# Virtual Integration of Heterogeneous Data and Data Model Unification

## Introduction: Basic Concepts

Sergey Stupnikov

Institute of Informatics Problems

Federal Research Center “Computer Science and Control“

Russian Academy of Science

sstupnikov@ipiran.ru

# Материалы к курсу

- https://synthesis.frccsc.ru/synthesis/student/BigData/master-course-integration/master-course-integration.html
  - synthesis.frccsc.ru ->
  - For MSU Students ->
  - Виртуальная интеграция неоднородных данных и унификация моделей данных

# Outline

- Course outline
- Interoperability
- Data, Information, and Knowledge
- Data Integration
  - Materialized Integration
  - Virtual Integration
- Subject Mediators
- Canonical Information Model
- Synthesis of the Canonical Model
- Unifying of a Data Model

# Course Outline (I)

❑ Виртуальная интеграция неоднородных данных и унификация моделей данных

- Технологии интеграции информационных ресурсов. Примеры систем интеграции ресурсов, сравнительный анализ
- Виртуальная интеграция в предметных посредниках: формальные основания. Переписывание запросов с использованием взглядов, поглощение запросов
- Каноническая информационная модель, исчисление спецификаций

# Course Outline (II)

- Виртуальная интеграция ресурсов в предметных посредниках: архитектура
  - Виды предметных посредников и методология интеграции неоднородных ресурсов в посреднике
  - Архитектура исполнительных механизмов (runtime) слоя предметных посредников
  - Инфраструктура предметных посредников для решения задач над множеством неоднородных ресурсов
- Синтез канонических моделей, унификация моделей ресурсов, метод доказательства сохранения информации и семантики операций при отображении моделей
  - Онтологические модели, OWL
  - Графовые модели
- Мультидиалектные инфраструктуры концептуальной спецификации и решения задач над неоднородными распределенными информационными ресурсами

# Отчетность

- Посещение лекций, семинаров секции SIGMOD, конференций 20%
- Домашние задания 30% (чем раньше, тем лучше)
- Экзамен 50%
- Ответы на вопросы на лекциях
- Факультативные задачи

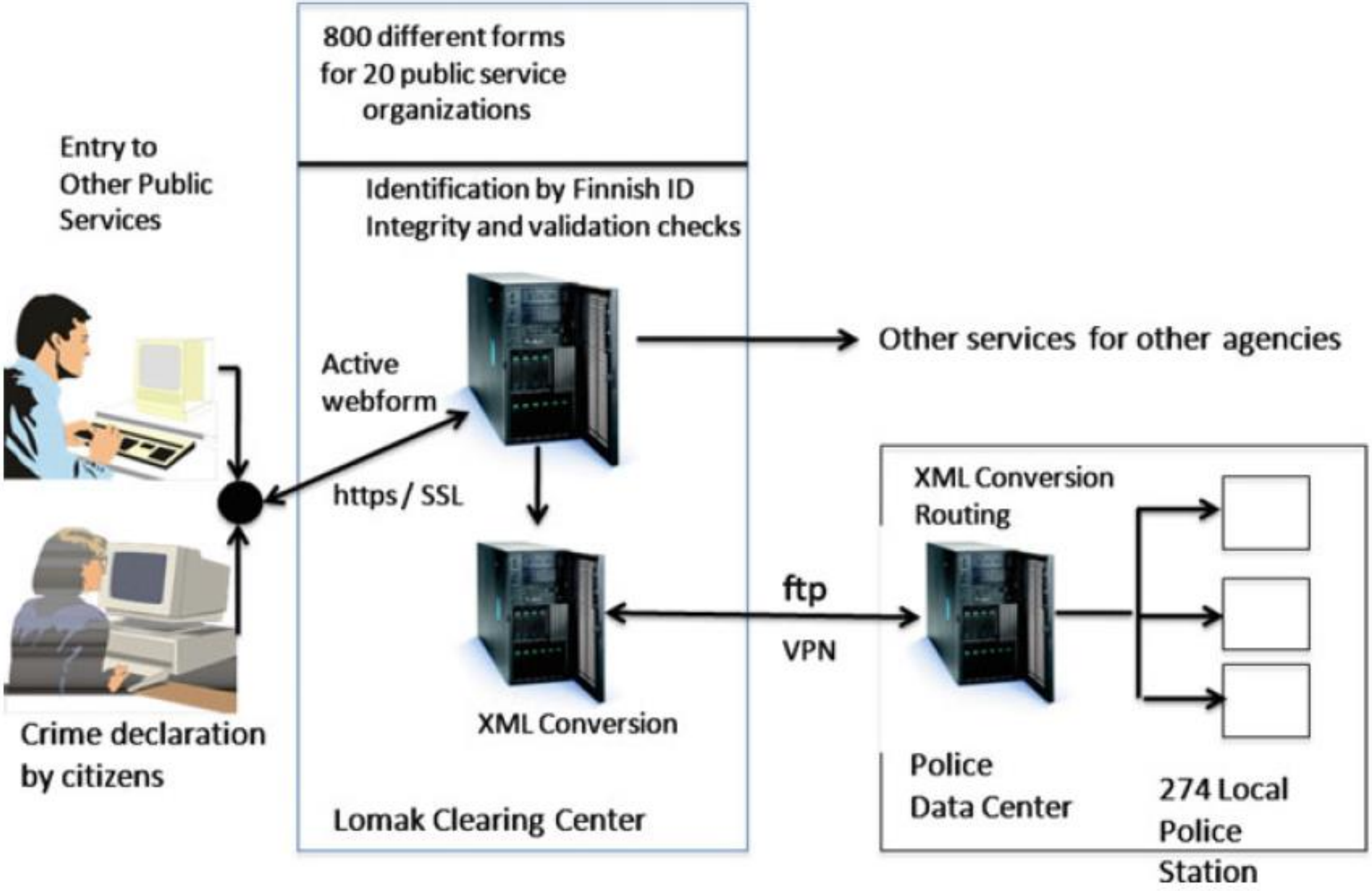
# Interoperability (I)

