

CERIF COURSE

Session2: Use of CERIF

Keith G Jeffery,

Director, IT CLRC

k.g.jeffery@rl.ac.uk

Anne Asserson,

University of Bergen

anne.asserson@ub.uib.no

- CRISs and their Usage
- Database, Data and Metadata
- Data Exchange
- Data Access over Heterogeneous Sources
- Data Model for an 'ideal' CRIS
- CERIF in Use

- CRISs and their Usage
- Database, Data and Metadata
- Data Exchange
- Data Access over Heterogeneous Sources
- Data Model for an 'ideal' CRIS
- CERIF in Use

- Current
 - Research
 - Information
 - System
-
- Current = of current interest, not necessarily ongoing (e.g. Einstein)

CRISs and their Usage

The Shape of CRISs

- CRISs
 - Usually one major focus for entry
 - Project
 - Person
 - Organisational unit
- Implementation
 - various
 - IR Systems
 - Hierarchic Systems
 - Relational DBMS
 - Hypermedia Systems

CRISs and their Usage

The Use of CRISs

- Research Funding Administration
- Research Output recording / measurement
- Intellectual Property broking for technology transfer & wealth creation
- Funding Opportunities for R&D
- Expertise for consultancy or reviewing
- Research Partners for R&D
- Media awareness of R&D

- CRISs and their Usage
- Database, Data and Metadata
- Data Exchange
- Data Access over Heterogeneous Sources
- Data Model for an 'ideal' CRIS
- CERIF in Use

- Database, data and metadata
 - DATA, INFORMATION & KNOWLEDGE
 - DATA DELUGE, INFORMATION EXPLOSION AND METADATA
 - USAGE OF METADATA IN CRISs

- DATA : 06032002
 - representation of observation of real world
 - A lexical string of characters or symbols

- INFORMATION : 06-03-2002
 - USA: 3rd June 2002,
 - UK: 6th March 2002
- Instead use:
 - Data : 20020603
 - Metadata:
 - yyyyymmdd : a 'format template' (and ISO standard)
 - Date : a type
 - **Structured data in context**

- KNOWLEDGE
 - Theories or hypotheses
 - Representation of:
 - Facts (i.e. information)
 - Rules (when a, if b, then x, else y)
 - Processing of them by inference:
 - Deduction, induction, abduction
 - Commonly accepted justified belief

Database, Data and Metadata

DATA, INFORMATION & KNOWLEDGE

Knowledge: Facts

Start-Time	Departureairport	Flight	Arrivalairport	End-Time
0800	LHR	BA123	FRA	1000
0900	LHR	BA125	FRA	1100
1000	LHR	BA127	FRA	1200
1100	LHR	BA129	FRA	1300
Etc etc				
1800	LHR	BA137	FRA	2000

Database, Data and Metadata

DATA, INFORMATION & KNOWLEDGE

Knowledge: Induction

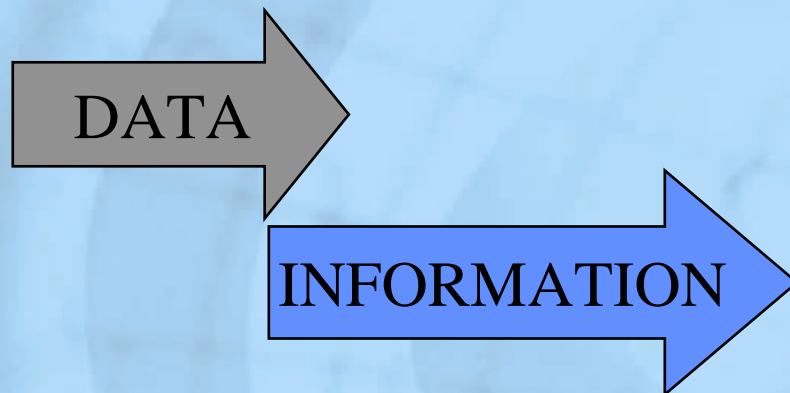
Start-Time	Departureairport	Flight	Arrivalairport	End-Time
0800	LHR	BA123	FRA	1000
0900	LHR	BA125	FRA	1100
1000	LHR	BA127	FRA	1200
1100	LHR	BA129	FRA	1300
Etc etc				
1800	LHR	BA137	FRA	2000

- between 0800 and 1800 **INDUCTION**
- every hour, on the hour (data mining)
- a BA flight leaves LHR for FRA

