RISIS
Research infrastructure for research and innovation policy studies

Annual week 2015

vrije Universiteit amsterdam

Entity-Centric Data Integration & Structured Data

Al Koudous Idrissou

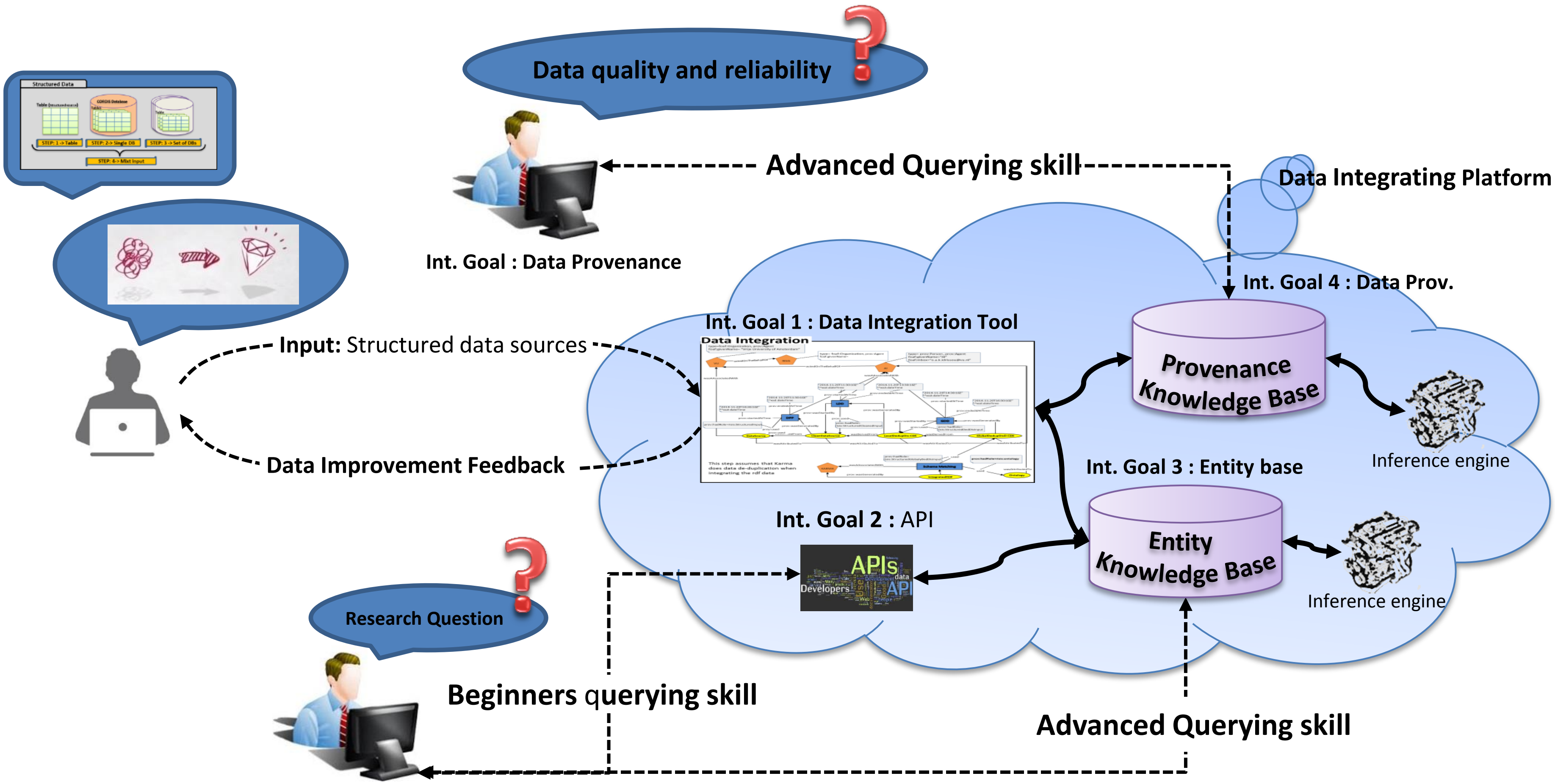
January 26, 2015

Overview

- Goal
- Problem
- Terminologies
- Data integration Tasks
- Tools towards data integration
- Plan

Goal

Increase researchers' capability by creating public single instance **Entity Store** and **Data Provenance Store,** and populate them with an **Automated Data Integrating Tool** using **structured data sources** as input. Among others, the Data Integration Platform ensures a **cleaned, structured, normalized, de-duplicated** and **disambiguated** Entity Knowledge-base and a mean to assess its data's **Quality** and **Reliability** through **data provenance**.