- **Interoperability** is the ability of making systems and organizations to work together (inter-operate)

- Technical interoperability
  - Systems are capable of communicating
  - . . . is usually associated with hardware/software components, systems and platforms that enable machine-to-machine communication to take place. This kind of interoperability is often centered on (communication) protocols and the infrastructure needed for those protocols to operate (European Telecommunication Standards Institute)
  - Protocols - TCP/IP, HTTPS, SMTP and SMIME, FTP and SSL
- Syntactic interoperability
  - Systems are capable of exchanging [of clearly defined classes of] data
  - . . . is usually associated with data formats. The messages transferred by communication protocols need to have a well-defined syntax and encoding (ETSI)
  - Data formats: HTML, XML, …

# Technical and syntactic interoperability in lomake.fi

Central Finnish portal for electronic forms in the public administration



# Interoperability (I)

- Semantic interoperability
  - Systems are capable
    - to combine received information with other information resources and
    - to process it in a meaningful manner
  - . . . is concerned with ensuring that the precise meaning of exchanged information is understandable by any other application that was not initially developed for this purpose (ETSI)
  - automatic recognition of the individual data exchanged
  - data becomes information (semantics is added)

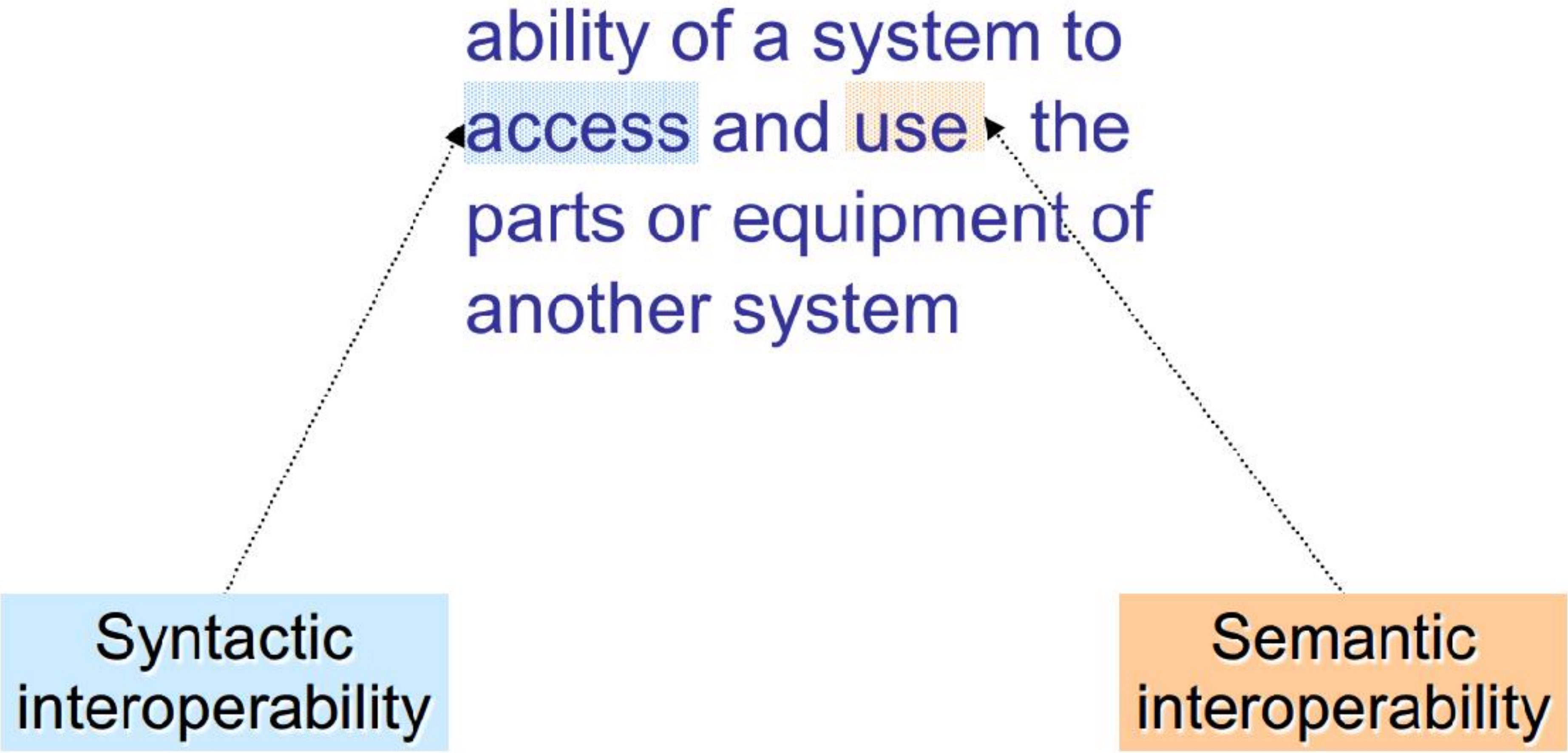
❑ Use case: Electronic invoices [date, number, price]

❑ Number semantics

- ❑ Universal Product Code (UPC)
- ❑ European Article Number (EAN)
- ❑ ISBN

❑ Solution: number → [code, number]

# Interoperability



ability of a system to access and use the parts or equipment of another system

Syntactic interoperability

Semantic interoperability

# Data, Information, and Knowledge (I)

- **Data** are the individual facts that are out of context, have no meaning, and are difficult to understand. They are often referred to as *raw data*
    - 3, 6, 9, 12
    - cat, dog, gerbil, rabbit, cockatoo
    - 161.2, 175.3, 166.4, 164.7, 169.3
- **Information** is a set of data in context that have meaning [with relevance to one or more people at a point in time or for a period of time]
    - When data is processed into information, it becomes interpretable and gains significance
        - 3, 6, 9 and 12 are the first four answers in the 3 x table
        - cat, dog, gerbil, rabbit, cockatoo is a list of household pets
        - 161.2, 175.3, 166.4, 164.7, 169.3 are the heights of 15-year-old students