Collecting Observed Facts



Structuring in Context

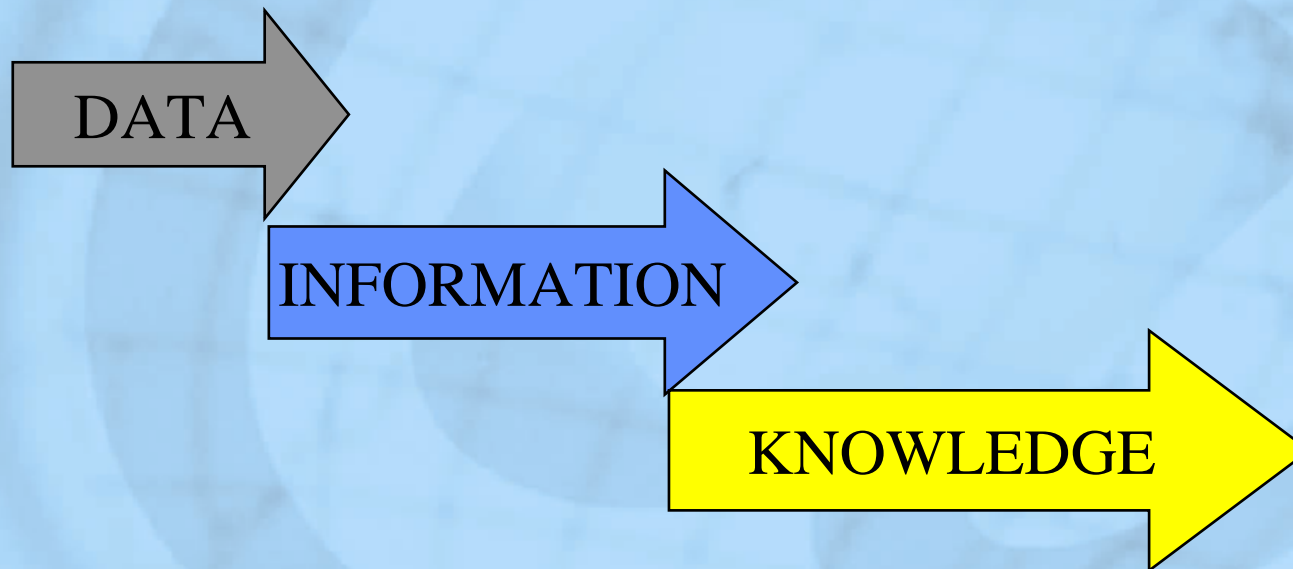


Database, Data and Metadata

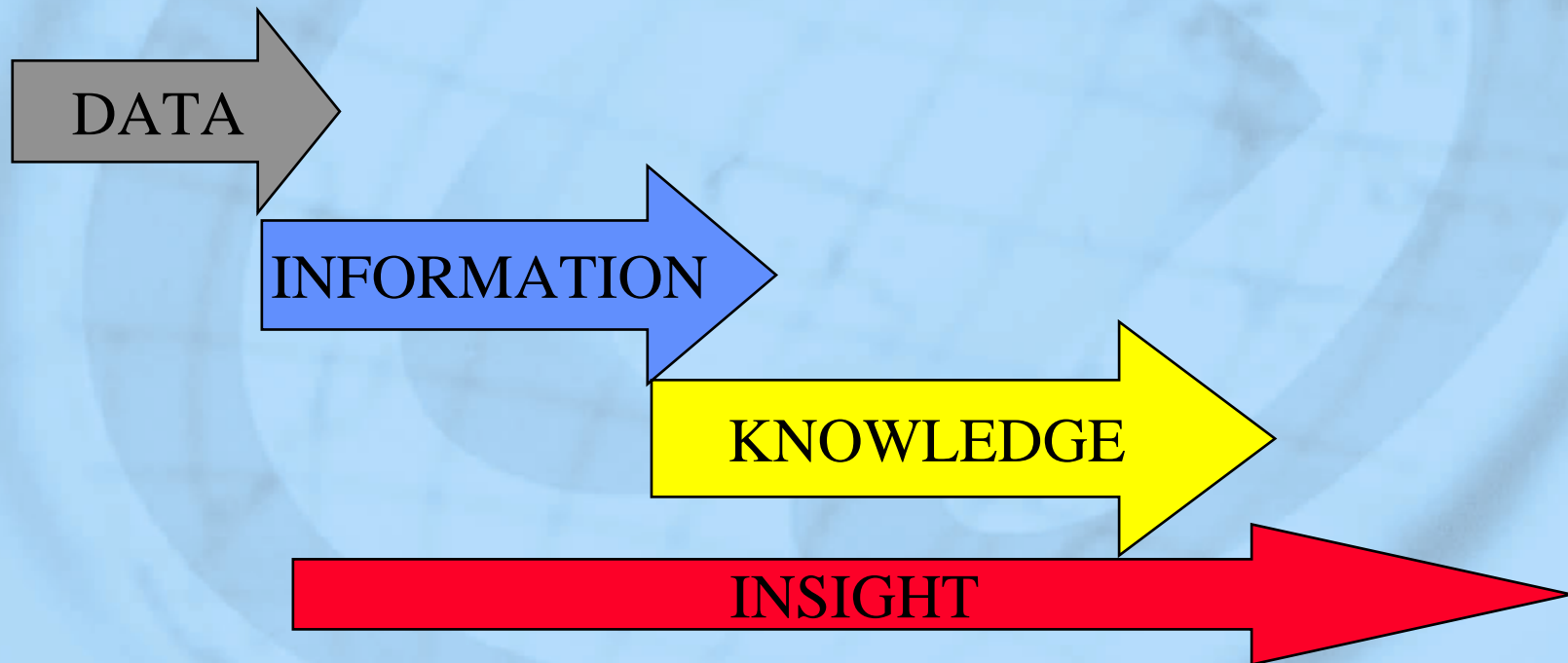
DATA, INFORMATION & KNOWLEDGE

Putting it together

Inducing commonly accepted belief



Value-Adding for Business Needs



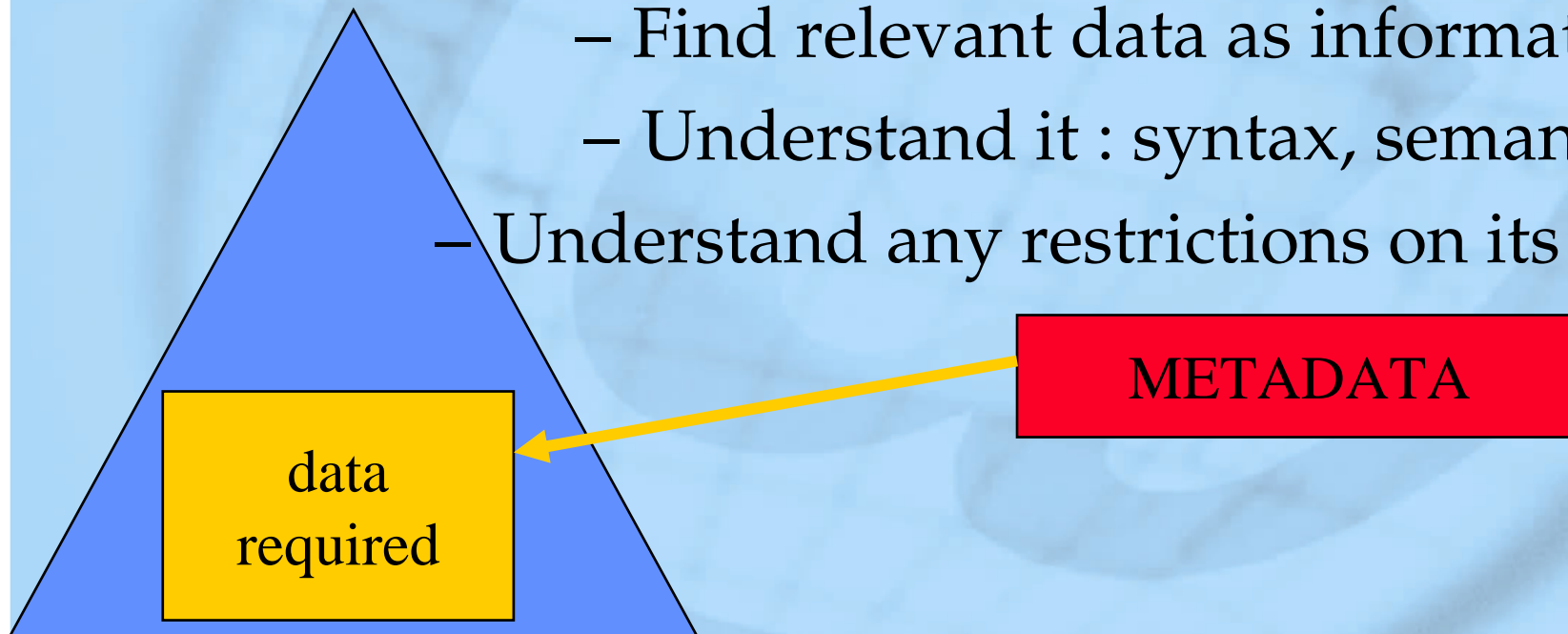
- Database, data and metadata
 - DATA, INFORMATION & KNOWLEDGE
 - DATA DELUGE, INFORMATION EXPLOSION AND METADATA
 - USAGE OF METADATA IN CRISs

Database, Data and Metadata

DATA DELUGE, INFORMATION EXPLOSION AND METADATA

Data & Metadata

- Much of this data is inaccessible
- Need to be able to
 - Find relevant data as information
 - Understand it : syntax, semantics
 - Understand any restrictions on its use