Data quality and reliability

Advanced Querying skill

Int. Goal : Data Provenance

Data Integrating Platform

Int. Goal 4 : Data Prov.

Int. Goal 1 : Data Integration Tool

Input: Structured data sources

Data Improvement Feedback

Provenance Knowledge Base

Inference engine

Int. Goal 3 : Entity base

Int. Goal 2 : API

Entity Knowledge Base

Inference engine

Research Question

Beginners querying skill

Advanced Querying skill

Goal: Increase researchers' capability

Output: Structured, Standardized, De-duplicated , Disambiguated & Integrated Data

Data in the real world

- Ideally, given 2 data sets about the same thing it should be relatively simple to combine them.

- It is not the case in real world.

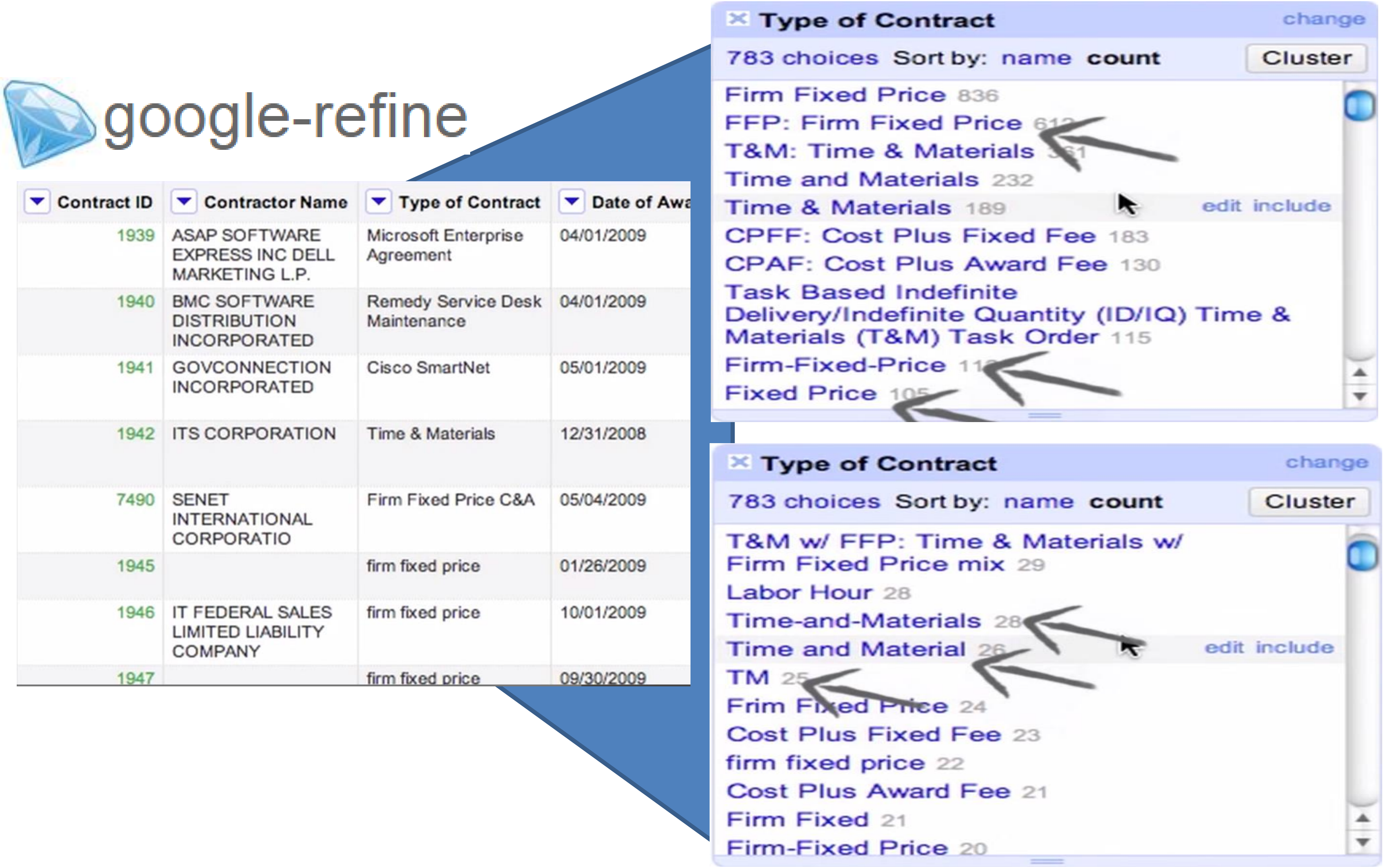
  – Real data is **messy**
  – Real data is **inconsistent**
  – Real data contains **ambiguity**
  – Real data is not **normalized**
  – Real data often has no **uniquely and globally identifiable entities**

