Data migration - A case based task

Nagamani Maddipatla[1], Dr. S. Vidyavathy[2], Dr. K. Mrutyunjaya Reddy[3]

[1]Research Scholar, JNTUH University, Hyderabad, India
[2]Associate Professor, JNTUH University, Hyderabad, India
[3]Dept. of Space National Remote Sensing Center (NRSC), Hyderabad, India

***Abstract*—Data migration is one of the vital tasks of Data integration process. It is always assumed to be most tedious as there will never be a systematic defined procedure. Each migration process is to be treated as unique as the input data sets will be different and the output format required is always unique based on the services provided as well as the user and data handler requirements. In the recent years data migration became the most vital process in various departments of public and private services due to technological advancements and big data handling requirements caused by the increase in acquired data volume.**

This paper discusses about data migration requirement, data migration strategy finalization and various stages of data migration process discussion of each stage and why complete automation of data migration is not feasible etc.

***Keywords*—Data migration; Quality assurance; legacy systems; GIS; Database.**

I. Introduction

Transforming one format to another is commonly referred as Data migration. It is to be noticed that data migration is not only the data storage systems upgradation, but it is the process to be performed to accommodate changing business conditions of the present world and also ensuring the customers about the quality of the data being shared or distributed. Today's world most of the time depends on the information (data) being provided. This process is required under various circumstances like

- When existing data is to be upgraded to make it compatible with latest technology improvements like systems, platforms and applications.
- When multiple data sets are to be combined to form and support existing process
- When various data sources are to be collectively used to acquire better information

In all these cases source data can be from a single source or multiple sources where data structure is entirely different. If this data includes geometric information then various other factors like below are also to be taken into consideration before planning data migration process.

- Scale of source data and desired destination or output data scale
- Units of measure of both source and destination data and conversion process accuracy
- In case of geographic data the mapping and projection models used to acquire the data and conversion algorithm accuracy

On the outlook, Data migration process looks to be simple and straight forward (transforming data from one format to other). But once it is started its complexity and intricate problems will be experienced and becomes very tedious and time consuming which may cause the company to endure losses and even goodwill. Though data migration is complex and intricate, it is a necessity to be in line with business requirements and improve business avenues.

Goals of Data migration

Data migration goals can be broadly explained as below from various points of view.

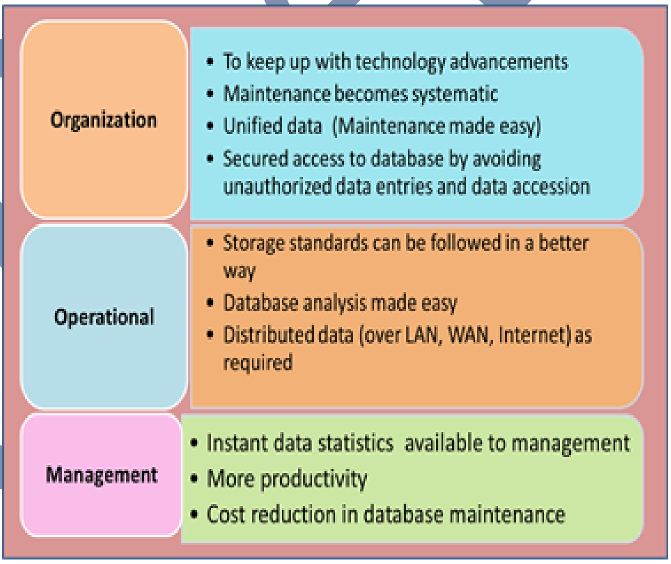

Organization
To keep up with technology advancements
Maintenance becomes systematic
Unified data (Maintenance made easy)
Secured access to database by avoiding unauthorized data entries and data accession
Operational
Storage standards can be followed in a better way
Database analysis made easy
Distributed data (over LAN, WAN, Internet) as required
Management
Instant data statistics available to management
More productivity
Cost reduction in database maintenance

Fig-1 goals of data migration

It can be seen that effective data migration provides improvement in the tasks of management, operational team and broadly in organizational betterment.

