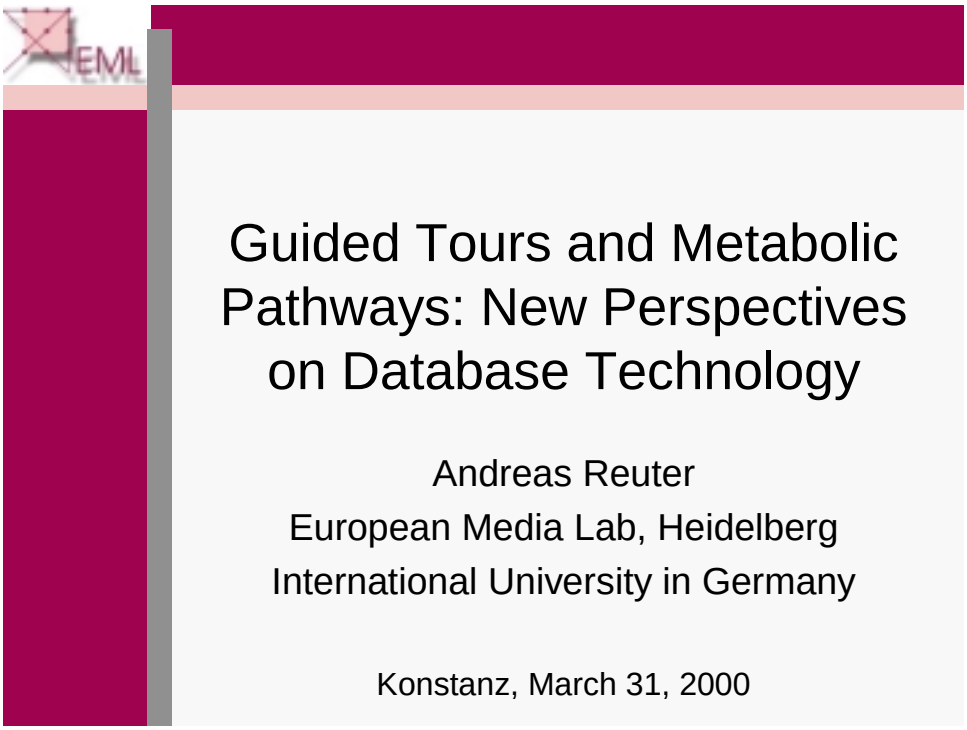
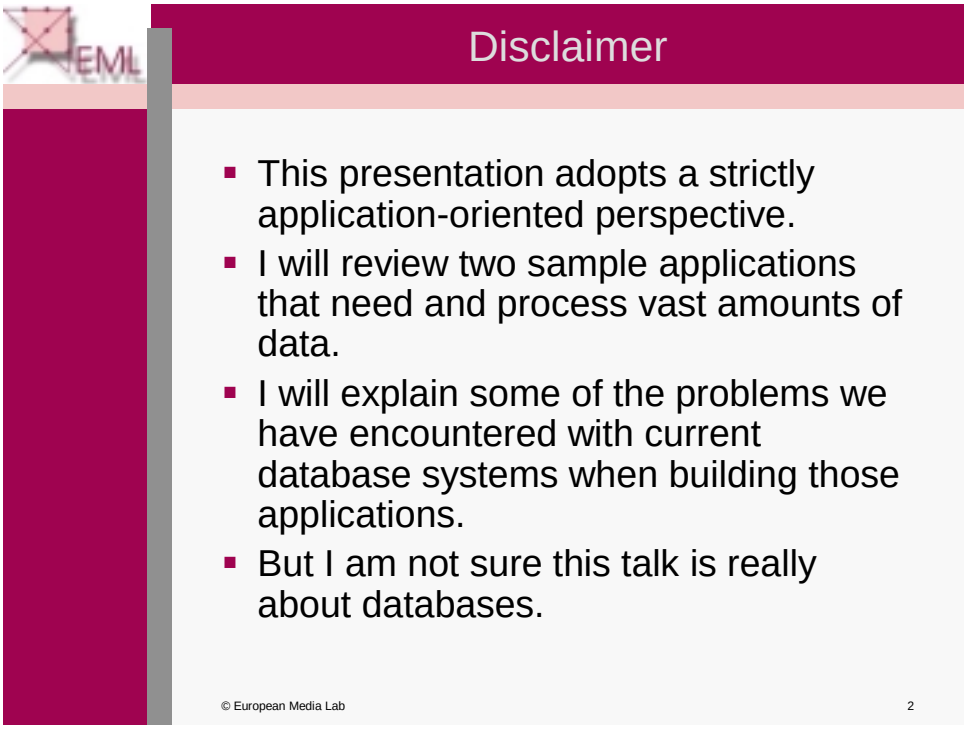