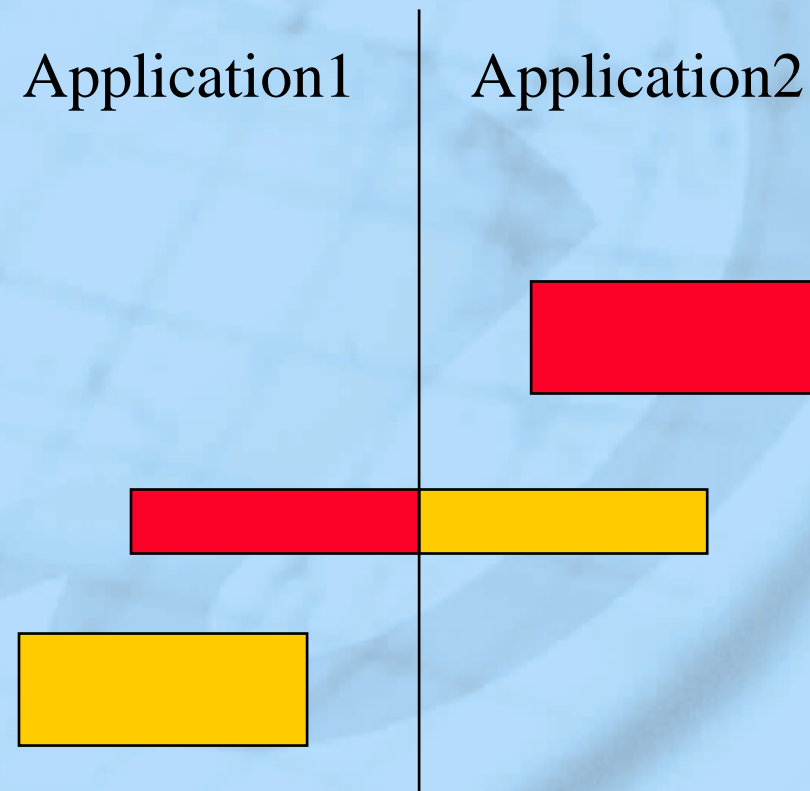


Database, Data and Metadata

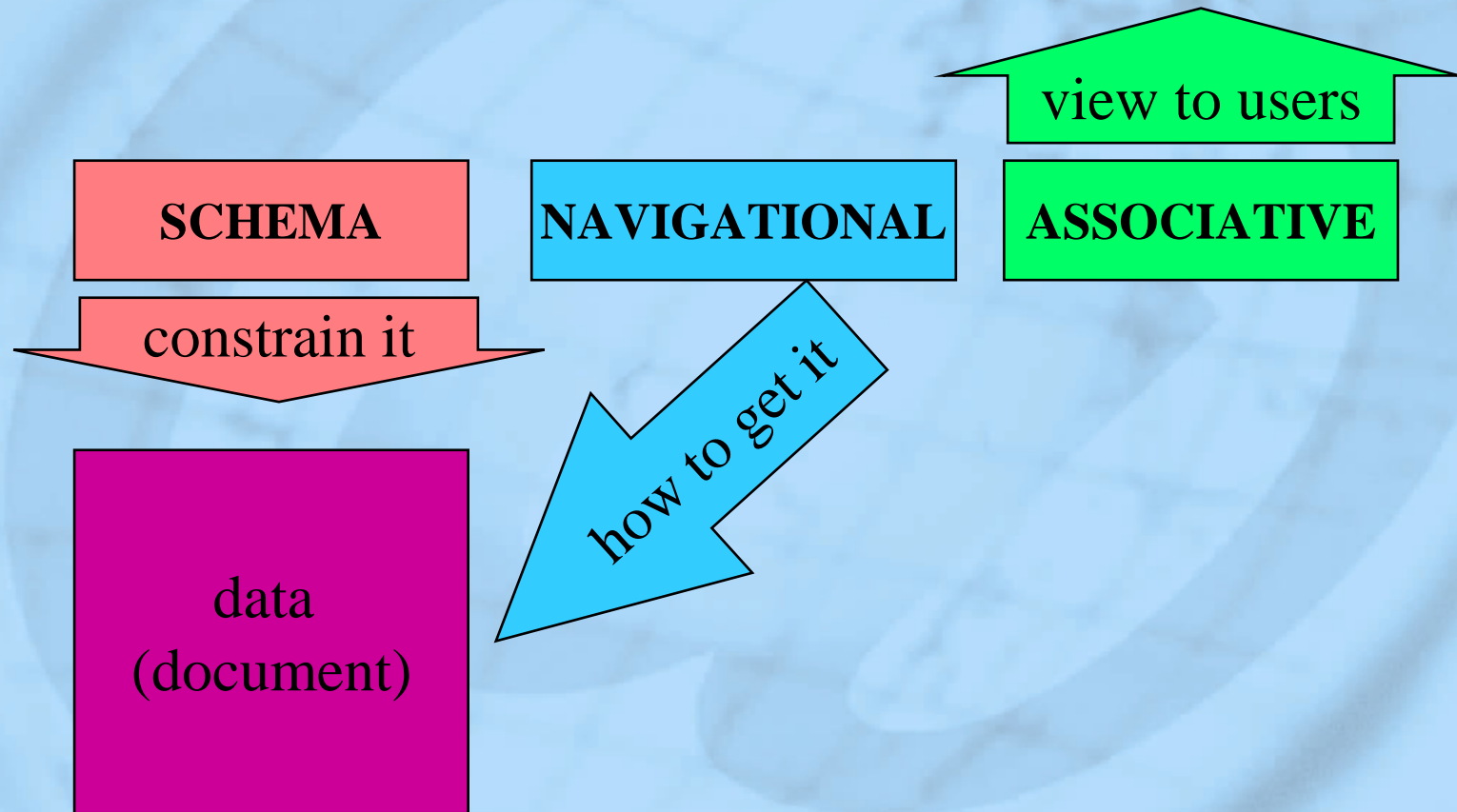
DATA DELUGE, INFORMATION EXPLOSION AND METADATA

Data & Metadata

- **Metadata** is data about **data**
- **Metadata** to one application is **data** to another



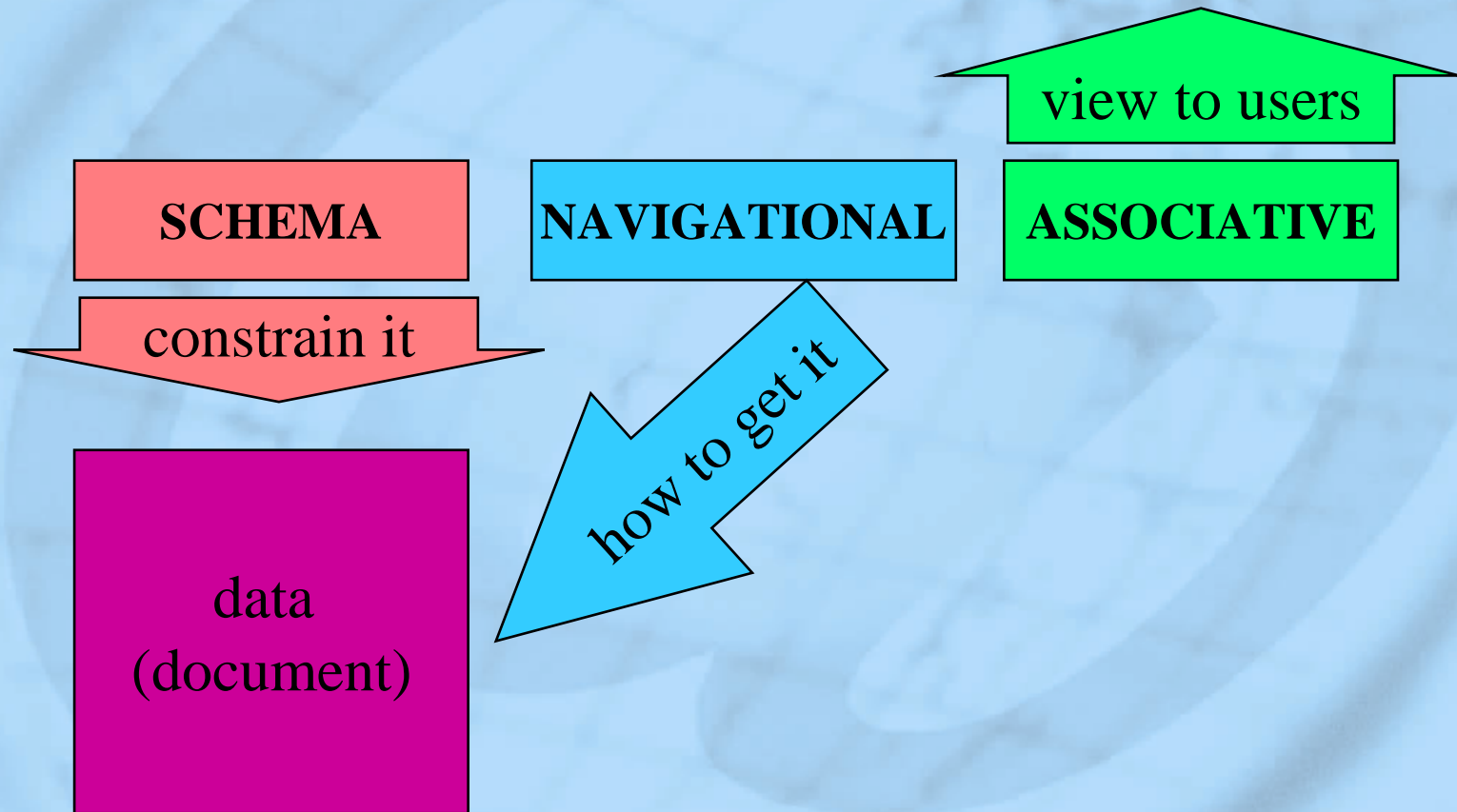
Three Kinds of Metadata



Metadata Kinds: Schema

- intensional description of extensional instances
 - database:
 - name
 - size
 - security authorisations
 - attributes:
 - name
 - type
 - constraints
- **formal logic relationship to data instances**