Data migration is generally of two types. When applications of a production chain are modified to meet customer requirement, then data migration is to be carried out to support such applications. Then it is called application driven data migration. In case of database driven data migration, requirement of applications does not change but the bulk of the data or the contents of the data may change to cater the customer needs better. In that case in addition to data migration applications upgradation is also to be carried out to support such data format and content changes.

In both the cases data structure may be modified to cater the needs of the present business requirements. The success of data migration process depends on various factors

- Migration strategy chosen
- Technology expertise
- Domain expertise
- Good risk management capability
- Good quality check system of migrated data

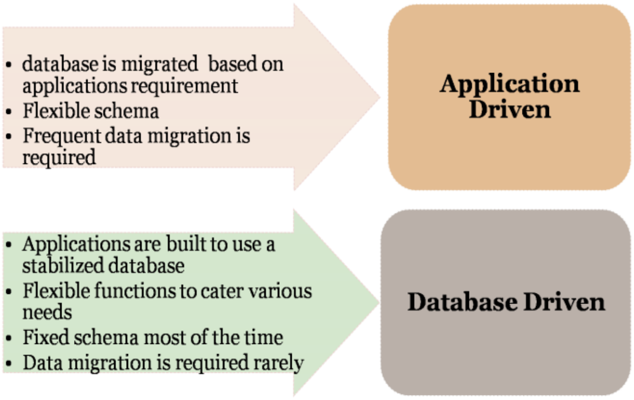

database is migrated based on applications requirement
Flexible schema
Frequent data migration is required
Application Driven
Applications are built to use a stabilized database
Flexible functions to cater various needs
Fixed schema most of the time
Data migration is required rarely
Database Driven

Fig-2 Models of Data Migration

Migration Strategy

Migration strategy is one of the vital factors of successful data migration process. Generally 2 types of strategies are in use.

.

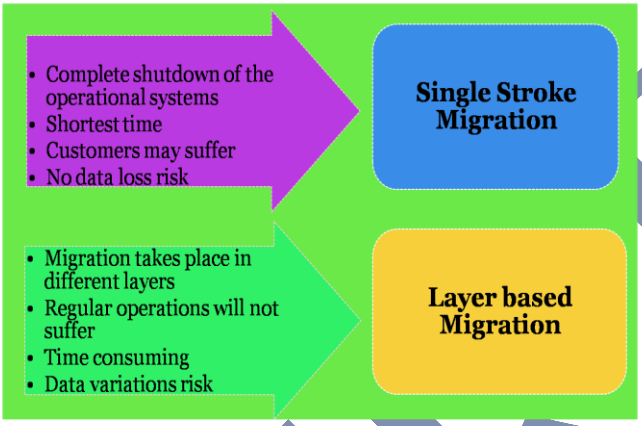


Fig-3Data migration strategies

In case of Single stroke migration complete operations shutdown is essential by which customers may suffer. But as it is of shortest time, it can be run after working hours and on holidays. But in real time operational business where customers support is required 24X7, it is risky to adopt this method. Failure risks can be minimized as operations shutdown takes place. Due to complete shutdown of operational systems, supporting staff will be left unengaged. To avoid this problem generally they will be trained on the migrated data systems during the shutdown time.

In case of Layer based migration, complete operational systems shutdown is not required. Migration will be carried out in a layered manner so that only a small part of the operations suffer for a time period. This may ensure continuous support of other operational systems and operational staff will be effectively engaged. Data loss risk is little high in this case as the migrated data may be used and modified on-fly by other operational systems. Sufficient care is to be taken to avoid this kind of data loss issue.

Technology Expertise

Technology expertise is another factor of successful migration process. The migration team should have sufficient expertise in handling the issues raised during the migration process and quick decisions are to be taken to avoid data loss as well as migration time. Training of operational staff on upgraded system also requires technology expertise.