Guided Tours and Metabolic Pathways: New Perspectives on Database Technology

Andreas Reuter
European Media Lab, Heidelberg
International University in Germany

Konstanz, March 31, 2000



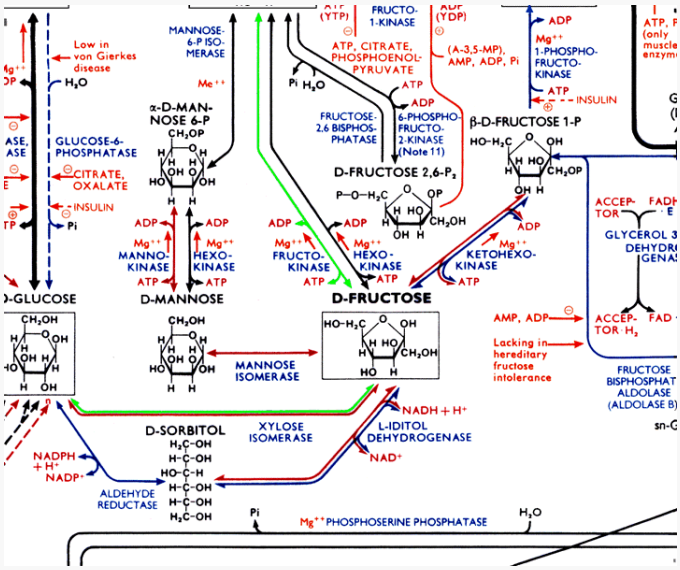
A Database System Should Understand About ...

	Agree	Disagree	
Accounts and transfers	☑	☐	OODBMS
Time	☑	☐	Temporal DBMS
Shapes	☑	☐	GIS
Compositional structure	☑	☐	OODBMS
Events	☑	☐	Active DBMS
Rules	☑	☐	Deductive DBMS
Exceptions	☑	☐	Deductive DBMS

Two Sample Applications

- The first example is a database that is to support simulation experiments in genetic research and bio-chemistry. It will contain information about static “facts” in this area, but also about kinetic properties of reactions.
- The second example is Deep Map, a portable tourist guide to Heidelberg. It will interact with the user via spoken language, visual displays, pointing devices and will be location-aware. Deep Map will contain information about the current situation of Heidelberg as well as historical data (events, persons, buildings, sights, etc.).

Metabolic Pathways



 5

Data for the Simulation of Metabolic Networks

- General data about biochemical reactions which stem from heterogeneous sources (different databases, formats, references, controversial data).
- Data about specific kinetic parameters which are typically published in paper form and not stored in databases.
- Making these data useful for simulations and other kinds of investigations requires a semantic model of the domain of discourse, which in this case is bio-chemistry, genetic research, micro-biology, etc.
- An complete semantic model for the domain of bio-chemistry and the subdomain of simulation can facilitate the retrieval of data from literature, the validation of data and the use of these data for the purpose of modeling and simulation of metabolic networks.

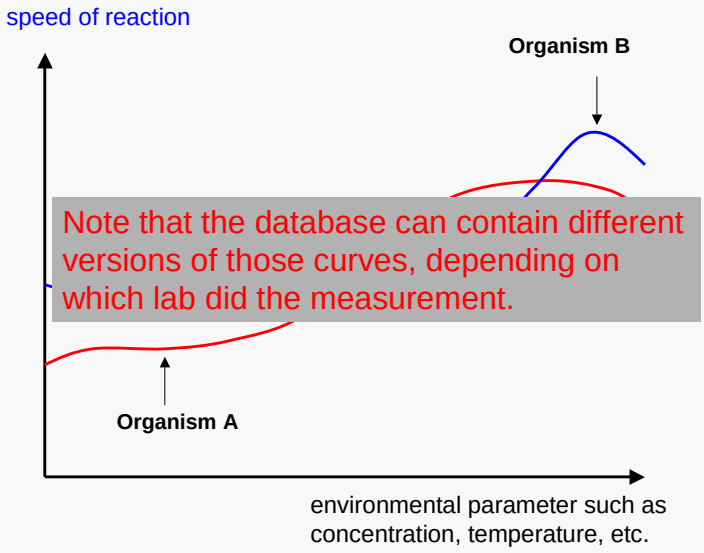
 6

PBE: Problems by Example

- There are a couple of hundred “databases” on genetic data, proteins, bio-chemical reactions etc. used and maintained in the research community and by companies.
- Most of these databases are flat files.
- There is no commonly agreed-on schema.
- The concepts underlying these databases are different.
- Often the really interesting information is “hidden” in text fields, further encoded by lab-specific jargon.
- But: Researchers would like to get transparent access to all these data sources.