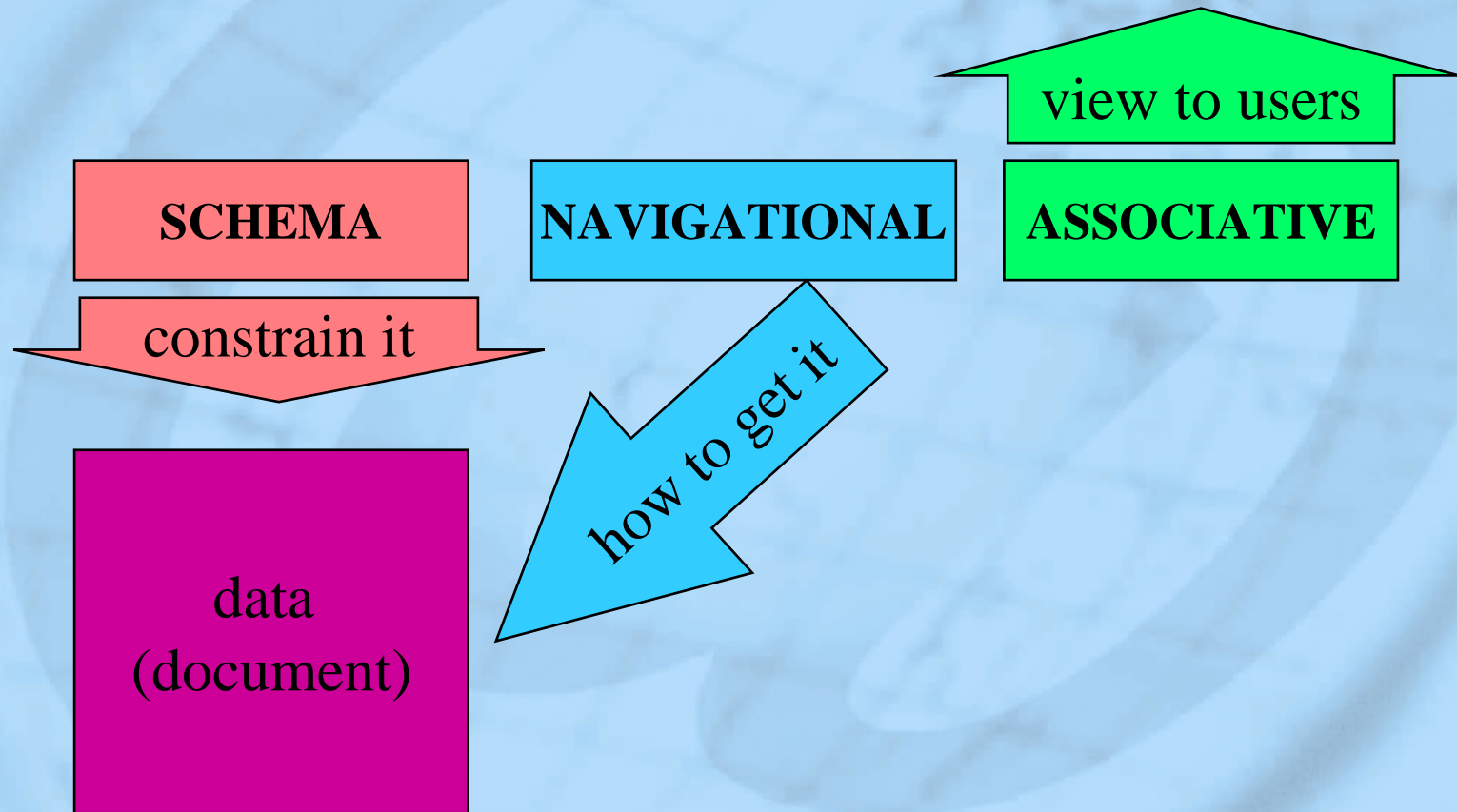
Three Kinds of Metadata



Metadata Kinds: Navigational

- How to get to information resource direct
 - filename
 - DB name + navigational algorithm
 - DB name + predicate (query)
 - URL
 - URL + predicate (query)
- or any of the above via
 - web indexing system (eg AltaVista, ExCite...)
 - local indexing system bookmarks or proxy server)

Three Kinds of Metadata



- Database, data and metadata
 - DATA, INFORMATION & KNOWLEDGE
 - DATA DELUGE, INFORMATION EXPLOSION AND METADATA
 - USAGE OF METADATA IN CRISs

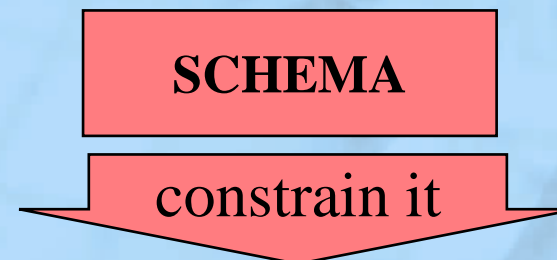
- Data quality
- Access
- Understanding answers
- Improving Queries
- Interoperability with other CRISs
- Interoperability with other Systems e.g.
 - Local management information systems
 - Bibliographic systems
 - Scientific data systems

Database, Data and Metadata

USAGE OF METADATA IN CRISs

Schema Metadata

- All CRISs based on
 - DB SYSTEM
 - IR SYSTEM
- Have schema metadata
- It may not be sufficient
 - To ensure integrity
 - To provide rich enough program interface
 - To ensure integrity in foreign key - primary key linkage to associated CRISs or other systems



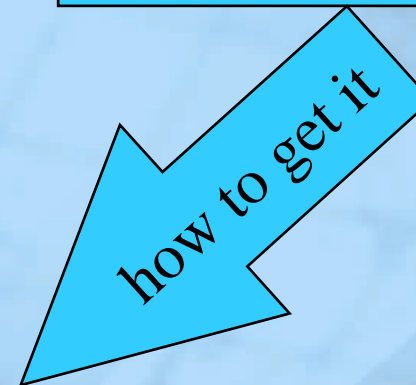
Database, Data and Metadata

USAGE OF METADATA IN CRISs

Navigational Metadata

- 'Base CRISs' may have navigational metadata
 - If provide raw information only: no
 - If provide URLs to e.g. publications, scientific datasets: yes
- 'Meta-CRISs' which act as catalogues or indexes to other CRISs do have navigational metadata

NAVIGATIONAL

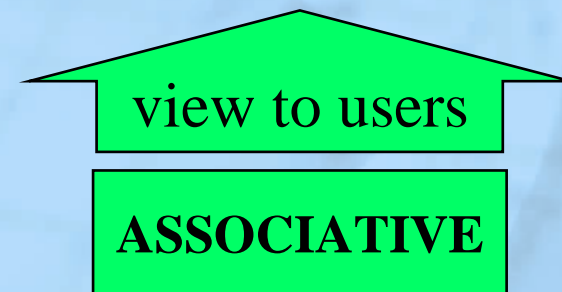


Database, Data and Metadata

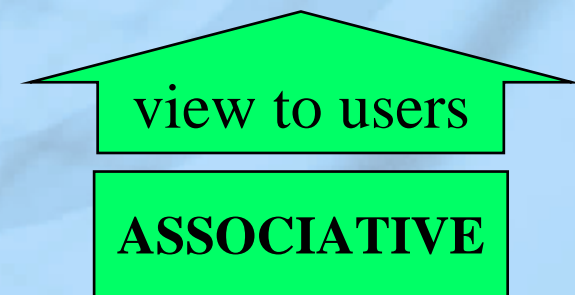
USAGE OF METADATA IN CRISs

Associative Metadata

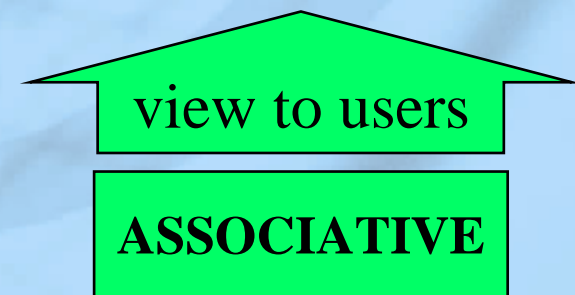
- AdM
 - Associative descriptive
- ArM
 - Associative restrictive
- AsM
 - Associative supportive



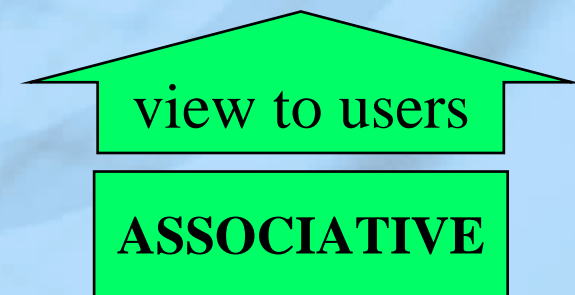
- CRISs have AdM if
 - Provide summary record of ≥ 1 {<project> | <person> | <orgunit>} and point to detailed records
 - The AdM provides machine-readable (syntax) and machine-understandable (semantics) information