Domain Expertise

Domain expertise helps in taking required decision during data migration plan finalization. Connectivity between various fields and tables of the data is only known to domain experts. Unless this is thoroughly known validation of migrated data cannot be done effectively.

Good risk management capability

To achieve this there should be complete understanding between technology experts and domain experts. Mistakes of one group are to be quickly corrected by other group wherever required without waiting for protocols. But at the same time log of all these corrections are to be properly maintained to avoid such risks further. To achieve this good project management team is also to be introduced in to the migration team. Close follow-up of the migration task can also be achieved by this team so that proper standards can also be followed during migration process.

Good quality check system of migrated data

Migrated data validation is the measure of successful data migration. Generally following are the quality checks to be made on migrated data.

- Proper access to data achieved or not
- File permissions are satisfactory or not
- Fields of all tables are properly defined or not
- Working conditions of the applications on migrated data
- Proper connectivity and parent-child relationship between features is maintained or not
- Proper business rule base establishment in migrated database

Various data migration procedures are in usage. A procedure used for one source data may not be optimal for other source data. Hence optimal procedure can be derived based on the source data complexity and heterogeneity.

Following is the generic flow chart of the migration process. It contains basically three steps each of which further contain multiple tasks.

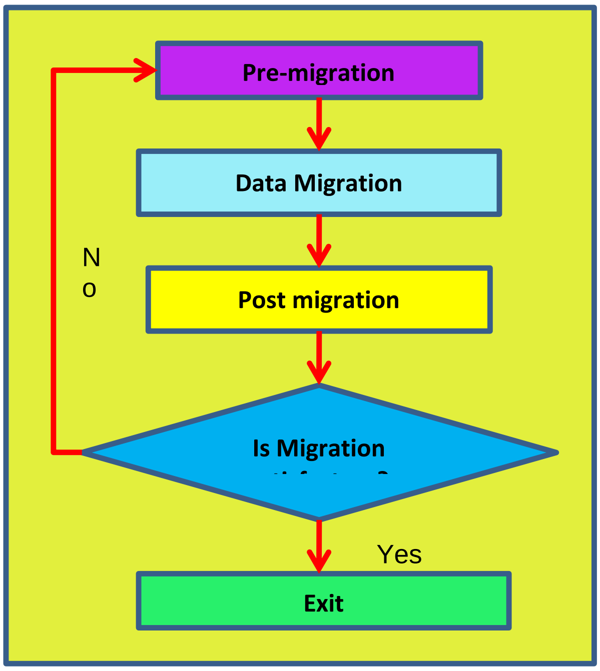


Fig-4Data migration generic flow chart

Following is the data migration methodology to be followed sequentially in general. Based on the data type and volume of the data some of the parts can be shortened but not eliminated.

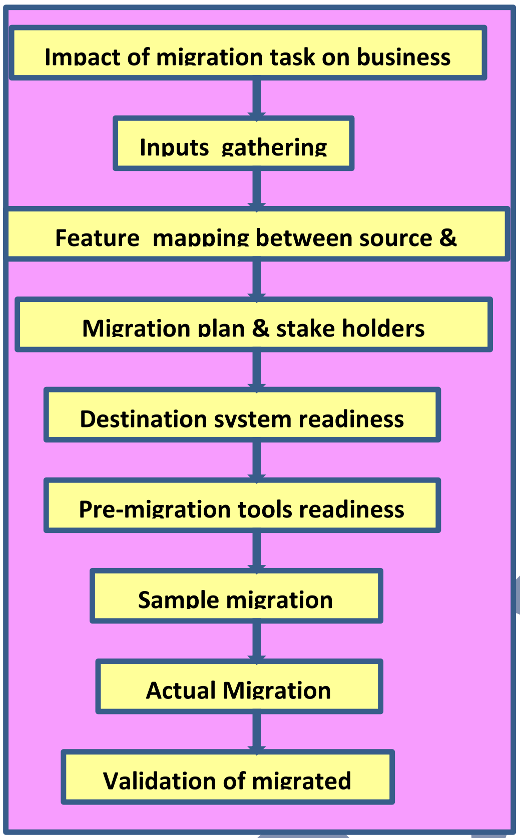

Impact of migration task on business
Inputs gathering
Feature mapping between source &
Migration plan & stake holders
Destination system readiness
Pre-migration tools readiness
Sample migration
Actual Migration
Validation of migrated