- **Difficult to link** the available data between different sources.

- Big limitation in gathering the **right information** in an acceptable time interval.

- Successfully gathered data often presents a lack of meta data about data's **origin**, **context** and **quality**

Structured Messy and Inconsistent Data



Contract ID	Contractor Name	Type of Contract	Date of Awa
1939	ASAP SOFTWARE EXPRESS INC DELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/2009
1940	BMC SOFTWARE DISTRIBUTION INCORPORATED	Remedy Service Desk Maintenance	04/01/2009
1941	GOVCONNECTION INCORPORATED	Cisco SmartNet	05/01/2009
1942	ITS CORPORATION	Time & Materials	12/31/2008
7490	SENET INTERNATIONAL CORPORATIO	Firm Fixed Price C&A	05/04/2009
1945		firm fixed price	01/26/2009
1946	IT FEDERAL SALES LIMITED LIABILITY COMPANY	firm fixed price	10/01/2009
1947		firm fixed price	09/30/2009

Type of Contract change

783 choices Sort by: name count Cluster

Firm Fixed Price 836
FFP: Firm Fixed Price 612
T&M: Time & Materials 361
Time and Materials 232
Time & Materials 189 edit include
CPFF: Cost Plus Fixed Fee 183
CPAF: Cost Plus Award Fee 130
Task Based Indefinite Delivery/Indefinite Quantity (ID/IQ) Time & Materials (T&M) Task Order 115
Firm-Fixed-Price 11[illegible]
Fixed Price 105

Type of Contract change

783 choices Sort by: name count Cluster

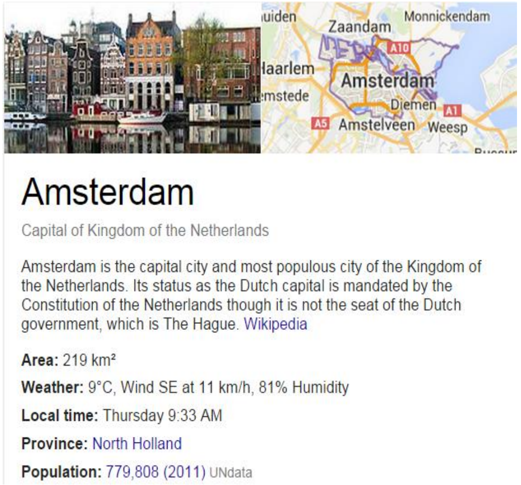
T&M w/ FFP: Time & Materials w/ Firm Fixed Price mix 29
Labor Hour 28
Time-and-Materials 28
Time and Material 26 edit include
TM 25
Frim Fixed Price 24
Cost Plus Fixed Fee 23
firm fixed price 22
Cost Plus Award Fee 21
Firm Fixed 21
Firm-Fixed Price 20

What is an **Entity**?