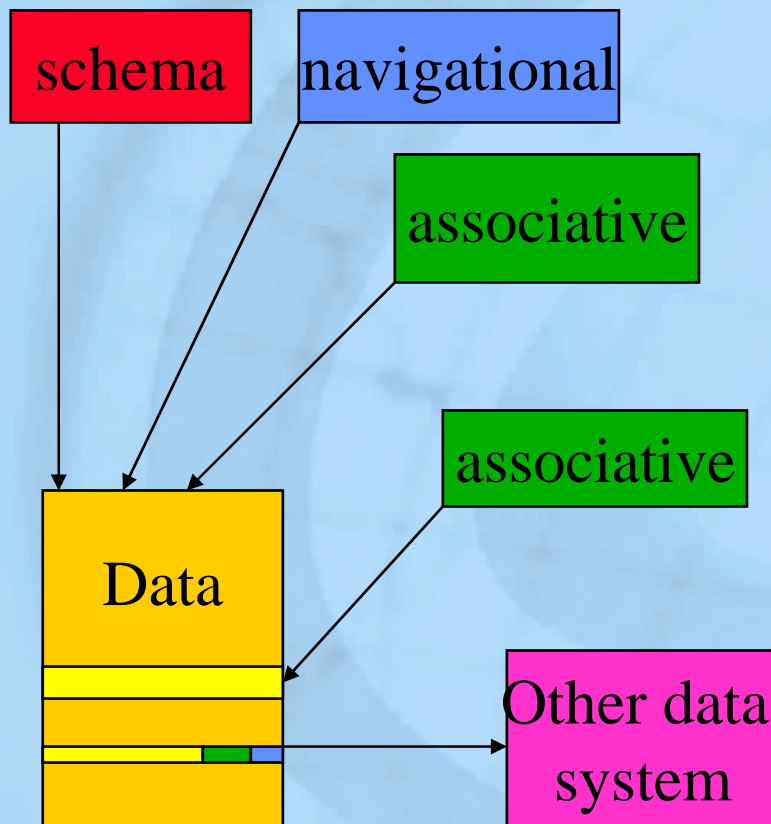
- CRISs have ArM if
 - Provide separate metadata record with information on access rights, copyright, IPR, 3rd party liability disclaimer, pricing
 - The ArM provides machine-readable (syntax) and machine-understandable (semantics) information



- CRISs have AsM if
 - Provide ≥ 1 {dictionary | hyperglossary | thesaurus | domain ontology}
 - The AsM provides machine-readable (syntax) and machine-understandable (semantics) information and / or knowledge



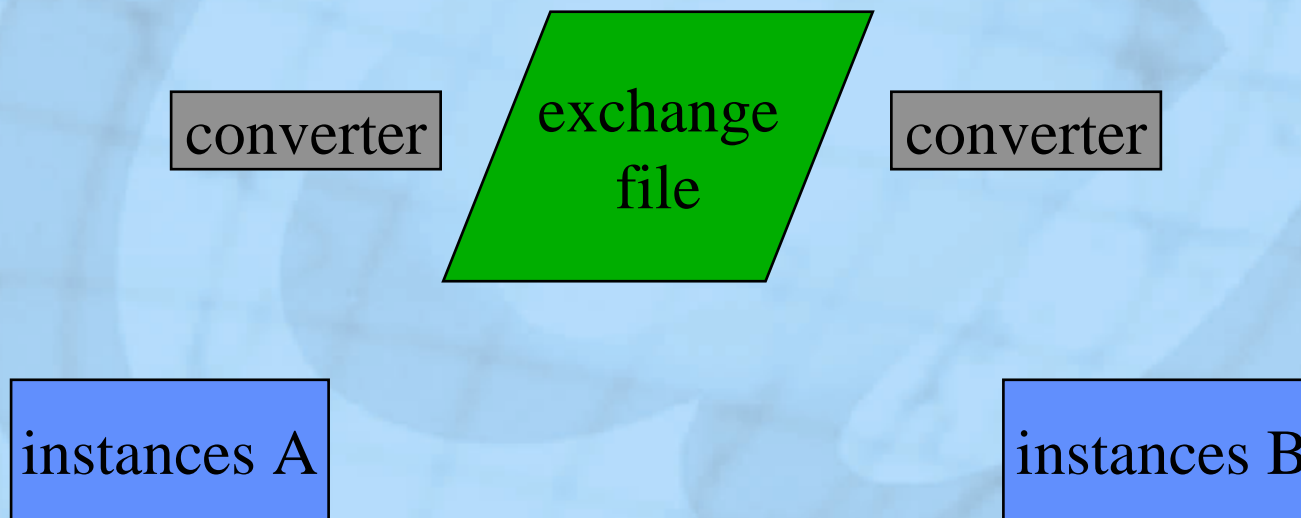
Typical CRIS and Metadata



- Metadata for whole collection of base CRIS data records
- Metadata for data record in base CRIS
- Metadata within base CRIS

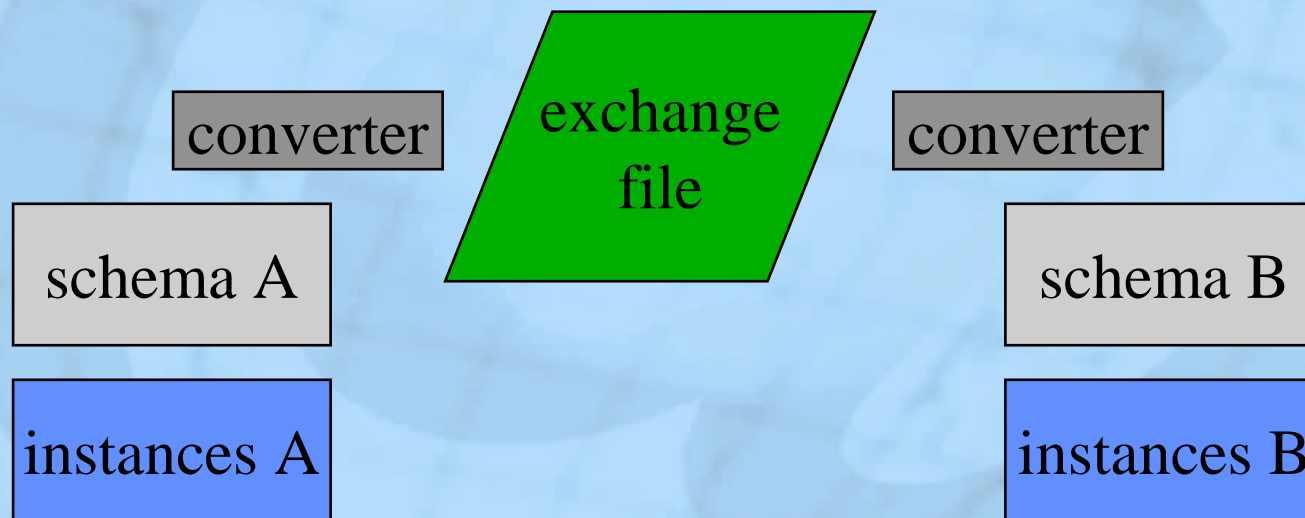
Structure of Session

- CRISs and their Usage
- Database, Data and Metadata
- **Data Exchange**
- Data Access over Heterogeneous Sources
- Data Model for an 'ideal' CRIS
- CERIF in Use