 7

A Typical Bio-Chemical Fact



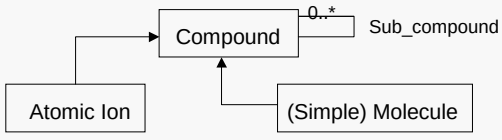
 8

Developing a Database Model for Biochemical Pathways

- The database definition should be based on knowledge of the characteristics of biochemical objects and their relations.
- The data model should be flexible, supporting the evolution of concepts and relationships, common in molecular biology.
- The system should offer the support for the retrieval and update of information from analysis and data management tools.
- The szstem should contain information about the origin of the data stored and references to related sources.
- The system should offer information that can help the scientist in the simulation of biochemical reactions, providing measured values of reaction and compound characteristics.

Some of the Difficulties Encountered

- Fuzzyness of definitions, e.g. gene, operon, among others.
- Exceptions and generalisations, e.g.:
  - "All organisms are mono- or multi-cellular" … and what happens with viruses? Are they organisms?
- Complex relations: e.g. a molecule is a compound, composed of atoms, when the atoms are charged they form atomic ions which can also act as compounds in a reaction.
- Behavioral constraints: a (complex) compound can have sub-compounds, but a molecule cannot be a sub-compound of a (simple) molecule.



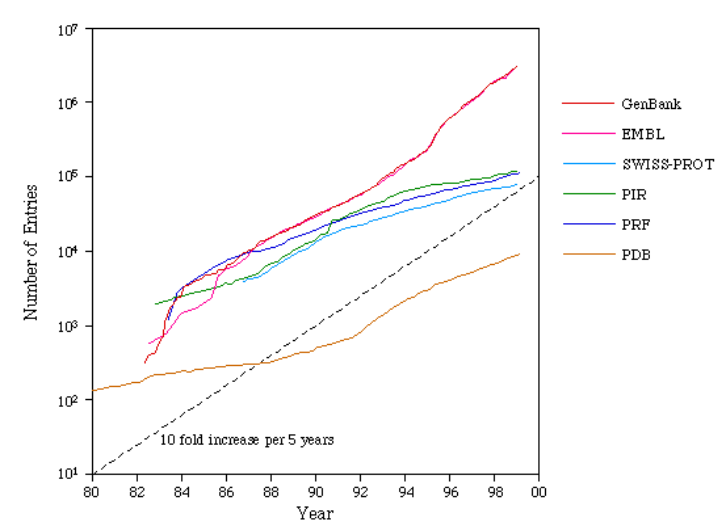
Some of the Difficulties Encountered

- Many of these “special cases” can be handled by introducing new sub-classes.
- This will lead to a complex structure of objects many of which will be hard to understand for the user.
- The other problem is that many of these exceptions and special cases are not static in the sense a schema is static.
- As bio-chemical research progresses, new concepts are introduced that are oblique with respect to the previous concepts. Some of them are abandoned, others are re-defined, etc.

Some of the Difficulties Encountered

- There is a constant need to define new concepts to correctly express the nature of the information being stored. A good example in case are genome fragments.
- Of course, all the existing information should not have to be re-formatted as a consequence of that.
- There is also the need to store meta-information about the basic data.
- There is the need to integrate data from multiple biochemical databases, each with different representations, completeness, scope, functionality and accuracy.

Growth of Bio-Chemical Databases



13

Our Current Approach

- We try to capture all relevant aspects: heterogeneity of sources, ambiguity of models, change in concepts and categories, need to extract information from text data by building an ontology for those parts of bio-chemistry that are needed for supporting the simulations.
- This ontology will be powerful enough to allow for reasoning and will express all the aspects that the schema itself cannot.
- The contents of the ontology can be used to automatically generate database programs for the relevant queries.

14

What is an Ontology?

