**Course: Advanced Bioinformatics**

**Module title: Problems with Data Integration**

**Module no. : 215**

**Problems with data integration**

* Data spread over multiple, heterogeneous data sources

• Databases (MySql, Oracle etc.)

• Semi-structured sequence files (text or XML)

• HTML format in Web sites

• Output of analytic programs (BLAST, PFAM, etc.)

* Use different terms and definitions
* Lack of standards in data representation

• XML is emerging as a standard for data transfer

* Data in the original data sources can change
* Poor data quality(errors) and incomplete data e.g. due to errors in Labs
* Patients’ rights
* Confidentiality
* Security
* Privacy