: a Benchmark for XML Schema Matching Tools

Fabien Duchateau[1], Zohra Bellahsene[1] and Ela Hunt[2]
[1]LIRMM, Univ. Montpellier 2-CNRS, [2]ETH Zurich

duchatea@lirmm.fr, bella@lirmm.fr, hunt@inf.ethz.ch

XBenchMatch uses as input the result of a schema matching algorithm (set of mappings and/or an integrated schema) and generates statistics about the quality of this input and the performance of the matching tool. A demo version of the prototype is available at *http://www.lirmm.fr/duchatea/XBenchMatch*.

GOALS: extensibility, portability, simplicity (ease of use), scalability, genericity, completeness

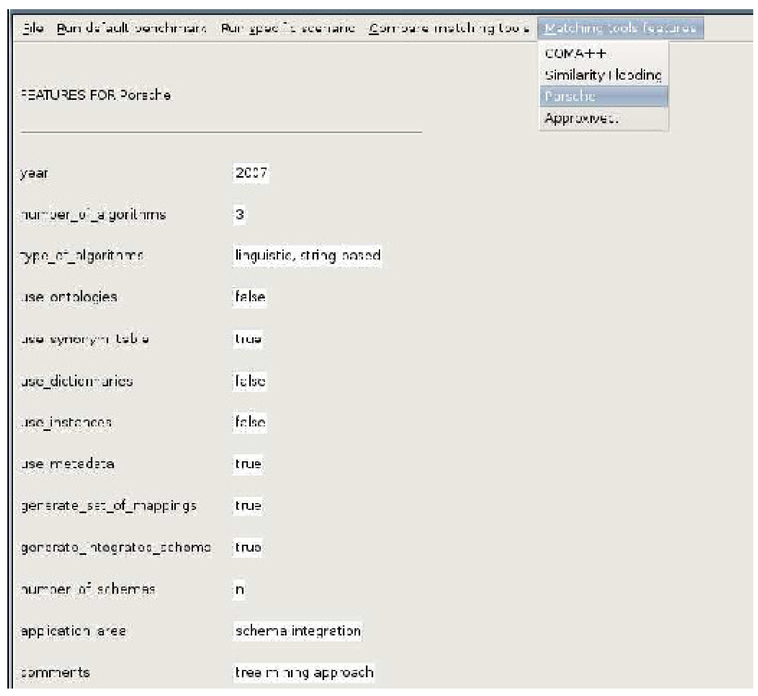
XBenchMatch FEATURES

- **Extensibility.** The benchmark should be able to be extended to include new measures and new format
- **Portability**. The benchmark should be OS-independent,
- **Simplicity.** since both end-users and schema matching experts are targeted by this benchmark tool.
- **Scalability** on two aspects: creating new benchmark scenarii is an easy task. And a benchmark composed of many scenarii should be easy to build and evaluate.
- **Genericity.** It should work with most of the available matchers.

KIND OF EVALUATION

1.Quality of Mappings and Quality of Integrated Schema
- based on the use of the metrics

2.Performance of Matching Algorithms (time).



MAPPING QUALITY MEASURES, ***T**map* are derived mappings, ***T**ex* are expert mappings, expressed as paths:

1. Precision = $|\boldsymbol{T}map \cap \boldsymbol{T}ex| / |\boldsymbol{T}map|$
2. Recall = $|\boldsymbol{T}map \cap \boldsymbol{T}ex| / |\boldsymbol{T}ex|$
3. $F_{measure}$ = (2 · precision · recall) / (precision + recall)

INTEGRATED SCHEMA QUALITY MEASURES
for an integrated schema Si, and an input schema Sg:

1. **Backbone measure, BM,** corresponds to the size of the largest common subtree of Sg and Si (measured in nodes), seen against the background of the integrated schema Si. **BM = | LCSub(Si, Sg) | / | Si |**
2. **Structural overlap** corresponds to the number of nodes shared by Si and Sg and included in a common subtree. **Sub** is the set of all disjoint subtrees (each containing a minimum of two nodes) common to Si and Sg. **kSub** is the total number of elements of all subtrees in Sub. **StructuralOverlap = kSub / |Si|**
3. **Structural proximity** considers the number of subtrees common to Si and Sg. **o** is the number of elements in Si that are not included in any common subtree, **o = | Si | - kSub**.
   StructuralProximity = kSub / sqrt(|Si|x|Sub| + o)



XBenchMatch: a Benchmark for XML Schema Matching Tools

duchatea@lirmm.fr, bella@lirmm.fr, hunt@inf.ethz.ch

Experiments

SCHEMAS

1. Person schemas are small and strongly heterogeneous.

2. Purchase orders, XCBL collection 3, demonstrate matching of a large schema to a smaller one.

3.University course schemas are from Thalia [4].

4. Biological schemas correspond to Uniprot protein DB, and GeneCards integrate data from over 100 databases.

TESTED MATCHERS

Porsche, COMA++ and Similarity Flooding.