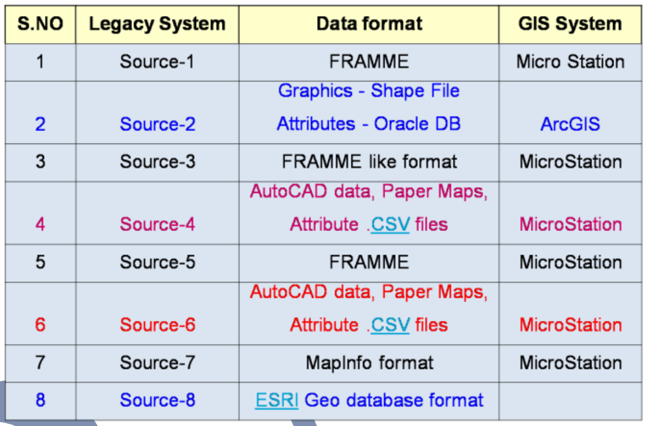
Fig-5 Data migration steps

Various sectors of organizational staff face inconveniences during migration process due to non accessibility of data. Hence care is to be taken to minimize this issue by following an accurate checklist of operations to be performed as a part of migration. Prior indications should be passed on to avoid non accessibility.

Inputs gathering step involves requirements of various business users and destination system planned to use. Source data analysis, finding issues with source data and planning to rectify them during migration process are few critical tasks of this step.

Feature and attribute mapping between source and destination data-models is a very critical step of data migration process. Each feature attributes are to be correctly defined and if required more details are to be added to improve productivity and automation of management requirements. For this a mapping chart preparation is done so that the same can be used readily wherever required.

Destination data-model finalization and system readiness is to be done prior to migration plan and stake holders finalization as concerned people will be made responsible and migration plan can be followed by all. Data migration is an iterative process due to in the data. If source data is from a single source then there will not be much risk. But if the source data is from multiple sources and of different formats then it is always better to bring all these formats to a common known format and then migrate the entire data in one stroke to destination system. This not only avoids post migration issues but also helps us to migration of various systems in parallel. This kind of scenario happens when big companies take up mergers and want to combine all the data to form an unified data model to support customers better. Here is an example where source data is of eight different databases which is to be transported to Intergraph G/Technology.



S.NO	Legacy System	Data format	GIS System
1	Source-1	FRAMME	Micro Station
2	Source-2	Graphics - Shape File Attributes - Oracle DB	ArcGIS
3	Source-3	FRAMME like format	MicroStation
4	Source-4	AutoCAD data, Paper Maps, Attribute .CSV files	MicroStation
5	Source-5	FRAMME	MicroStation
6	Source-6	AutoCAD data, Paper Maps, Attribute .CSV files	MicroStation
7	Source-7	MapInfo format	MicroStation
8	Source-8	ESRI Geo database format	

Fig-6Source data details for migration

In the above situation where GIS data is also involved in migration process positional accuracy is vital for success of the migration process. G/Technology is the latest platform and data-model understanding and mapping itself will take lot of time. If source data is to be directly transported to G/Tech from each data source, then if by any change an issue occurs in between then migration process is to be initiated for all the data sources again. Otherwise any future migration error identification may result in lot of time, manpower and money wastage.

Instead if an intermediate platform is used with similar destination database to migrate each data source, migration process will become much simpler and fast. Below is the block diagram of a middleware which allows data import, cleanup, correction and validation.

