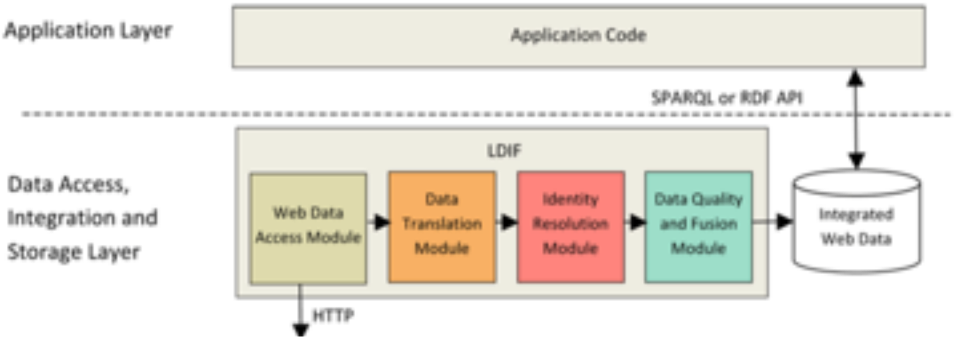
Data Fusion research methodology description.

General Overall architechture for data fusion from multi provenance data using LDIF and SIEV.



Pre ranking steps: -

1) Collect Data: import data by crawling or sparql query
2) Map to schema. - R2R framework for mapping data set.
3) Resolve Identities. - LDIF employs the Silk Link Discovery Framework [1] to find different URIs that are used within different data sources which identify the same real-world entity.
4) Quality Assessment and data fusion. - Using sieve data quality and fusion framework. We present different mechanism for sieve to learn and ranking of data sources.
5) Output - Cleaned data with provenance information

Different Quality assessment method :-

1) Machine learning on SIEVE strategy. - sieve takes URI with properties and multiple properties and each property having different provenance and applies strategy manually assigned to it and gives the output. So we can have a data set of truths and a set of XML based strategy to be employed. Then we choose the function which gives the least min threshold error on gold standard. [2]
2) Measuring accuracy as a function of probability of being copied and source authenticity.[3]

Refrences:-

[1] LDIF - A Framework for Large-Scale Linked Data Integration
Andreas Schultz Web-based Systems Group Freie Universität Berlin, Germany a.schultz@fu-berlin.de Andrea Matteini mes|semantics Berlin, Germany a.matteini@mes-info.de
Robert Isele Web-based Systems Group Freie Universität Berlin, Germany mail@robertisele.com
[2] Learning Conflict Resolution Strategies for Cross-Language Wikipedia Data Fusion
Volha Bryl, Data and Web Science Group University of Mannheim
Christian Bizer , Data and Web Science Group University of Mannheim
[3] Data Fusion: Resolving Conflicts from Multiple Sources
Xin Luna Dong^1, Laure Berti-Equille^2, and Divesh Srivastava^3