FROM THE "DATABASE" TO "DATA" COURSES

Tatiana Malyuta
tmalyuta@citytech.cuny.edu
NYC College of Technology
300 Jay Street Brooklyn, NY 11201

Abstract: The paper discusses different roles of data in today's applications, Information Systems (IS) and the Web, the new approaches and data technologies that support these roles, and the necessity to reflect these issues in the curricula of computer majors. Databases remain the most important topics of data courses, which have to be updated with coverage of the tremendous advances in database technology that have taken place during the last few years. However, discussion of data has to go beyond traditional databases and include such important topics as design and implementation of data for exchange between applications and for presenting knowledge about particular business areas, as well as data models and technologies used for these purposes. Just adding these issues to curricula of the database programs makes it difficult to provide systematic coverage of data technologies. The paper suggests transforming *database* curricula into *data* curricula that will allow for developing a more consistent view of today's data technologies and better teaching approaches, and will provide the possibility for easier upgrading of data programs in response to advances in data technology.

Such *data* curricula is demonstrated for a special modular undergraduate program of the Computer Systems Technology (CST) Department of the New York City College of Technology (NYCCT).

New roles of data and data technologies come together with the changed paradigm of the IS architecture and the new ways of using Web resources, which calls for changes in other computer major courses as well.

Key words: data, database, metadata, ontological data, XML technologies.

Introduction:

From traditionally being the *object* of processing by the Information System (IS) and discussed as a database, data more and more have been becoming the *subject* of processing as the system's metadata and the *skeleton* of the system with the Service-Oriented Architecture (SOA) [1], [2]. Furthermore, as ontological data, now it goes beyond a separate application or IS in the form of the *Semantic Web* and allows for linking various Web resources [3]. All these purposes of data are supported by the corresponding data models: semi-structured and object-oriented, and technologies: eXtensible Markup Language (XML), Web Service Language Description (WSLD), Resource Description Framework (RDF), Ontology Web Language (OWL), and others.

The curricula of computer majors traditionally include the discussion of the concepts of design, implementation and management of a database as the object of processing of an application, as well as the relevant technologies—Database Management Systems (DBMS). Coverage of the new data technologies is added to the database textbooks and database courses, however, it is often superficial, and the purpose of these new technologies is not clearly explained. To present a systematic picture of data and data technologies, the curricula of computer majors have to move from the "database" courses to "data" courses, which discuss the multiple roles of data in the IS, and the design and implementation of data of each role with respect to the employed technologies.

Database technologies themselves have undergone important changes in the last few years, and these changes have to be reflected in the database topics of the data programs.

Current state of database curricula:
The published textbooks reflect the most common curricula in the corresponding area. Judging by the database textbooks, traditional database curricula cover such issues as design, implementation and management of databases. Relational databases, as the most widely used ones, get most of the attention. In the last years, the new editions of the textbooks add discussion of database application development, distributed databases, data warehouses, and XML technologies. There are several problems with adding the new issues to those previously included in the curricula: the absence of systematic approaches and insufficient explanations of the role of the new models and technologies, and superficial coverage of the new topics. These problems are quite understandable—usual conservatism and inertia of curriculum development especially in computer majors, that cannot keep up with rapid changes in IT, reluctance to perform radical changes in curricula and textbooks, and the impossibility to cover the spectrum of today's existing approaches and technologies in a limited number of courses. From our point of view, it is time for reconsidering and rebuilding *database* programs of computer majors and transforming them into *data* programs, in which the roles of data in IS are clearly defined, and the methodologies of design and implementation, as well as the corresponding technologies, are explained in these roles' context. Such an approach, in our opinion, will not only provide a systematic coverage of data technologies, but will also allow for easier and more consistent upgrading of data curricula in the future. Discussing databases within data curricula does not diminish their role and place in IS, and they will still get most of the attention—respectively to their importance, usage, and the need in special situations..

This paper makes general recommendations on curricula changes that will allow for providing both a systematic and deeper coverage of data-related issues. It tries to systemize the data related topics and suggests how to incorporate them into the special baccalaureate program of the CST Department of NYCCT.