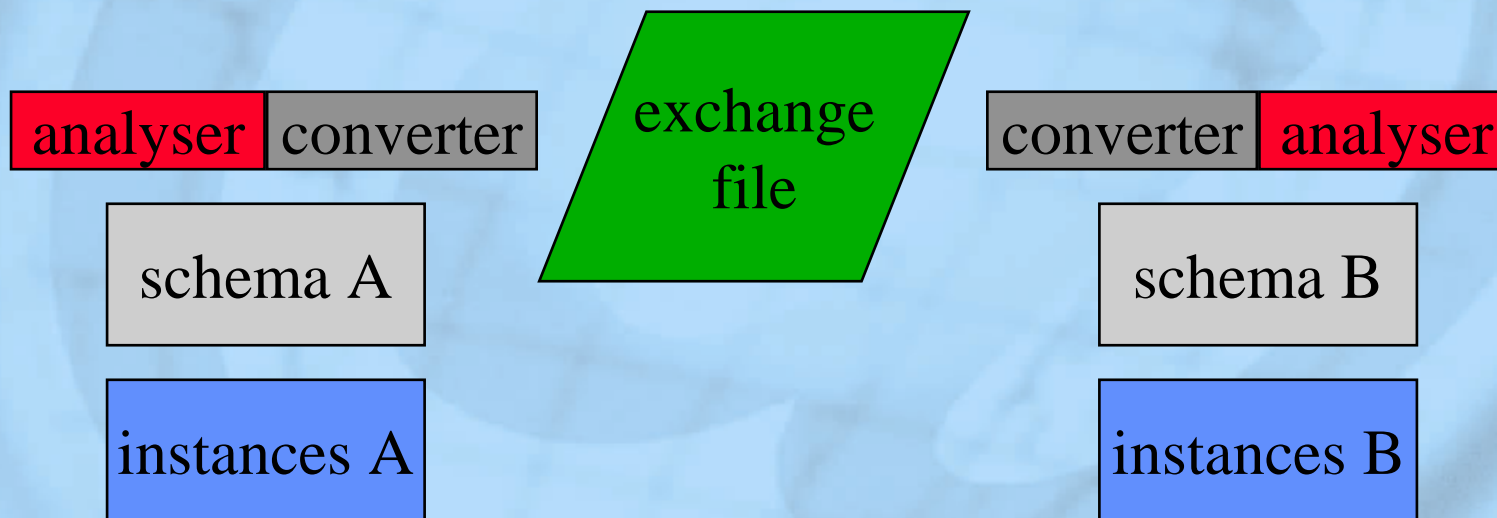


Data Exchange Standard Schema, Standard Structure

- Only content (values of instances) varies
- used for well-defined and agreed exchanges
- e.g.
 - standard reports
 - financial transactions
 - certain industries (e.g. oil, borehole data)
- converter 'hard-wired'



- Structure not predefined
- Negotiated exchange (within limits)
- e.g.
 - person plus skills
 - person plus skills plus job history
- converter has to make decisions within a closed world (standard schema)



- Only the schema language (DDL) is known
- e.g.
 - first attempt at exchange
 - negotiated flexible exchange
- converter has to match exchange schema to native schema

- The Problem -
- only information available is that in schema of exchange file
- and we know how poor schema information can be at logical level
- usually have to add human intelligence
- work on adding intelligence to schema

- Copernicus-funded project 1995-1998
- RAL, MU-ICS, T-Soft, Amis, Elas, MDS
- Use of hyperlinked multimedia as exchange format (represented by graphs)
- Use of logic on arcs
- Use of object-oriented technology at nodes
- Provides maximum flexibility

Structure of Session

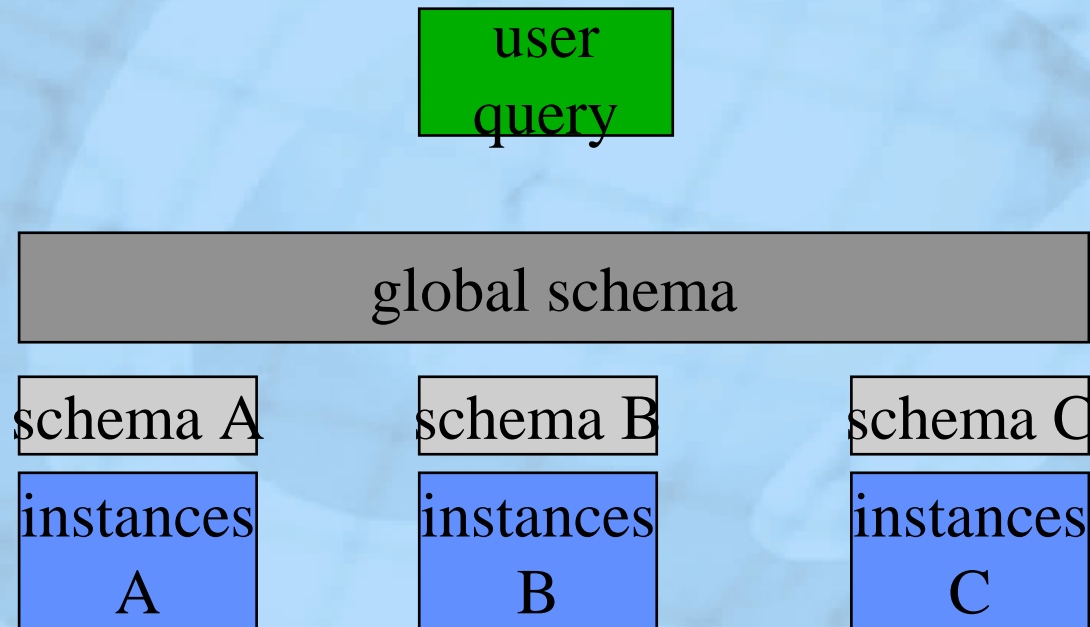
- CRISs and their Usage
- Database, Data and Metadata
- Data Exchange
- Data Access over Heterogeneous Sources
- Data Model for an 'ideal' CRIS
- CERIF in Use

- Global Schema
- Catalog
- Hyperstructures
- Meta-Translation
- Object Equivalencing
- Mediation
- Intelligent Cooperative Systems

- All techniques rely on schema equivalencing - problems:
 - attribute names (syntax, semantics)
 - domains
 - constraints
 - calibration and units / conversion
 - nulls, uncertainty, probability
 - keys - syntax and semantics (structure)

- as well as matching attributes:
- matching structures
 - complex objects
 - roles and sub-entities
 - textbase
 - graphics, images, sound, video.....

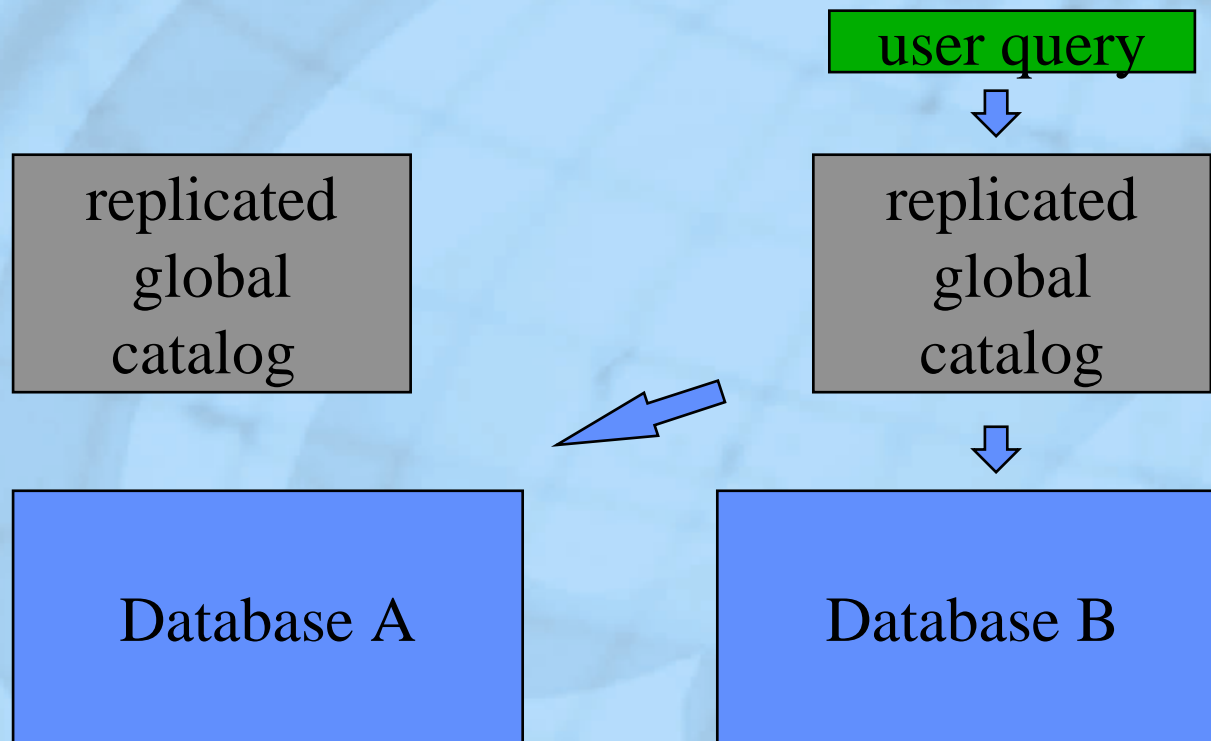
- Schema integration / reconciliation is the major task
- there is a need to reconcile transactions and processes
- there is a need to reconcile events
- there is a need to reconcile constraints (link process and data)



