# Postdoc position: Practical Algorithms for Big Data Curation

**Skills and profile:**

We are seeking excellent candidates with background on database systems and database theory.
The position is available immediately. The candidates should contact us by sending a CV with a detailed list of publications, a statement of research interests, and names and contact information of three references by e-mail with Subject 'Postdoc application'.
Priority will be given to applicants who apply by June 30th, 2016.

**Duration**: 12 months (possibility of extension)
**Where**: University of Lyon 1/Liris CNRS
**Contact**: Angela Bonifati (angela.bonifati@univ-lyon1.fr)
**Starting date**: the position can be filled immediately, and not later than September 1, 2016.
**Salary**: 2320 euros gross/month.

**Topic**:

Massive datasets are pervasive in today's user applications and their curation is of paramount importance to ensure the quality of results in data analytics, decision making and knowledge discovery. Despite the abundance of techniques for data repairing and data cleaning [4,5], practical feasible algorithms are still missing in the literature. The size of involved datasets, the number of involved constraints, the variety of data formats often affect dramatically the performances of existing algorithms for data repairing and integration, based on the classical chase procedure [6].
Recently, the notion of causality has been proposed for tuples [1,2] and cells [3] to detect the causes of errors in the data. Such a detection is ensured by polynomial-time algorithms [2]. The notion of chase repairs could be key to the use of constraints to propagate information curated by experts to other data that depends on such information via those constraints. Such a propagation has to work in tandem with the repairing itself and has to achieve practical performances. This also brings up the question of defining new provenance models that can be used for data quality and that can be easily adapted to the data variety often found in Big Data, such as for instance for graph-shaped data. In a sense, we believe that by leveraging these models, one can find the same benefits of query explanation in query optimizers and thus optimize the repairing process.

*The ideal candidate should have the following qualifications:*

- *a Ph.D. in computer science or closely-related field*
- *a solid background in the area of database systems*
- *ability in designing methods and formalisms for Big Data curation. Both mixed (theory + applications), and purely theoretical contributions are welcome.*
- *a good command of spoken and written English.*

[1] A. Meliou, W. Gatterbauer, J. Y. Halpern, C. Koch, K. F. Moore, and D. Suciu, Causality in Databases. IEEE Data Eng. Bull., 33(3), pp. 59-67, 2010.
[2] B. Salimi and L. E. Bertossi, From Causes for Database Queries to Repairs and Model-Based Diagnosis and Back. ICDT, pp. 342-362, 2015.
[3] M. Debosschere, F. Geerts. Cell-Based Causality for Data Repairs. In TaPP, 2015.
[4] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. Schema Matching and Mapping. Springer Berlin Heidelberg, 2011.
[5] X. L. Dong, D. Srivastava: Big Data Integration .Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2015, pp. 1-198
[6] A. Bonifati, I. Ileana, M. Linardi. Functional Dependencies Unleashed for Scalable Data Exchange. In SSDBM, 2016. (to appear)