Person

Organization

Products



Amsterdam

Capital of Kingdom of the Netherlands

Amsterdam is the capital city and most populous city of the Kingdom of the Netherlands. Its status as the Dutch capital is mandated by the Constitution of the Netherlands though it is not the seat of the Dutch government, which is The Hague. Wikipedia

Area: 219 km²

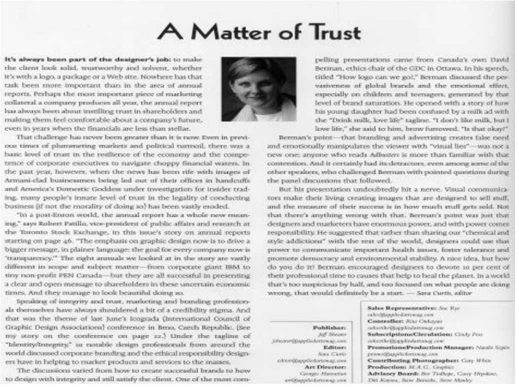
Weather: 9°C, Wind SE at 11 km/h, 81% Humidity

Local time: Thursday 9:33 AM

Province: North Holland

Population: 779,808 (2011) UNdata

City



A Matter of Trust

It's always been part of the designer's job: to make the client look solid, trustworthy and solvent, whether it's with a logo, a package or a Web site. Nowhere has that task been more important than in the area of annual reports. Perhaps the most important piece of marketing collateral a company produces all year, the annual report has always been about instilling trust in shareholders and making them feel comfortable about a company's future, even in years when the financials are less than stellar.

That challenge has never been greater than it is now. Even in previous times of plummeting markets and political turmoil, there was a basic level of trust in the resilience of the economy and the competence of corporate executives to navigate choppy financial waters. In the past year, however, when the news has been rife with images of Armani-clad businessmen being led out of their offices in handcuffs and America's Domestic Goddess under investigation for insider trading, many people's innate level of trust in the legality of conducting business (if not the morality of doing so) has been vastly eroded.

"In a post-Enron world, the annual report has a whole new meaning," says Robert Patillo, vice-president of public affairs and research at the Toronto Stock Exchange, in this issue's story on annual reports starting on page 46. "The emphasis on graphic design now is to drive a bigger message, in plainer language: the goal for every company now is 'transparency.'" The eight annuals we looked at in the story are vastly different in scope and subject matter—from corporate giant IBM to tiny non-profit PEN Canada—but they are all successful in presenting a clear and open message to shareholders in these uncertain economic times. And they manage to look beautiful doing so.

Speaking of integrity and trust, marketing and branding professionals themselves have always shouldered a bit of a credibility stigma. And that was the theme of last June's Icograda (International Council of Graphic Design Associations) conference in Brno, Czech Republic. (See my story on the conference on page 22.) Under the tagline of "Identity/Integrity," 10 notable design professionals from around the world discussed corporate branding and the ethical responsibility designers have in helping to market products and services to the masses.

The discussions varied from how to create successful brands to how to design with integrity and still satisfy the client. One of the most compelling presentations came from Canada's own David Berman, ethics chair of the GDC in Ottawa. In his speech, titled "How logo can we go?," Berman discussed the pervasiveness of global brands and the emotional effect, especially on children and teenagers, generated by that level of brand saturation. He opened with a story of how his young daughter had been confused by a milk ad with the "Drink milk, love life" tagline. "I don't like milk, but I love life," she said to him, brow furrowed. "Is that okay?"

Berman's point—that branding and advertising creates false need and emotionally manipulates the viewer with "visual lies"—was not a new one; anyone who reads *Adbusters* is more than familiar with that contention. And it certainly had its detractors, even among some of the other speakers, who challenged Berman with pointed questions during the panel discussions that followed.

But his presentation undoubtedly hit a nerve. Visual communicators make their living creating images that are designed to sell stuff, and the measure of their success is in how much stuff gets sold. Not that there's anything wrong with that. Berman's point was just that designers and marketers have enormous power, and with power comes responsibility. He suggested that rather than sharing our "chemical and style addictions" with the rest of the world, designers could use that power to communicate important health issues, foster tolerance and promote democracy and environmental stability. A nice idea, but how do you do it? Berman encouraged designers to devote 10 per cent of their professional time to causes that help to heal the planet. In a world that's too suspicious by half, and too focused on what people are doing wrong, that would definitely be a start. — Sara Curtis, editor

Publisher:
Jeff Shearer
jshearer@appliedartsmag.com
Editor:
Sara Curtis
editor@appliedartsmag.com
Art Director:
Georges Haroutiun
art@appliedartsmag.com