# Data, Information, and Knowledge (II)

- **Knowledge** is cognizance (понимание, осведомленность), cognition (восприятие?), the fact or condition of knowing something with familiarity gained through experience or association. Knowledge is information that has been retained with an understanding about the significance of that information
  - *Explicit knowledge* - acquiring and remembering a set of facts
  - *Tacit knowledge* - the use of information to solve problems
  - Information + application or use = Knowledge

- Applying information to gain knowledge
  - 4, 8, 12 and 16 are the first four answers in the 4 x table (because the 3 x table starts at three and goes up in threes the 4 x table must start at four and go up in fours)
  - The tallest student is 175.3cm.
  - A lion is not a household pet as it is not in the list and it lives in the wi

# Data, Information, and Knowledge (II)

- Data
  - ↓ processing
- Information
  - ↓ acquiring facts, understanding how to solve problems: knowledge discovery
- Knowledge

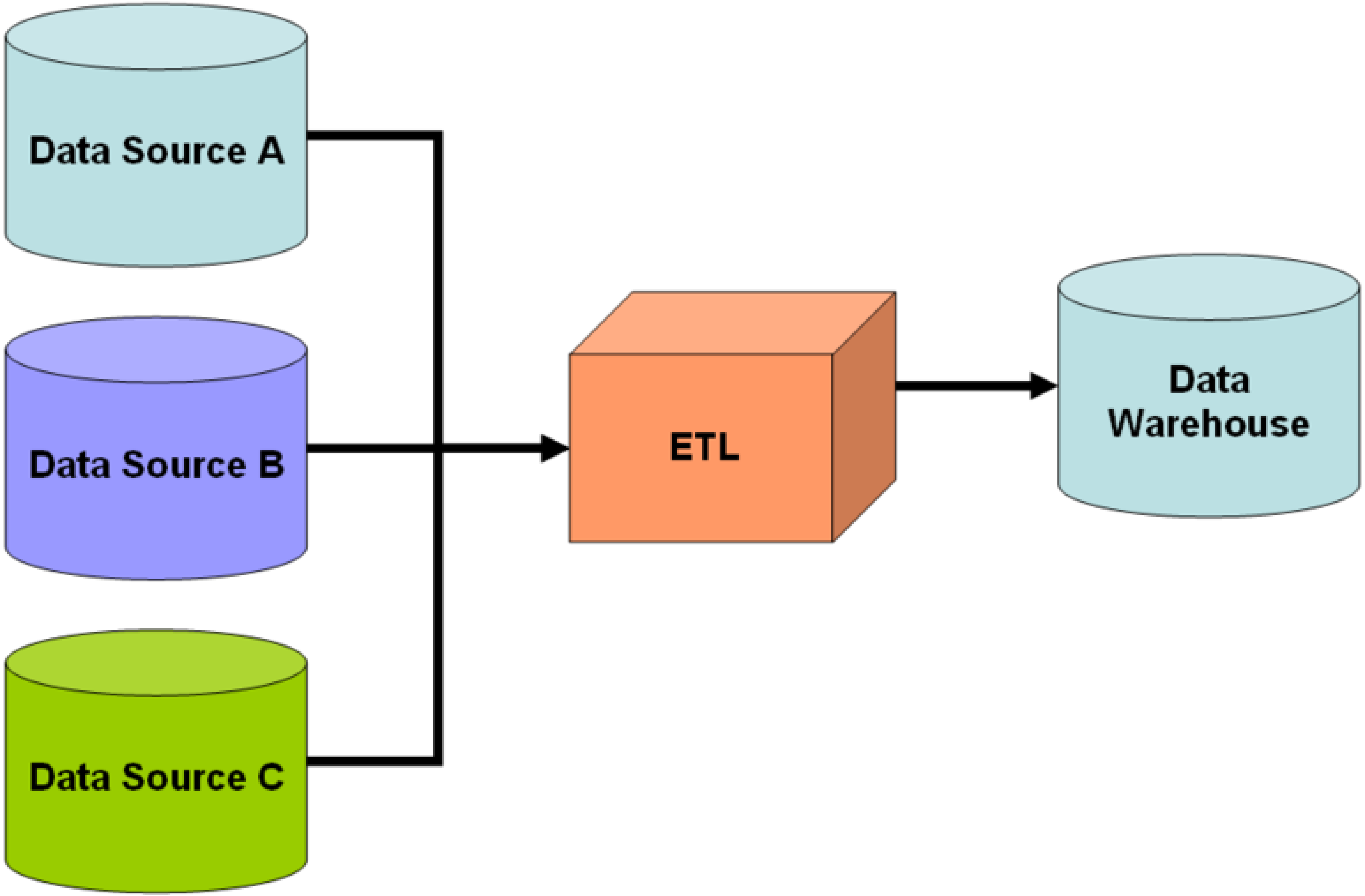
- ❑ Knowledge discovery
  - ❑ Data mining - finding patterns in large bodies of data and information using statistics, machine learning

➢ **?** Data, information, and knowledge within the machine learning pipeline

# Data Integration

- Combining heterogeneous data sources under a single query interface
  - Commercial use case: two similar companies need to merge their databases
  - Scientific use case: combining research results from different bioinformatics repositories
- Types of data integration
  - Materialized integration (Data warehousing)
  - Virtual integration