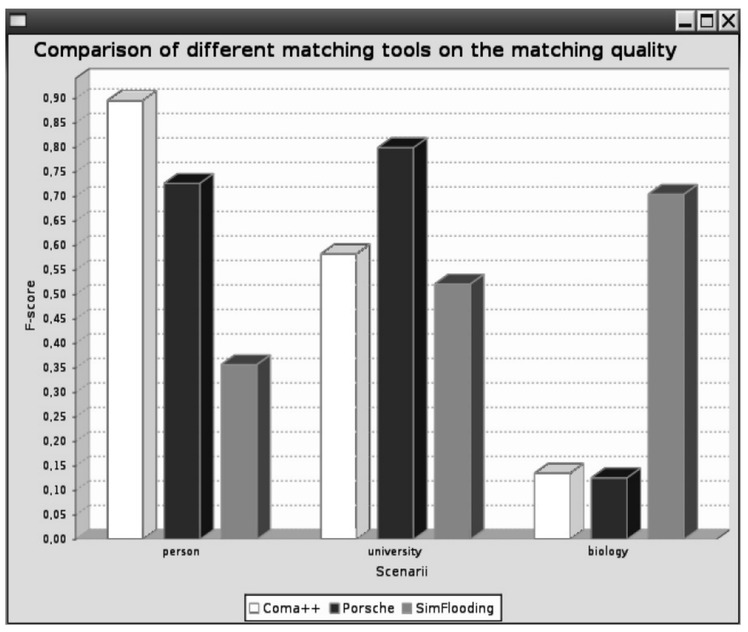


Figure 2. Matching precision on the three scenarios for three schema matchers.

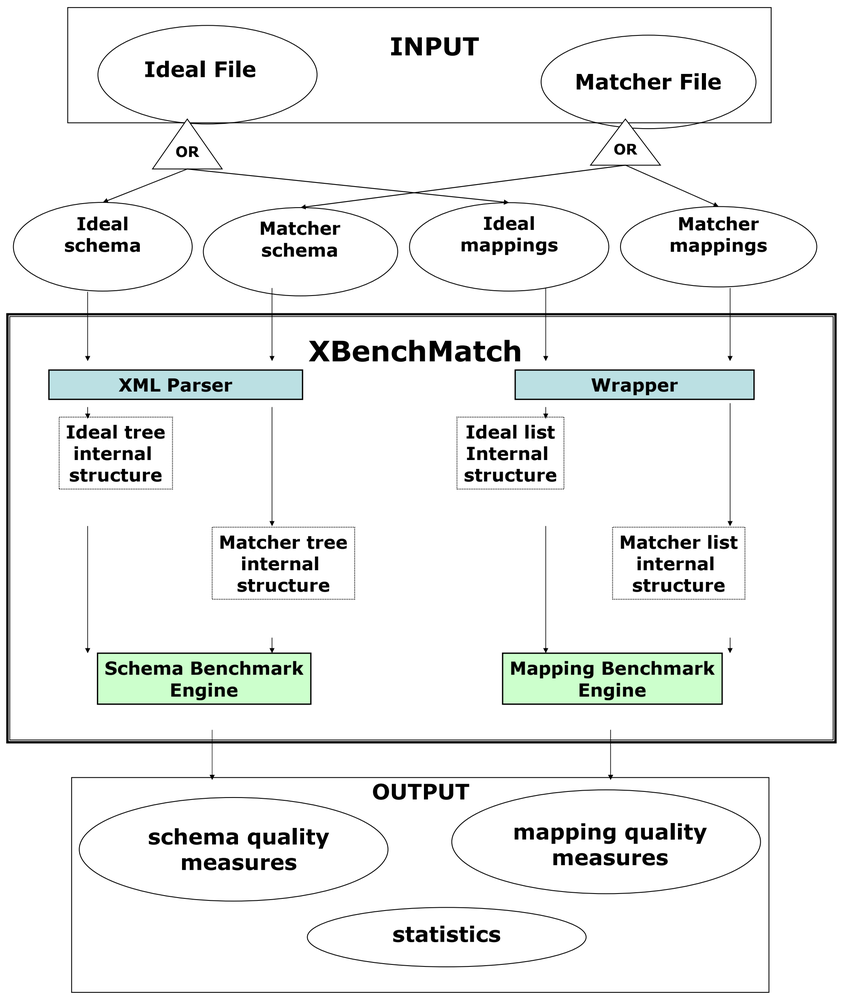


Figure 1. Architecture of XBenchMatch

	Person	University	Order	Biology
NB nodes (S_1 / S_2)	11 / 10	18 / 18	20 / 844	719 / 80
Avg NB of nodes	11	18	432	400
Max depth (S_1 / S_2)	4 / 4	5 / 3	3 / 3	7 / 3
NB of Mappings	5	15	10	57

Table 1: Summary of four evaluation scenarios.

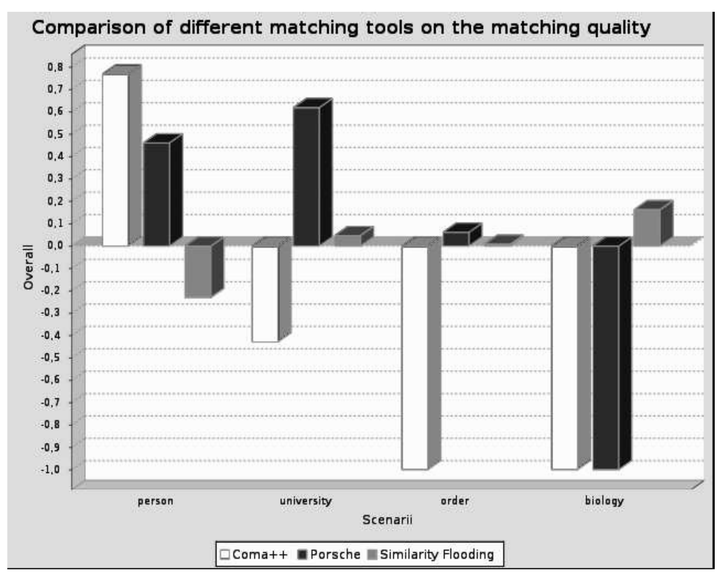


Figure 3. Matching quality of the four scenarios for three schema matchers.