Data and data technologies—important changes:
The main changes in data and database technologies were directed to support automated data integration and exchange, elimination of data duplication, and management of rapidly growing databases and the increasing number of database users. These changes are not reflected or reflected insufficiently in the current curricula:

- *Databases design*. Discussion of database design in most textbooks is provided using the old examples and case studies for different companies or organizational units of companies, e.g. departments and offices. Such an approach that has been used for years resulted in multiple databases containing data about the same business entities. Today, the business is emphasizing design of databases for business processes. Emergence of Business Process Modeling (BPM) and its application in database design has not received enough attention in curricula and textbooks.
- *Database technology*. Database technology and DBMSs have undergone tremendous changes in the last years in response to the needs of business and the growth of databases—"data tsunami." The new features of DBMSs include extended SQL functionality, object features, data warehouses support, support and processing of media data and semi-structured and

unstructured data, self-managing features, new approaches to data storage, data security, optimization of query processing [4], [5]. These features are not discussed sufficiently and their role and importance are not explained enough.

- *Data warehouses*. Data warehouses highlight many problems of today's data technology and at the same time they demonstrate progress in resolving these problems. Their design, the dimensional model of data, the impact of data design on the database performance and ease of use, integration of data from different data sources, and some other issues can serve as excellent examples for discussing important issues of databases and data. However, these topics do not get enough attention in database curricula.
- *Data exchange and integration*. The Internet started a new era in B2B communications and created new possibilities for data and functionality exchange. However, the new possibilities came at a huge cost (consumption of technology in Fortune 500 companies has grown at least 10 percent every year for the last five years). The cost of integration can be explained by the complexity of implementation and support of interfaces between applications and integration of data. XML as a standard of data exchange and the subsequently developed XML technologies such as WSDL and OWL promise a cheaper, less painful and more flexible integration approach. The promises of XML are defined by two main factors: standardization and openness of the tools. The development of XML technologies went through the following steps: the standard way to exchange data, the standard way to exchange functionality, and the standard way to describe any resource. Essentially, any interface implemented within the XML approach is a standard, and, therefore, understandable to everyone or any application description of it with the help of data. From being a part of the information system and the object of the system's processing, data penetrated all parts of the system and became the medium in which the parts co-exist. Explanation of the changed, more important, and different roles of data in information systems, of technologies that support the changes, the issues of design, implementation and support of data are insufficient in most database curricula.

  Limiting coverage of data and data technology to the traditional discussion of data in databases leaves the curricula of computer majors behind important advances in IT. The curricula should move from narrower database topics to the wider data topics.

Database program at NYCCT—special features and potential:

We want to describe a flexible database program within the baccalaureate program of the CST Department of the NYCCT and show how it can be developed to include the important topics of data design and management depending on the role of data in the IS. The baccalaureate program is based on specialization in three (out of seven) areas (directions) of information technologies. Each area is represented with a module—a group of courses of three consecutive levels (http://www.citytech.cuny.edu/academics/deptsites/cst/programs.shtml). One of the modules is "Databases." Currently, the program consists of four courses[1] that allow for a more detailed discussion of databases than in one or two courses programs (see Figure 1). Though most of the discussed topics are database related, the program has been evolving and changing to reflect the important achievements of the data technology. Currently, the program covers in depth many

[1] Associate Program includes a separate course on the programming language of relational databases—SQL.

advanced issues that are not traditionally included in undergraduate computer programs. Due to its unique and special structure, the program is very flexible and can react faster to the changes in the data technology.

The "Database Design" course is similar to database courses in most undergraduate computer programs. It covers the concepts of database design, the role and place of the database in the IS, includes a detailed discussion of the relational data model and implementing a relational database[1].

The "Distributed Database Design" course is not typical for college undergraduate programs. It concentrates on physical design of data (distribution of data is one of the ways to build a physical data model) and implementation of databases. These issues require understanding of the database technology, which is why a significant part of the course is dedicated to the features and functionality of DBMSs. The objectives of the course make it very dynamic—during last few years it has undergone many changes in response to the changes of database technology. Today the program of the course includes coverage of the most advanced features of DBMSs used for implementation of the physical data model and support of scalability, security, performance, concurrent access, and reliability of databases [5].

The course "Database Administration" is aimed to continue the discussion of functionality of DBMSs started in the "Distributed Database Design" course. It provides a more detailed and in-depth coverage of the features of database technology. Similar to its predecessor course "Distributed Database Design", this course is very dynamic.