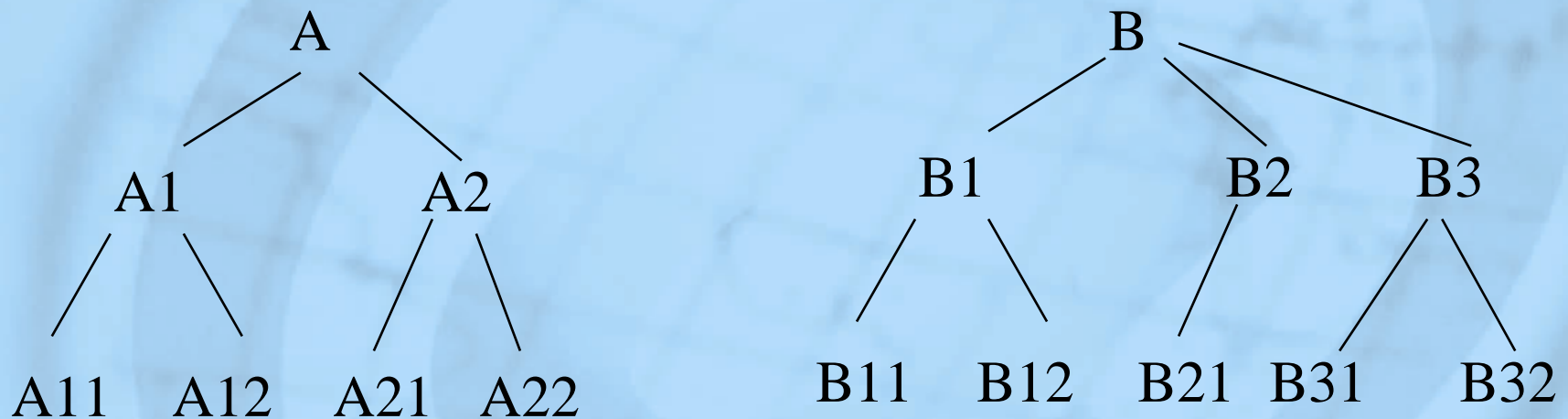
- Create Global Schema
- Common subset of attributes from schemas
- Add non-common attributes
- Ready for queries

- (Felix Saltor)

Data Access Catalog (e.g. EXIRPTS)



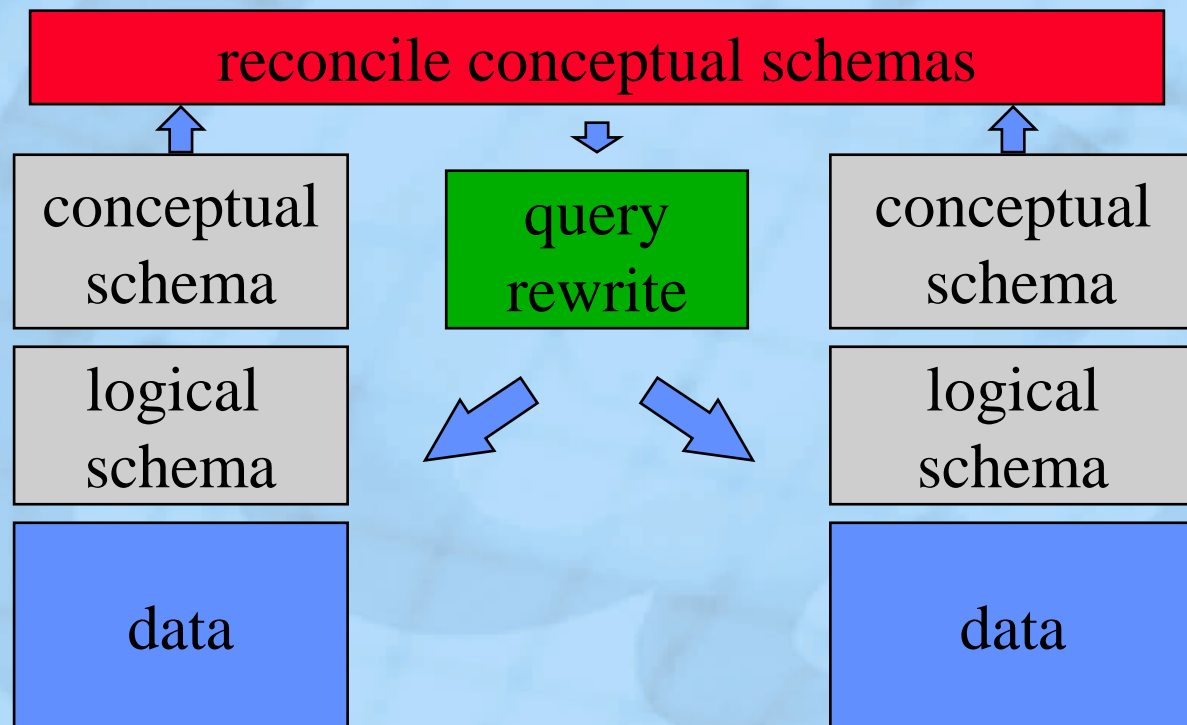
- Desirable (some common) subset of attributes defined
- instances extracted from all databases and converted to common format
- unioned into catalog, replicated to all sites
- queries two-stage:
 - on catalog
 - to obtain all available data indexed by catalog entries
- (Keith Jeffery et al 1988)



Is A2 and substructure = B2 or B3?

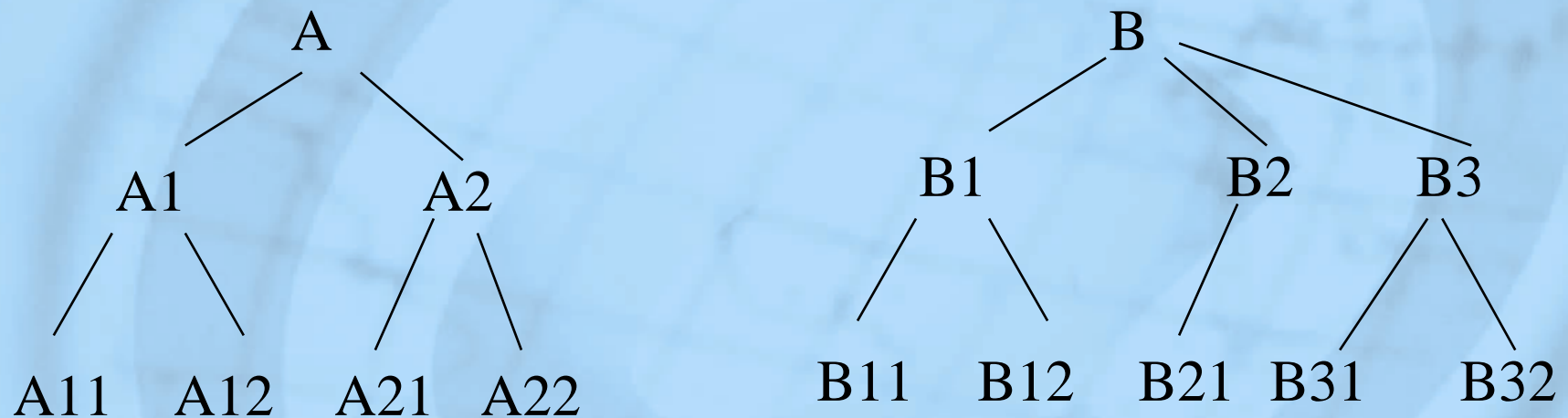
- Complex structure / content can be represented by hyperstructures
 - intensional level
 - extensional level
- Compare hyperstructures and find common structural / content subset
- Link (or exchange) on subset
- Rather like catalog technique

- (Keith Jeffery et al 1994)



- Reverse engineer logical schema to conceptual schema for each database
 - Reconcile schemas at conceptual level
 - Re-write queries to map from conceptual level to each logical level
-
- (Alex Gray, Cardiff)