Sales Representative: *Sue Rye*
sales@appliedartsmag.com
Controller: *Rita Ozkaynak*
rokaynak@appliedartsmag.com
Subscriptions/Circulation: *Cindy Fisi*
[illegible]@appliedartsmag.com
Promotions/Production Manager: *Natalie Ségin*
promo@appliedartsmag.com
Contributing Photographer: *Gary White*
Production: *M.A.G. Graphics*
Advisory Board: *Ben Tollefson, Casey Hrynkow, Dirk Kanou, Steve Brisitz, Steve Mosley*

Article

Real World Things

OR

Anything

W3C® Provenance Ontology defines entity as a physical, digital, conceptual, or other kind of **thing with some fixed aspects; entities may be real or imaginary**

.

What is **Data Provenance**?

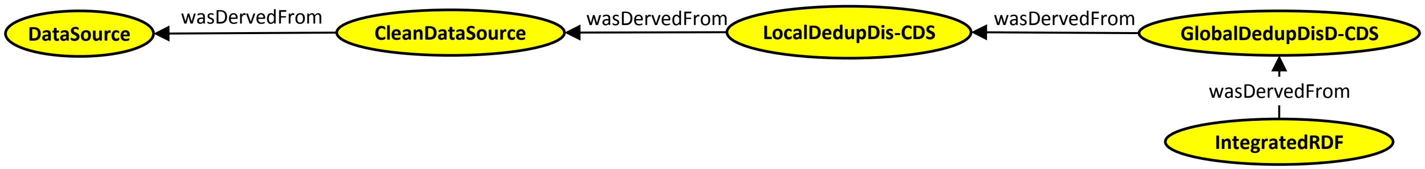
Def: “**Provenance** is to **electronic data** what a record of ownership is to a work of art.” [Moreau et Al, 2007]

W3C **Goal**: Provenance captures information about entities, activities and people involved in producing a piece of data or thing which can be used to form assessments about it quality, Reliability or trustworthiness

Purpose

- **Understand** how data was collected
- Determine **ownership** and **wrights** over an object
- Making judgment about information to determine whether to **trust it**.
- **Verify** the **process** and steps used to obtain a result given a set of **requirements**
- **Reproduce** how something was generated