The course "Data Warehouses" reflects the latest tendencies in data technologies. Data warehouses represent very large integrated databases. Currently the course discusses design of very large databases and application of database technologies for implementation and support of these special databases. In addition to the advanced database design topics, the latest enhancements of SQL for analytical queries, and features of DBMS for support of dimensional data design, the program includes discussions of integration of databases and applications, the role of metadata in the IS, XML technologies, and SOA.

The special IT project is aimed on giving students an additional in-depth and hands-on experience in different areas of database development and maintenance. It enhances the program's flexibility and allows for the testing of discussions and teaching of the new concepts of data technology before introducing them in the courses of the program.

Database program at NYCCT—future:
We find it difficult to continue the program's improvement by keeping it database-centered and by adding discussion of the new topics and technologies into separate courses. By reorienting the program and making it data-centered, we will be able to maintain a sequence of courses that discuss data technologies in the logical and consistent manner. We see it possible to perform initial changes within the existing courses:

[1] The department is a member of the Oracle Academic Initiative. Oracle database is used for demonstrations of various database concepts and practical work.

- "Database Design". This course will be dedicated to the explanation of the roles of data in the today's IS and renamed to "Data Design." Data as an object of data processing is stored in databases within some particular structure, e.g. relational, and database design will continue to be the important part of the course. We plan to emphasize the design of databases for business processes. Above this primary data layer, the IS can contain data that defines the processing itself, or, we say, data serves as or defines the subject of processing. We call such data metadata. Today, metadata is usually presented by the semi-structured data model and implemented with the help of XML. Design and implementation of metadata have to be discussed separately. Above the tier of metadata, the IS may contain data that explains metadata: its structure, purpose and relationships. This tier helps to integrate data and functionality of the IS with other applications. Such data is called ontological data. It is modeled with the help of the object-oriented data model and implemented in XML technologies: RDF, OWL, and others. The design of such data and approaches to its implementation have to be discusses as well.
- "Design of Distributed Databases". In addition to the issues of physical design of data in databases, the course has to contain discussion of the physical model and implementation of XML files and ontologies. Comprehensive coverage of the latest database technologies will be supplemented by discussions of data technologies: XML, RDF, OWL, and others.
- "Data Warehouses". Data warehouses present an excellent opportunity to explore the different roles of data in the IS and discuss in more detail the design and implementation of different data parts of the system within one course. The course will cover data integration and exchange, metadata and ontological data, as well as technologies that support these different data components of the data warehouse.

Changing of the program's orientation at this point does not require radical changes of the courses. Discussion of databases and database technology will still dominate other issues. We want to emphasize though, that within the new program, coverage of data and data technologies will be more consistent, and from the beginning students will realize that they are working with the technology for a particular role of data in the IS.

The program can also be changed by adding alternative courses. We see that such courses can be added at the last, third level and offer specialization in different advanced topics. However, in the future, the program can develop subprograms, where specialization will start earlier.

New roles of data in the IS are defined by changes in perception of the IS architecture and interfaces between IS parts, which promoted new paradigms of IS and development of supporting technologies. This will also require changes in other computer programming courses.

References:

1. **T. Malyuta, "*Data on the Web*",** http://eportfolio.citytech.cuny.edu/~tmalyuta, Professional Development: Presentations, 2003**.**
2. **T. Malyuta, "*New Paradigm of the Information System*",** http://eportfolio.citytech.cuny.edu/~tmalyuta**,** Professional Development: Presentations, 2004.
3. **T. Malyuta, "*Semantic Web*",** http://eportfolio.citytech.cuny.edu/~tmalyuta, Professional Development: Presentations, 2006.
4. **Y. Dayen, T. Malyuta, "*Teaching Advanced Database Technology*,"** Workshop, ASEE conference, 2006.

5. **T. Malyuta, "*Physical Design and Implementation of Relational Databases—A Practical Approach*",** Pearson Custom Publisher, 2006.