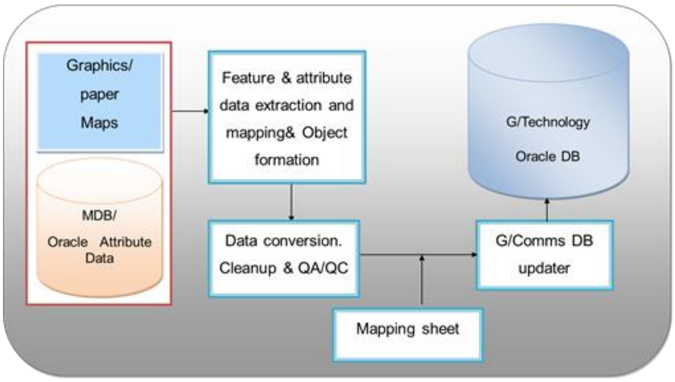

Graphics/ paper Maps
MDB/ Oracle Attribute Data
Feature & attribute data extraction and mapping& Object formation
Data conversion. Cleanup & QA/QC
Mapping sheet
G/Comms DB updater
G/Technology Oracle DB

Fig-7 Block diagram of middleware based migration

Following are the advantages observed during middleware based data migration

- All data sources can be migrated in parallel as destination databases are similar but different.
- Mapping can be handled better for each data sources as complexity will be reduced

- Data reconciliation tools during pre-migration stage can be simpler as each data source will be handled separately.
- Migration time will be reduced exponentially.

There are few limitations also for this kind of migration process which are listed below.

- Validation tools are to be developed at two levels (intermediate level and on G/Tech platform)
- Manpower requirement will be very high as one team is to be identified for each data source.
- Domain expertise should be of highest order as G/tech is the latest and still unknown to many.

Below is the sample data after migration and verified with respect to Google Earth raster data. If this level of verification is done then the percentage of failure can be very minimal.

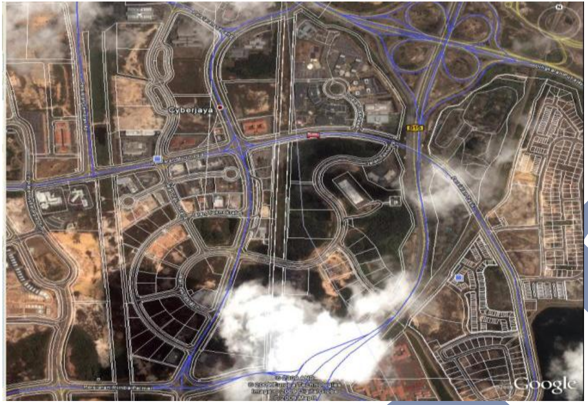


Fig-8 Electrical network overlaid on Google image

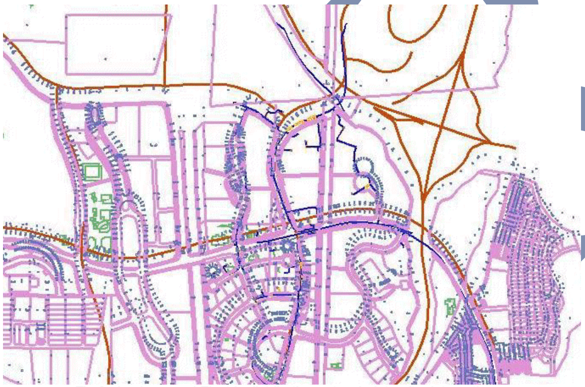


Fig-2. Lanbase and Electric distribution network data after integration

Fig-9 Electrical network overlaid on Landbase data

When we compare migrated data with images of Googlemaps we may not be able to identify the errors completely as network GIS data like (electrical or telecom)will be of centimeters accuracy. But with reference to landbase data (which is of highest accuracy of millimeters , we can easily identify the migration issues.