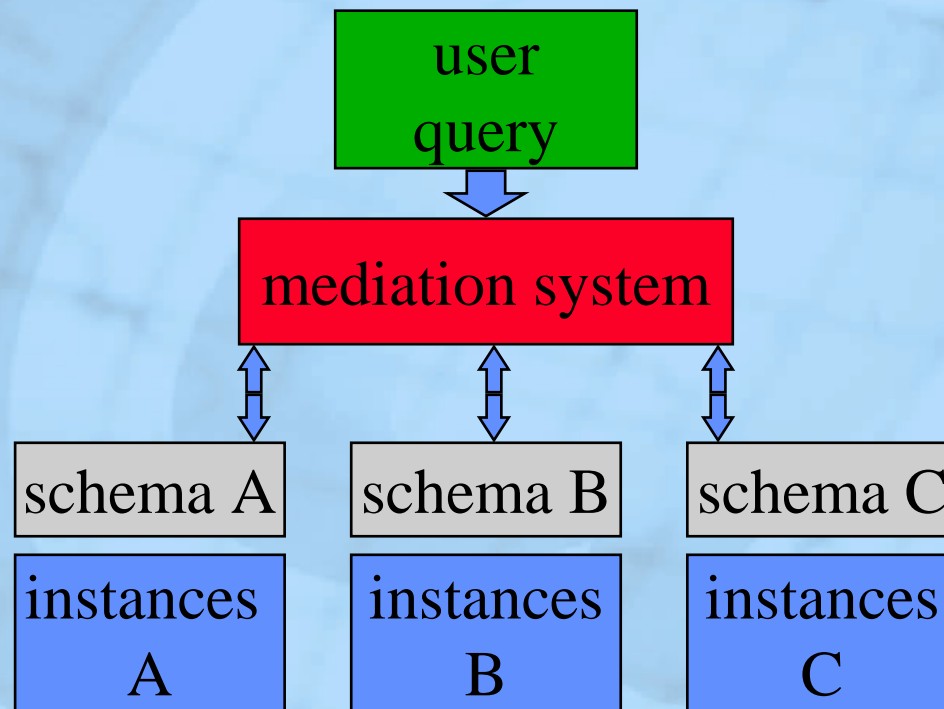
Data Access Object Equivalencing



Is A2 and substructure = B2 or B3?

Object Equivalencing

- Like hyperstructures
- Only works for integrating O-O DBs
- Relies on logical level schema reflecting exactly conceptual level schema
- Unless object attributes match, very difficult
- Encapsulation - so powerful in other contexts, impedes integration
- (many authors, including Patrick Valduriez, INRIA)



- Mediate logical level schemas to conceptual level
- Equivalence at conceptual level
- Requires much domain semantic knowledge attached to mediator
- (Gio Wiederhold, Stanford)

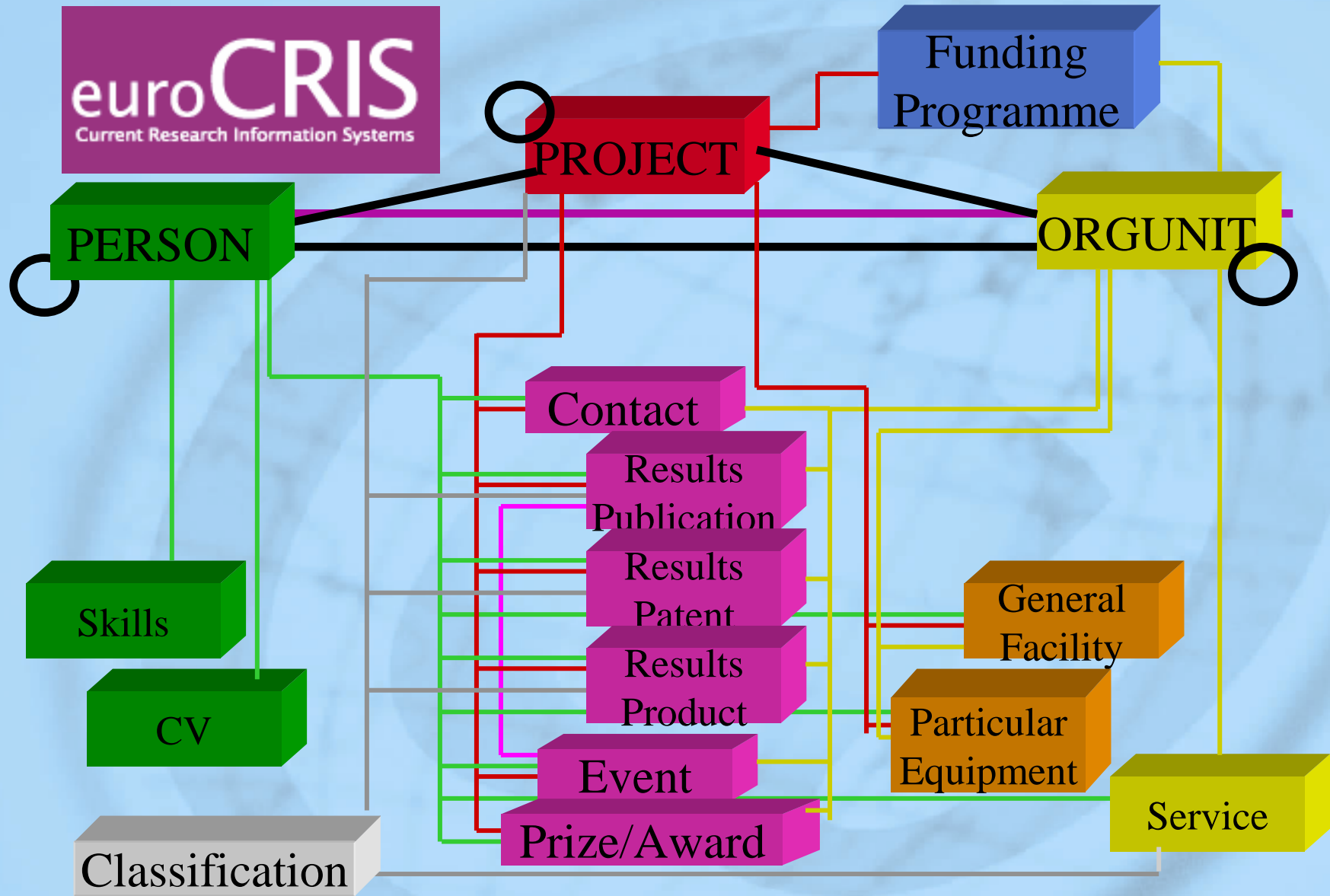
- Each system has intelligence
 - Use intelligence to mediate / negotiate
 - Requires each system to have domain semantic knowledge
 - (extension of mediation technique)
-
- (Mike Papazoglou, QUT)

- CRISs and their Usage
- Database, Data and Metadata
- Data Exchange
- Data Access over Heterogeneous Sources
- Data Model for an 'ideal' CRIS
- CERIF in Use

- An organisation just starting to build a CRIS
- An organisation with one or more legacy CRISs and wishing to evolve to a new single one
- Are in the market for an 'ideal' CRIS
- CERIF provides a template

Ideal Data Model CERIF2000 A Template

- CRIS can be implemented using subset or superset of full CERIF model:
 - for projects (management)
 - for people (expertise)
 - for organisations (capabilities)
 - for publications, patents , products (output)
 - for services (offerings)
 - for facilities, particular equipment (offerings)
- with role-based relationships



Ideal Data Model The Advantages

- Neutral Architecture
- Data Model can be implemented:
 - relational
 - object-oriented
 - information retrieval (including WWW)
- Process model can be implemented
 - DBMS and query; centralised or distributed;
 - html web / harvesting / IR-query;
 - advanced knowledge-based technology

- CRISs and their Usage
- Database, Data and Metadata
- Data Exchange
- Data Access over Heterogeneous Sources
- Data Model for an 'ideal' CRIS
- **CERIF in Use**

The use of CERIF2000 today

- ICERIS (IS) Access to Information on Icelandic Research Projects & R&D Results
- AURIS-MM (AT) Provides access to Austrian University Research extended with multimedia
- SICRIS (SL) Access to University Research in Slovenia
- HUNCRIS Access to R&D in Hungary
- SRIS (GB) Scottish Research Information Systems, public research in Scotland
- CRIS-MER (EC) Research information on Migration and ethnic Relations (planned)
- Corporate model, CRLC (UK)
- METIS (NL) previously OZIS, currently used by majority of Dutch Universities
- Fdok (NO) University of Bergen, results

- CERIF is based on a knowledge of:
 - Previous CRIS systems throughout Europe and wider
 - The R&D in relevant information technologies
- CERIF2000 (and its subsequent developments) is used already in systems
 - And new ones are starting up all the time