Title: Integration and Provenance of Cereals Genotypic and Phenotypic Data

Authors: Domenico Beneventano, Sonia Bergamaschi, Abdul Rahman Dannaoui (1), Nicola Pecchioni (2)

Emails: firstname.lastname@unimore.it

Affiliations: University of Modena and Reggio Emilia

(1) Dept. of Computer Science - via Vignolese 905, 41125 Modena, Italy

(2) Dept. of Agricultural and Food Sciences, via Amendola 2, 42122 Reggio Emilia, Italy

ABSTRACT:

This paper presents the ongoing research on the design and development of a Provenance Management component, PM_MOMIS, for the MOMIS Data Integration System [1,2]. MOMIS has been developed by the DBGROUP of the University of Modena and Reggio Emilia (www.dbgroup.unimore.it). An open source version of the MOMIS system is delivered and maintained by the academic spin-off DataRiver (www.datariver.it).

PM_MOMIS aims to provide the provenance management techniques supported by two of the most relevant data provenance systems, the "Perm" [3] and "Trio" [4] systems, and extends them by including the data fusion and conflict resolution techniques provided by MOMIS. PM_MOMIS functionalities have been studied and partially developed in the domain of genotypic and phenotypic cereal-data management within the CEREALAB project [5]. The CEREALAB Data Integration Application integrates data coming from different databases with MOMIS, with the aim of creating a powerful tool for plant breeders and geneticists. Users of CEREALAB played a major role in the emergence of real needs of provenance management in their domain.

In [6] we defined the provenance for the "full outerjoin-merge" operator, used in MOMIS to solve conflicts among values;   this definition is based on the concept of "PI-CS-provenance" of the "Perm" system; we are using the "Perm" system as the SQL engine of MOMIS, so that to obtain the provenance in our CEREALAB Application. The main drawback of this solution is that often conflicting values represent alternatives; then our proposal is to consider the output of the "full outerjoin-merge" operator as an uncertain relation and manage it with a system that supports uncertain data and data lineage, the "Trio" system.

References

[1] Bergamaschi, S.,Castano, S.,Vincini, M., Beneventano,D.:Semantic integration of heterogeneous information sources. Data Knowl. Eng. 36(3), 215–249 (2001)

[2] Bergamaschi,S.,Beneventano,D.,Guerra,F.,Orsini,M.: Data integration. In: Handbookof Conceptual Modeling: Theory, Practice and Research Challenges. Springer Verlag (2011)

[3] Glavic, B., Alonso, G.: Perm: Processing provenance and data on the same data model through query rewriting. In: ICDE '09. pp. 174–185 (2009)

[4] Benjelloun,O.,Sarma,A.D.,Hayworth,C.,Widom,J.:An introduction to ULDBs and the TRIO system. IEEE Data Eng. Bull. 29(1), 5–16 (2006)

[5] Milc, J., Sala, A., Bergamaschi, S., Pecchioni, N.: A genotypic and phenotypic information source for marker-assisted selection of cereals: the Cerealab database. Database (2011)

[6] Beneventano,D.,Dannoui,A.R.,Sala,A.:Data lineage in the Momis data fusion system. In: ICDE-Workshops, April 11-16, 2011, Hannover, Germany. pp. 53–58 (2011)