II. GUIDELINES FOR SUCCESSFUL DATA MIGRATION

A systematic approach to data migration will result in an error free process and output. Prior to data migration following key-points are to be thoroughly examined

Destination system study: Appropriate destination system selection is vital for successful and long lasting migrated data or system. Following are some of the factors to be taken into consideration during destination system selection

a) System procurement should include both hardware and software items and it should be within budget limits.
b) Understanding with the vendor about free supply of upgrades within warranty period
c) LAN configuration support
d) Training needs are also to be included as a part of procurement process if required.
e) Installation procedure is to be thoroughly understood and relevant documents are to be procured from vendor
f) Maintenance related understanding is to be established with vendor based on requirement
g) documentation and manuals are to be part of procurement
h) Online help for operations and customization should be available as a part of software.

If GIS data is to be migrated then additional points are also to be considered. Appropriate GIS System selection improves user's operational efficiency and effective budget usage. GIS users need to be aware of different GIS software products during system selection and beyond.Informed choice is the best way to select the best GIS

a) A GIS is often defined not for what it is but for what it can do.
b) If the GIS does not match the requirements for a problem, solution will notbe forthcoming.
c) A GIS may have overcapacity.
d) Six critical functions of GIS (data capture, storage, management, retrieval, analysis and display) are to be verified before selecting

- **Coordinate system to be adopted for the migrated data:** When single source data set is being handled, destination coordinate system selection at a later stage of data migration process may not cause any issue as projection transformation will be carried out only once. But when multiple heterogeneous data sources are to be migrated into a single destination system, it is wise to use an intermediate platform where finalized coordinate system is implemented. This not only reduces transformation issues but also provide a common platform for iterative data transportation process. From this intermediate platform, data can be transported to destination data model in a single step.
- **Type of facilities (end usage) expected from the migrated data**: Nowadays all the data contains geographic information also. For example in case of electric network data migration, user may have many other functional requirements in view once the migration task is completed like migration of distribution management system, billing system, load management system etc. This may require additional tables inclusion to destination data model and additional data inclusion in already designed tables and relationships creation between these tables. Adding these functionalities at a later date will be more tedious and may need profound changes in data model. Hence it is always better to keep in view all the foreseeable future plans also during destination data model creation. This not only provides a comprehensive data model but also reduces budget requirements.

Following are few pointers which will help in formulating efficient methodology for data migration

- Complete source data analysis
- Sample data migration and verification in terms of elements transportation as well as their positional accuracy.
- Revision of process timeframe based on the above

- Finalization of destination data model at the earliest possible
- In case of multiple heterogeneous source data sets, all of them are to be analyzed before finalizing destination data model
- Segregation of source data elements based on their properties and behavior so that common quality assurance tools can be developed/ customized to process in bulk
- Care is to be taken to confirm that the units of linear, area features and feature orientation angles are same for source and destination data models. If not proper translation mechanism is to be adopted

Sometimes positional accuracy of the migrated data may not be uniform throughout the area of interest (as shown below). In such cases to align the data with reference data we may have to use vector conflation (localized accuracy improvement) concept. In the below figure inaccuracy can be clearly observed. This will be visible with reference to large scale cadastral data only.

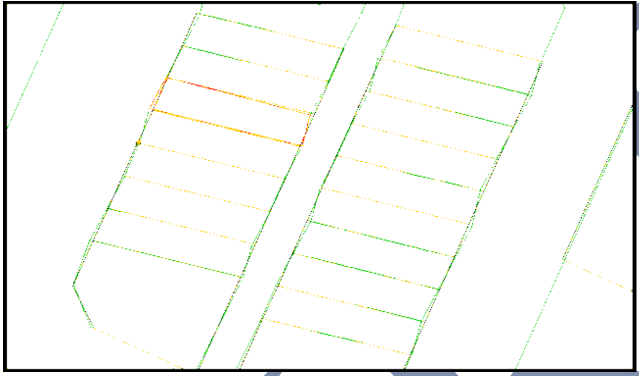


Fig-13 Vector conflation issue

Hence source data validation is to be done with reference to cadastral land-base data before data migration process itself. Data conflation is a time consuming process hence sufficient time is to be allotted to this process for achieving better results.