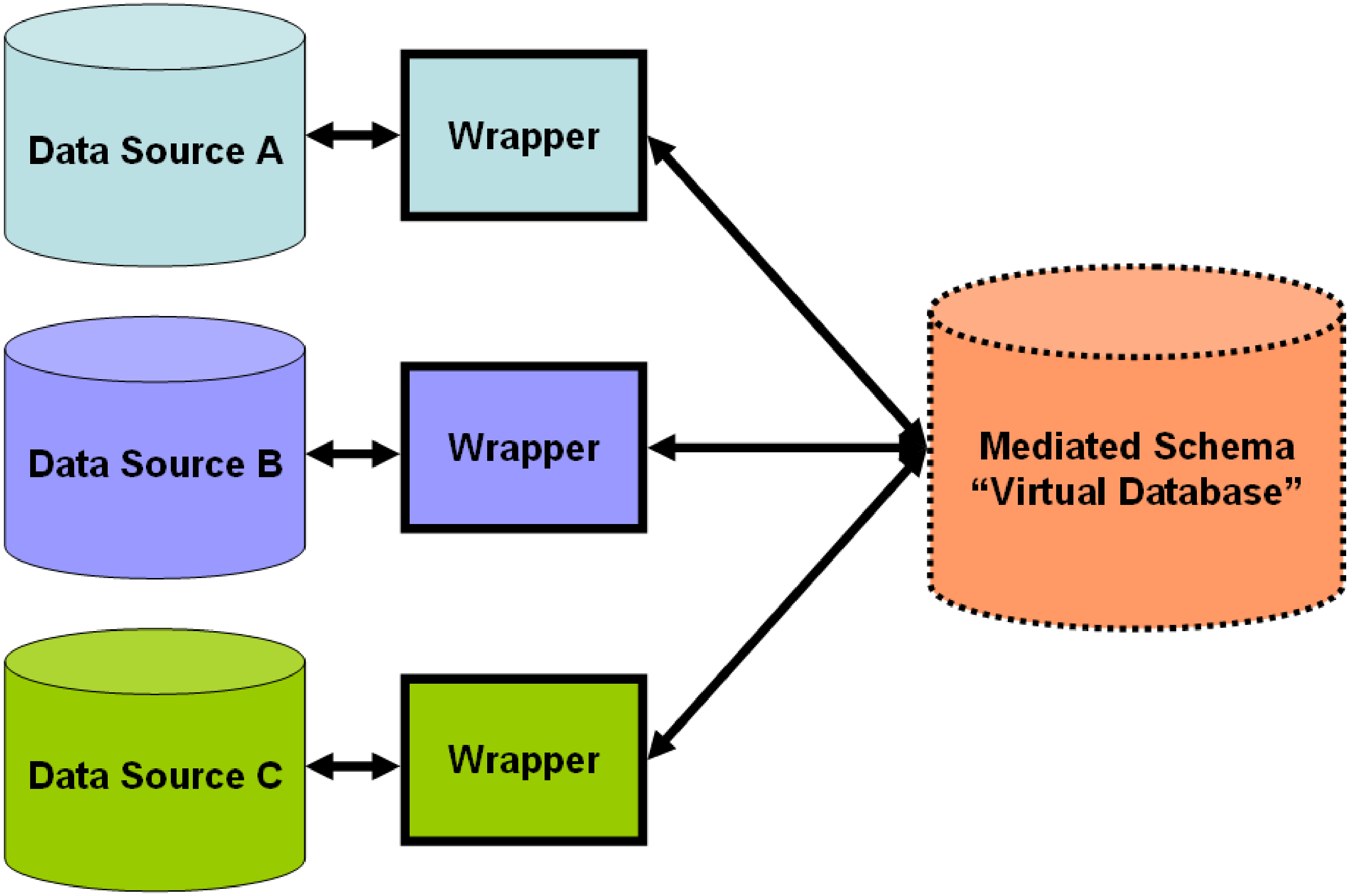
# Data Warehousing (I)



# Data Warehousing (II)

- **Data warehouse** – a database that consolidates data from multiple sources
- Each resource may have a DB schema that differs from the warehouse schema. So data has to be reshaped into common warehouse schema
- Extract-Transform-Load (ETL) tools
  - cleansing operations
  - reshaping operations