Figures:

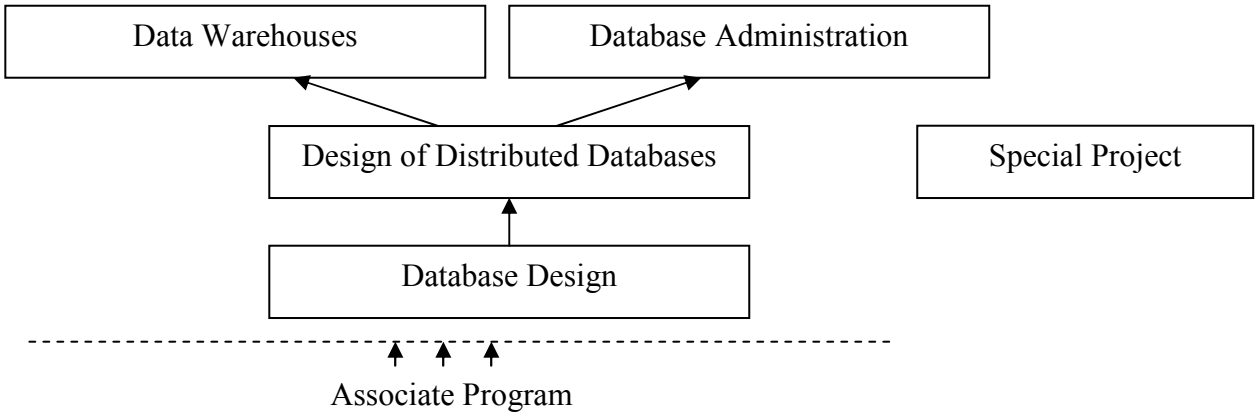

Figure 1. Structure of the database program.

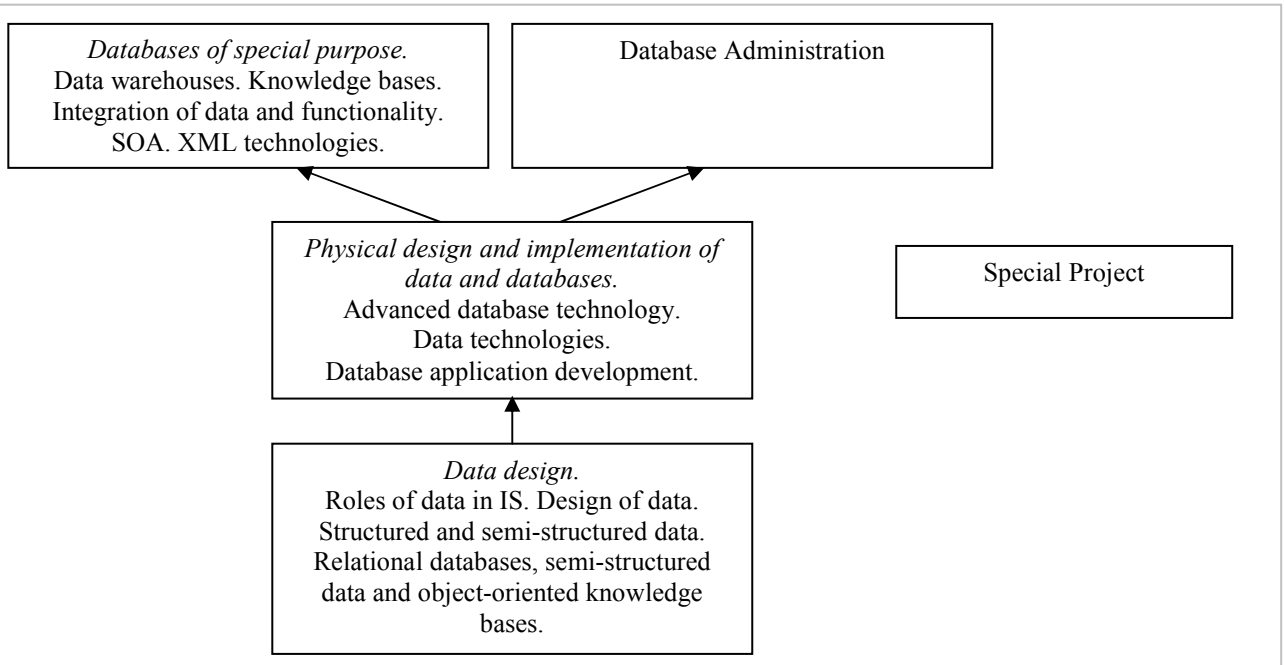

Figure 2. Development of the database program.