- Formalization of domain content:
  - General (substances, processes, graphs, etc)
  - Specific (chemistry, reactions, pathways)
- Type and concepts - their definition, properties and relationships --- representation rigorous enough to support reasoning.
- Ontologies as semantic specifications.
- We could express facts like:
  - a simple molecule cannot be a sub-compound of an atomic ion
  - (⇒ (and (subcompound ?Comp1 ?Comp2) (Simplemolecule ?Comp1)) (not (AtomicIon ?Comp2))

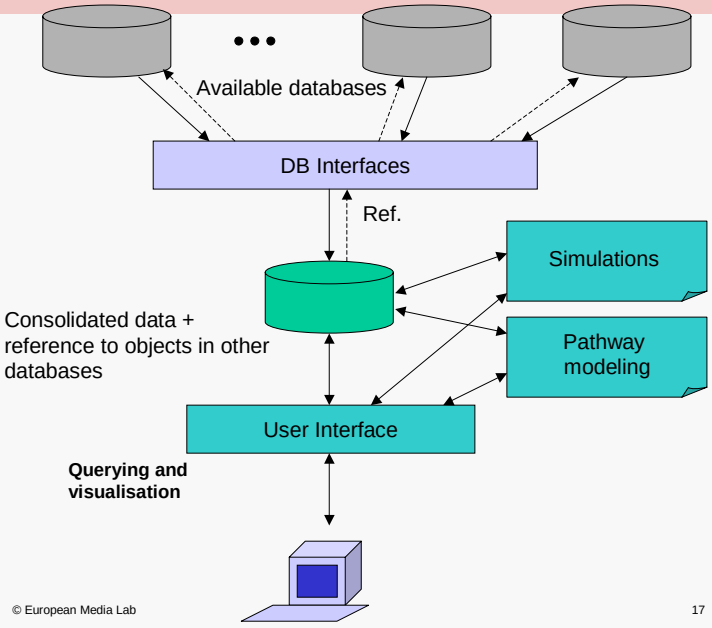
 15

Ontology-Based Models

- We cannot say that ontologies per se create better databases. A database designer has an ontology in his mind.
- Most of this ontology never makes it into the database design, mainly because of the lack of tools to express this in an easy and comprehensible manner.
- Ease of use is the key issue. DBMSs are not the problem.
- If the tools exist to turn ontologies into database applications, then database designers will be willing to put more of their (implicit) ontology into a design.
- Ontologies are hard to build, due to their level of generality. But that level of generality is just what makes them useful.

 16

Data Alive - ELSA



17

Purpose of the Simulation of Bio-chemical Processes

- The complexity of the biochemical network demands computer supported analysis of data and hypotheses.
- Simulations can validate models and thereby accelerate experimental work.
- Simulations are a useful and comparatively inexpensive tool for educating students.

18

Introducing the Second Application

Time for a movie!

Goal of the Deep Map Project

- Deep Map is going to be a portable device that can be used as a tourist guide in Heidelberg.
- It will know the preferences of the user and create appropriate tours.
- It will interact with the user through spoken language, it will display images, and it will have a pointing device.
- It will have a GPS or other location sensors, and it will have access to the Internet.
- It will provide information from a large number of databases in a variety of formats.
- Ideally, it will look like a camera and be as simple to use.

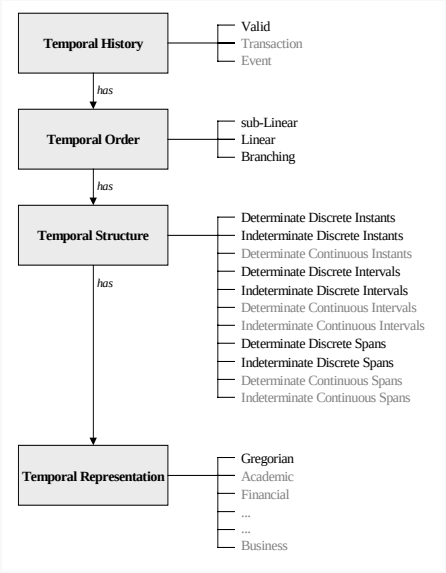
The Big Challenges in Deep Map

- How to handle time?
- How to handle topological and spatial references?
- How to recognize, maintain and exploit context?
- How to include the user's location (and his movements) into query handling and planning?
- How to optimize access to different databases, given the constraints of wireless connections?

Temporal Primitives

- discrete deterministic instants: April 22, 1975, September 13, 1993, 12:23h
- discrete non-deterministic instants: the moment of Caesar's death
- discrete deterministic intervals: January 1932 - October 1943
- discrete non-deterministic intervals: the siege of Vienna
- discrete deterministic time spans: 5 days; 23 years+2 months+12 days+7 hours
- discrete non-deterministic time spans: the length of the tourist season
- In addition there are descriptive references to time intervals which are inherently fuzzy: the middle ages, renaissance, the McCarthy era, etc.
- And: Is „the French revolution" an instant or an interval?