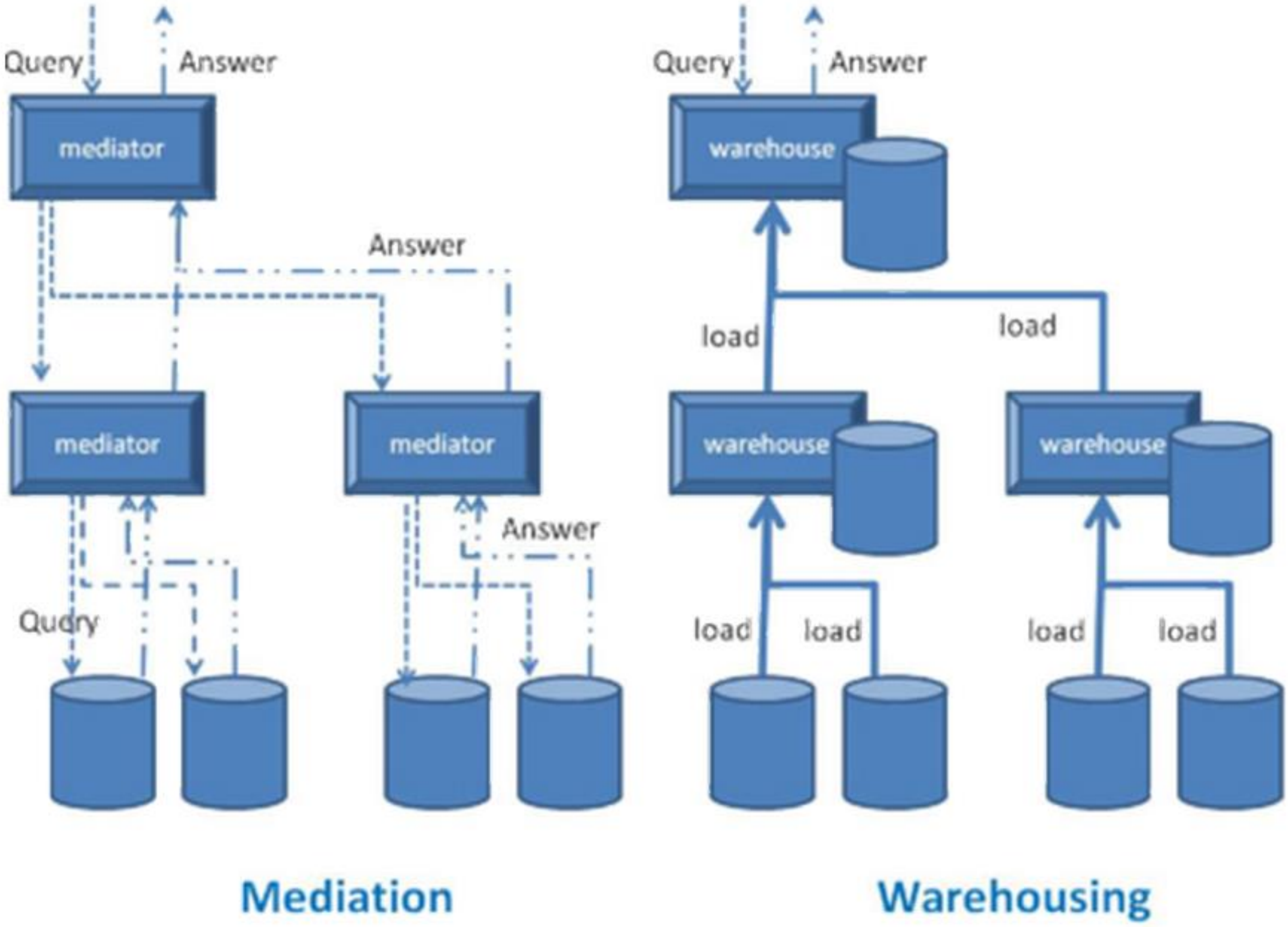
# Virtual Data Integration (I)



# Virtual Data Integration (II)

- Gives the illusion that data sources have been integrated without materializing data
- Offers a mediated schema against which users can pose queries
- The implementation, often called a query mediator system, translates the user’s query into queries over the data sources and integrates the result of those queries so that it appears to have come from a single integrated database
- Resources are heterogeneous in that they may use different database systems and structure the data using different schemas