ACKNOWLEDGMENT

Authors are very much thankful to Sunil Kumar Makkena, Senior vice president, Infotech Enterprises Ltd, Hyderabad for approval of information usage and encouragement in preparation of this paper. We are thankful to Sri. M.V. Nagireddy, Intergraph, Hyderabad for reviewing this paper and providing valuable suggestions. We are also thankful to Sri. Dr.M.Anji Reddy. Director, Directorate of R&D, JNTUH, Hyderabad for his cooperation during research period and providing valuable insights in proceeding with the research work.

REFERENCES

Very few papers were published regarding this topic so far. Almost all papers published were from IEEE only. Hence effort was put to get maximum information not only from journals but also from various websites.

Websites referred

[1] http://trac.osgeo.org/proj/

[2] http://www.epsg.org./

[3] http://mathworld.wolfram.com

Article in a journal:

[4] Serge Abiteboul, Sophie Cluet, Tova Milo, PiniMogilevsky, Jerome Sim´eon and Sagit Zohar, “Tools for Data Translation and Integration”, IEEE Computer SocietySpecial Issue on Data Transformations, vol. 22, issue 1, pp 3-9, March 1999.

[5] Philip A. Bernstein and Thomas Bergstraesser, “Meta-Data Support for Data Transformations Using Microsoft Repository”, IEEE Computer Society

[6] Special Issue on Data Transformations, . 22, issue 1, pp 10-15, March 1999

[7] Marco Carrer, Ashok Joshi, Paul Lin, and Alok Srivastava, “Metadata Transformation and Management with Oracle”, IEEE Computer SocietySpecial Issue on Data Transformations, vol. 22, issue 1, pp 16-19, March 1999.

[8] Kajal T. Claypool and Elke A. Rundensteiner, “Flexible Database Transformations: The SERF Approach”, IEEE Computer Society

[9] Special Issue on Data Transformations, . 22, issue 1, pp 20-25, March 1999

[10] Susan B. Davidson and Anthony S. Kosky, “Tools Specifying Database Transformations in WOL”, IEEE Computer SocietySpecial Issue on Data Transformations, vol. 22, issue 1, pp 26-31, March 1999.

[11] Laura Haas, Renee Miller, Bartholomew Niswonger, Mary Tork Roth, Peter Schwarz, and Edward Wimmers, “Transforming Heterogeneous Data with Database Middleware: Beyond Integration”, IEEE Computer SocietySpecial Issue on Data Transformations, . 22, issue 1, pp 32-38, March 1999

[12] Sandra Heiler, Wang-Chien Lee, and Gail Mitchell, “Repository Support for Metadata-based Legacy Migration”, IEEE Computer SocietySpecial Issue on Data Transformations, vol. 22, issue 1, pp 39-44, March 1999.

[13] Joseph M. Hellerstein, Michael Stonebraker, and Rick Caccia, “Independent, Open Enterprise Data Integration”, IEEE Computer Society

[14] Special Issue on Data Transformations, . 22, issue 1, pp 45-51, March 1999

Article in a conference proceedings:

[15] Lenzerini, M., “Data Integration: A Theoretical Perspective “,Symposium on Principles of Database Systems:
Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART, Vol 3, issue 5, pp 233-240, 2002

[16] Erhard Rahm, Hong Hai Do, “Data Cleaning: Problems and Current Approaches”, IEEE Data Engineering Bulletin 2000

[17] YannisPapakonstantinou, Hector Garcia-Molina, Jennifer Widom, “Object Exchange Across Heterogeneous Information Sources“,IEEE Data Engineering Bulletin 2000

[18] David DeHaan, David Toman, Mariano P. Consens, M. Tamer Ozsu, “A Comprehensive XQuery to SQL Translationusing Dynamic Interval Encoding”, University of WaterlooSchool of Computer ScienceWaterloo, Canada, IEEE Data Engineering Bulletin 2000.

[19] Ronald Fagin,AmnonLotemand MoniNaor, "Optimal aggregation algorithms for middleware", IBM Almaden Research Center, 650 Harry Road, San Jose, CA, IEEE Data Engineering Bulletin 2000.