Relationships Among Temporal Aspects



 23

Additional Problems With Temporal Data

- When using object-oriented models, there is the problem of identity.
- Consider an object that exists for some time, gets destroyed, and then is re-built. Is it the same object? If you are not sure, consider the Frauenkirche in Dresden as an example.
- If an existing building is dedicated to a new purpose - is it the same building? As an example consider the Festival Hall in Baden-Baden.
- For such buildings, how would they be treated in a query such as: What is the oldest building in town?

 24

Problems with User Queries

As an example consider a tourist on one of Heidelberg's squares, after 3 hours of patiently following Deep Map's route suggestions.

All of a sudden she asks: "Is there a Pizzeria anywhere near?"

Which answer should the system generate?

More Technical Problems

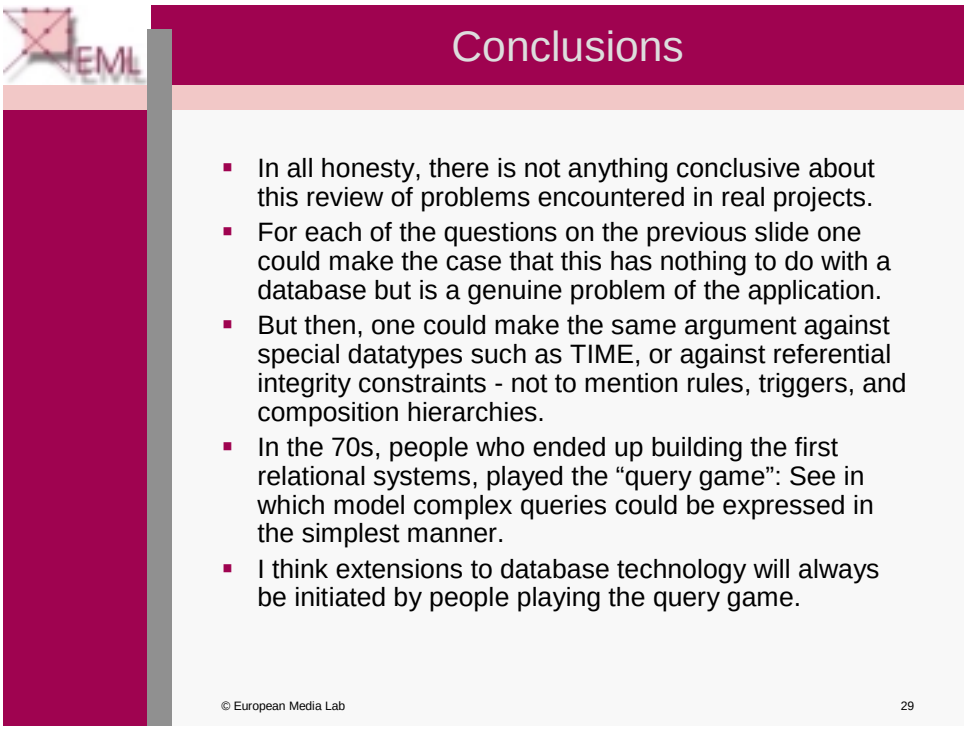
- For orientation purposes, the system has to answer queries based on visual input. Example: "What is this?"
- At the same time, the system has to display fairly complex visual information (renderings of old buildings etc.).
- All this requires data-intensive interaction with a server, which is a problem in a wireless setting.
- The system should do intelligent pre-fetching.

What Should a DBMS Have?

- Schema versions (temporal and otherwise)
- Interpolations, extrapolations, hypotheses
- Location awareness
- Integration of pub/sub mechanisms as basic functions
- The possibility to operate without a schema in some parts. Data are only organized by ontologies.
- Query modification and expansion based on history.
- Exceptions, controlled rule violations.

Let's Try Again: A DBS Should Understand About ...

	Agree		Disagree
History	☐	?	☐
Topology	☐	?	☐
Hypotheses and uncertainty	☐	?	☐
User intentions	☐	?	☐
Bio-chemistry	☐	?	☐
Laws of nature	☐	?	☐
Dynamic processes	☐	?	☐


EML
Conclusions
In all honesty, there is not anything conclusive about this review of problems encountered in real projects.
For each of the questions on the previous slide one could make the case that this has nothing to do with a database but is a genuine problem of the application.
But then, one could make the same argument against special datatypes such as TIME, or against referential integrity constraints - not to mention rules, triggers, and composition hierarchies.
In the 70s, people who ended up building the first relational systems, played the “query game”: See in which model complex queries could be expressed in the simplest manner.
I think extensions to database technology will always be initiated by people playing the query game.
© European Media Lab
29