# Virtual versus Materialized data integration

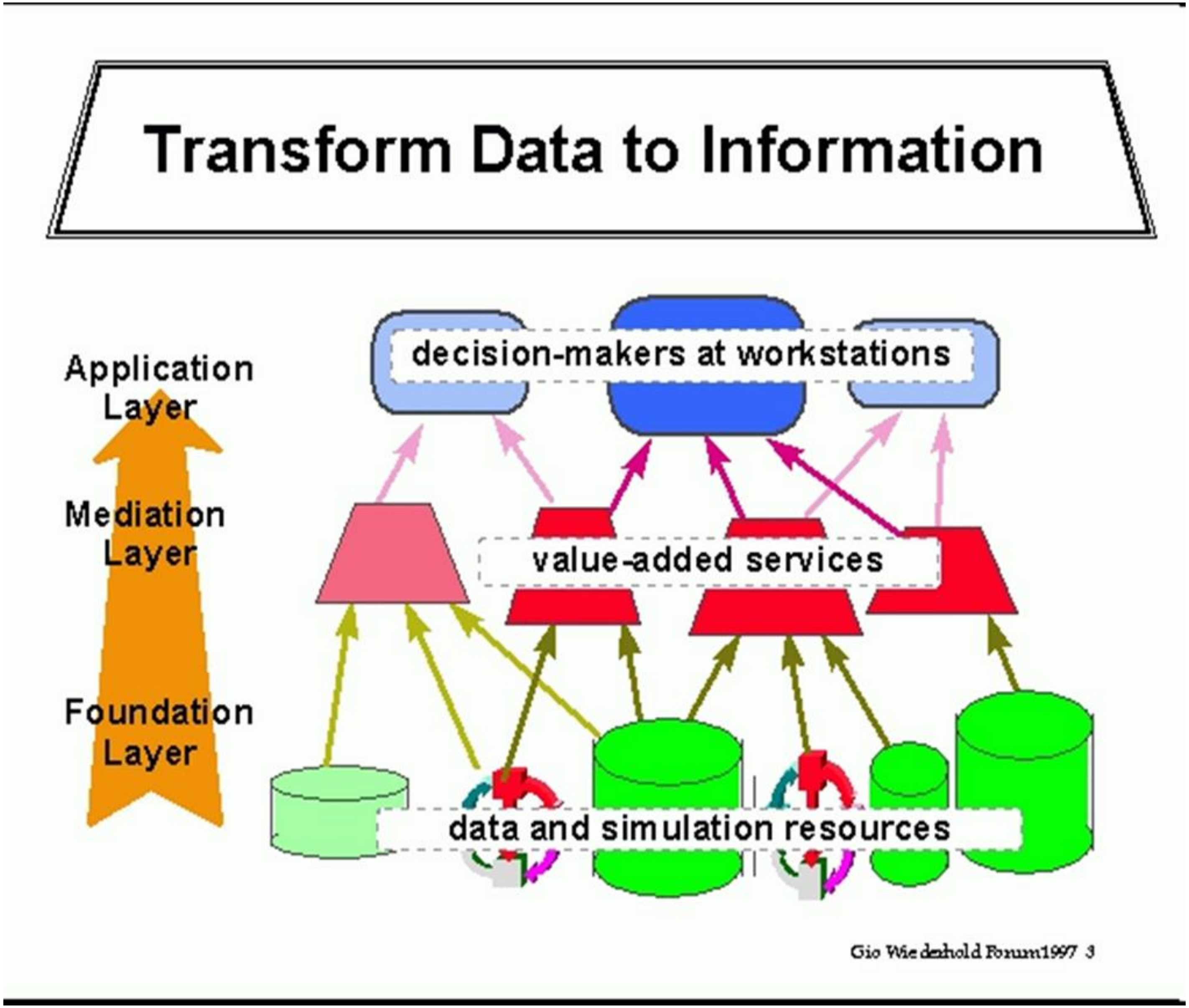


# Mediator according to Gio Wiederhold (Stanford Un.)

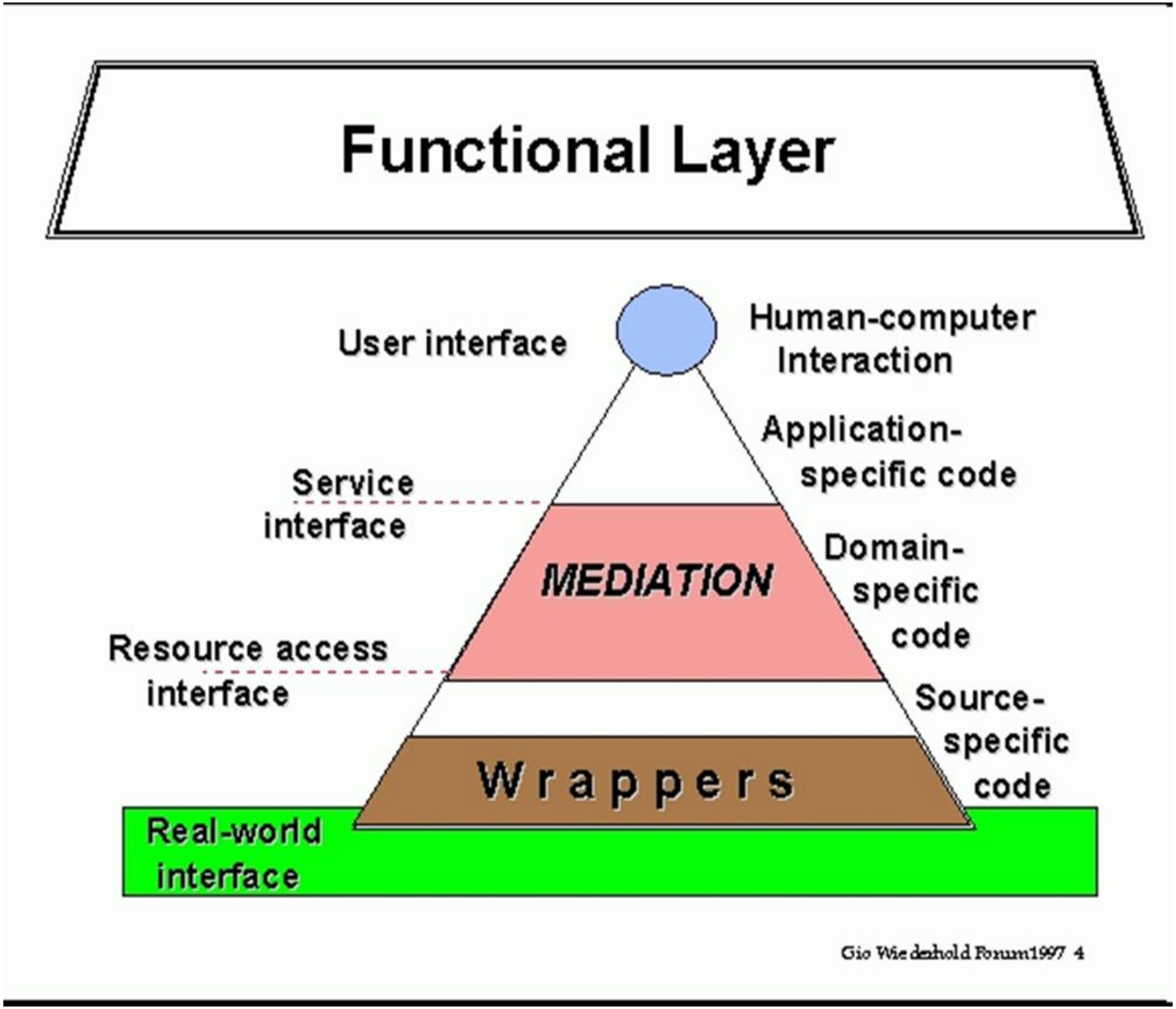
- **Data describes specific instances and events**. Data may gathered automatically or clerically. The correctness of data can be verified vis-a-vis the real world
- **Knowledge describes abstract classes**. Each class typically covers many instances. Experts are needed to gather and formalize knowledge. Data can be used to disprove knowledge
- A mediator is a software module that exploits encoded knowledge about some sets or subsets of data to create information for a higher layer of applications

- ❑ An important objective of the architecture is the ability to utilize a variety of information sources without demanding that they be brought into a common format and with only minimal requirements on their interfaces
- ❑ Mediator modules comprise a layer of intelligent middleware services in information systems linking data resources and application programs

# Subject Mediators (I)

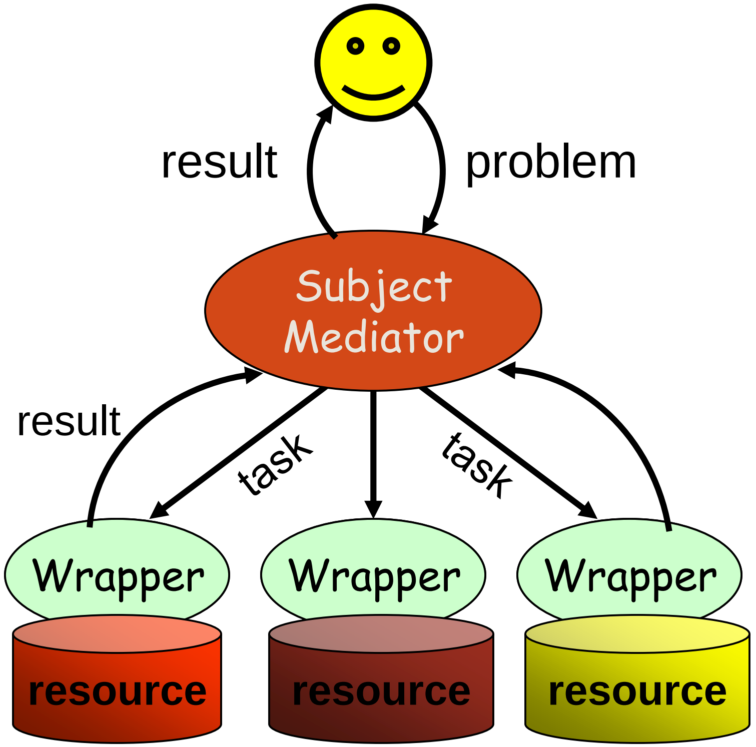

Transform Data to Information
Application Layer
Mediation Layer
Foundation Layer
decision-makers at workstations
value-added services
data and simulation resources
Gio Wiedeshold Forum1997 3

# Subject Mediators (II)



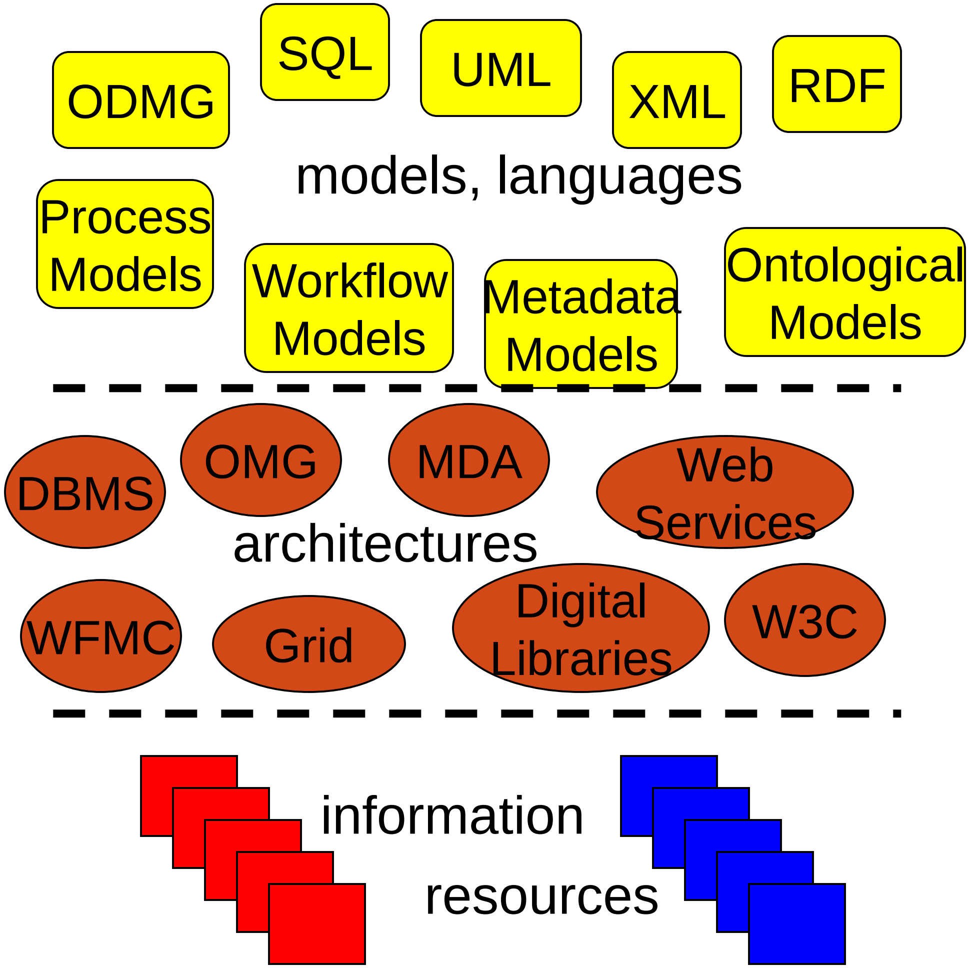
# Виртуальная интеграция в предметных посредниках

- Задача формулируется в терминах схемы посредника, затем
- трансформируется в набор подзадач (запросов) к ресурсам, зарегистрированным в посреднике;
- подзадачи исполняются на ресурсах, результаты возвращаются в посредник;
- результаты объединяются и представляются пользователю.



# Diversity of Information Models and Resources

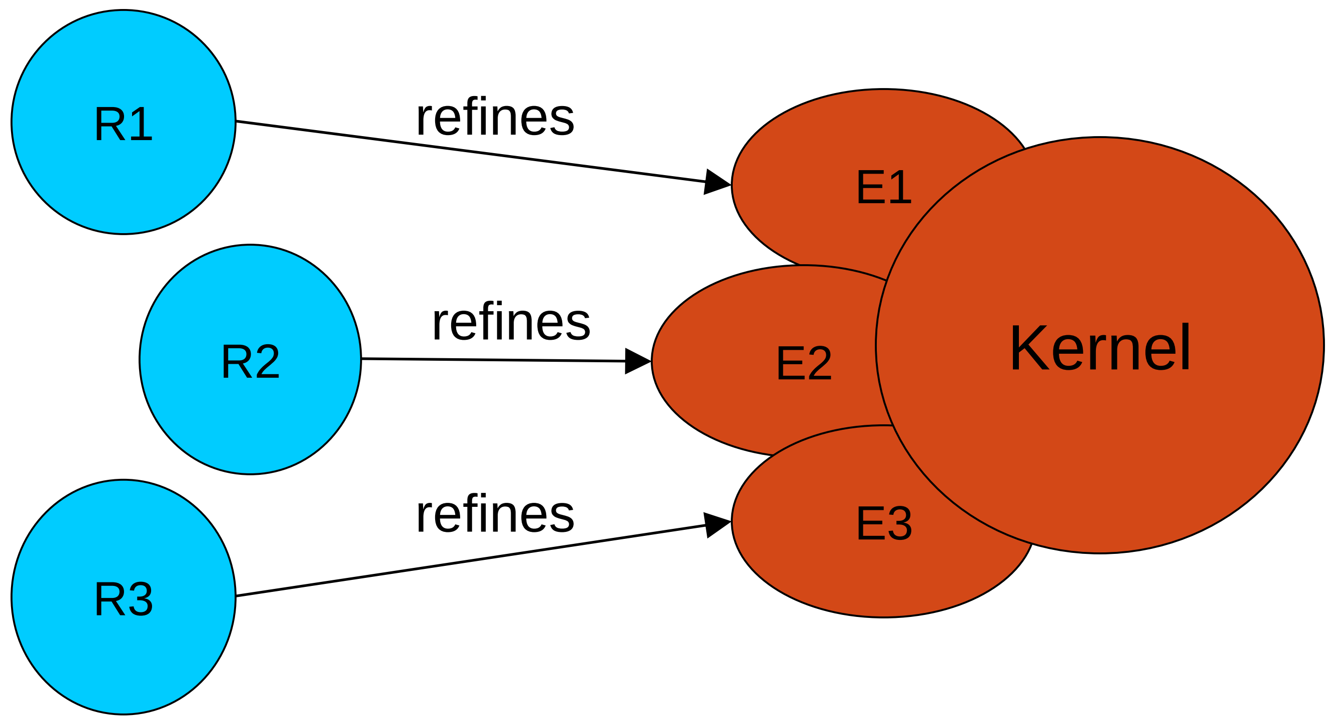
- **diversity of information models**
- **need for integration, reuse and composition of information resources**
- **accumulation of heterogeneous information resources**



# Canonical Data Model

- Canonical model as a set of language facilities sufficient for the IS conceptual modeling
- The canonical model plays a role of a unifying model, in which the resource information models can be represented without loss of information
- A transformation into the canonical model of the resource information models (languages) is required (to map resource schemas into the canonical model)
- Creating the transformations of the resource models into the canonical one (*resource models unification*) is a pre-requisite of resource schema mapping

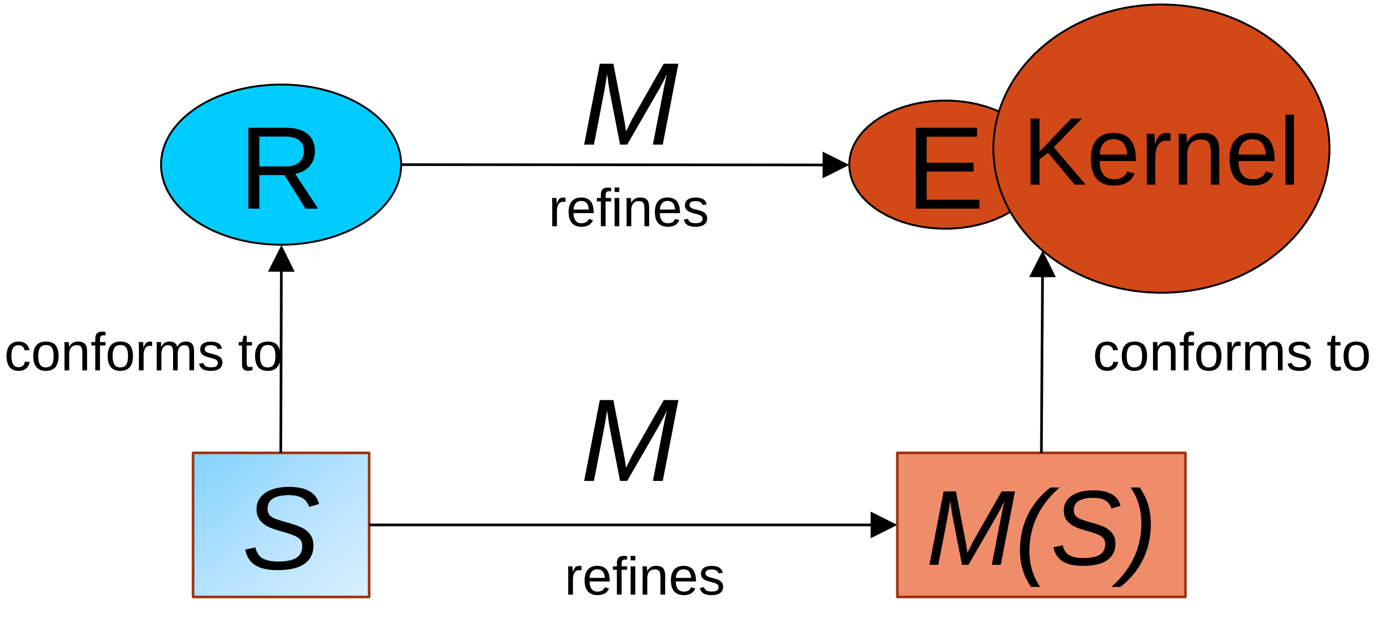
# Synthesis of the Canonical Model



*Source data models*

*Canonical Model*

# Unifying of a Data Model



- creation of extension **E** of the canonical model kernel
- creation of mapping ***M*** of a source model R into extended canonical one
- providing a possibility of proving that any ***S*** of R refines its image ***M(S)***