Prov Data flow



DataSource

wasDervedFrom

CleanDataSource

wasDervedFrom

LocalDedupDis-CDS

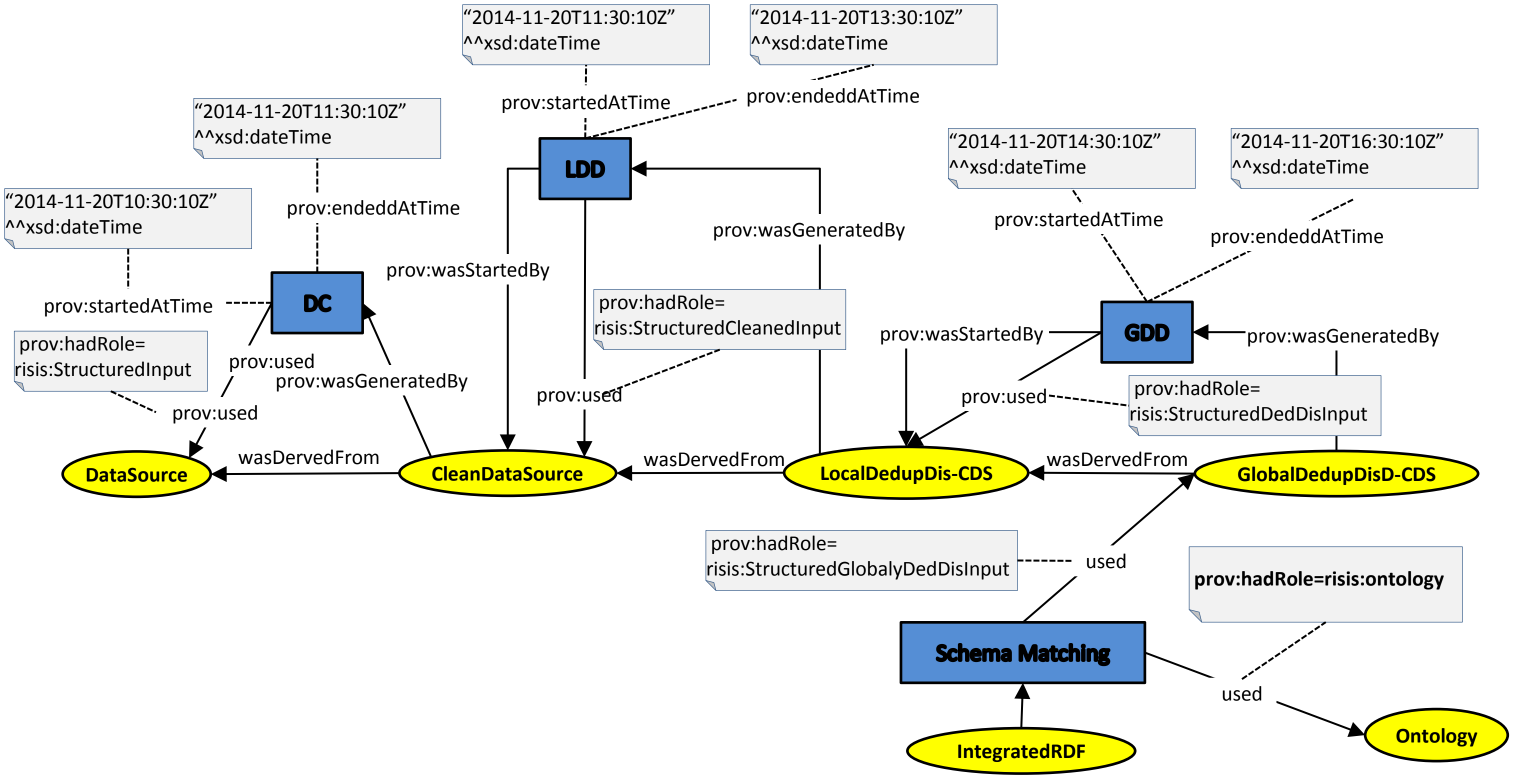
wasDervedFrom

GlobalDedupDisD-CDS

wasDervedFrom

IntegratedRDF

Prov Process flow



"2014-11-20T11:30:10Z" ^^xsd:dateTime

"2014-11-20T13:30:10Z" ^^xsd:dateTime

prov:startedAtTime

prov:endeddAtTime

"2014-11-20T11:30:10Z" ^^xsd:dateTime

LDD

"2014-11-20T14:30:10Z" ^^xsd:dateTime

"2014-11-20T16:30:10Z" ^^xsd:dateTime

"2014-11-20T10:30:10Z" ^^xsd:dateTime

prov:endeddAtTime

prov:wasGeneratedBy

prov:startedAtTime

prov:endeddAtTime

prov:wasStartedBy

prov:startedAtTime

DC

prov:hadRole= risis:StructuredCleanedInput

prov:hadRole= risis:StructuredInput

prov:wasStartedBy

GDD

prov:wasGeneratedBy

prov:used

prov:wasGeneratedBy

prov:used

prov:used

prov:hadRole= risis:StructuredDedDisInput

prov:used

DataSource

wasDervedFrom

CleanDataSource

wasDervedFrom

LocalDedupDis-CDS

wasDervedFrom

GlobalDedupDisD-CDS

prov:hadRole= risis:StructuredGlobalyDedDisInput

used

prov:hadRole=risis:ontology

Schema Matching

used

Ontology

IntegratedRDF

What is a **Knowledge-base System ?**

- **Data** raw signals” **[**ex: . . . - - - . . .]
- **Information** attaches **meaning** to data **[**Ex: S O S**]**
- **Knowledge** attaches **purpose** and **competence** to information **[**Ex: emergency alert -> start rescue operation **] ->** potential to generate action

Fact: Alice is an animated character

A **knowledge base** (**KB**) is a technology used to sore complex information (**facts about the world**) used by a computer system.

➔ While a **knowledge-base** stores facts about the world, an **Inference engine** reasons about facts to infer new facts.

A **knowledge-base system** consists a knowledge-base and an inference engine

Why Data Integration?

Need for seamlessly getting a complete and accurate view on the same thing.

Thing: Professors in the Netherland?
Resources: VU – UVA – INHOLLAND ….

Problem

Based on the same requirements, 2 databases designed by 2 different persons will look significantly different (Schema & representation)

Employee Database [VU]

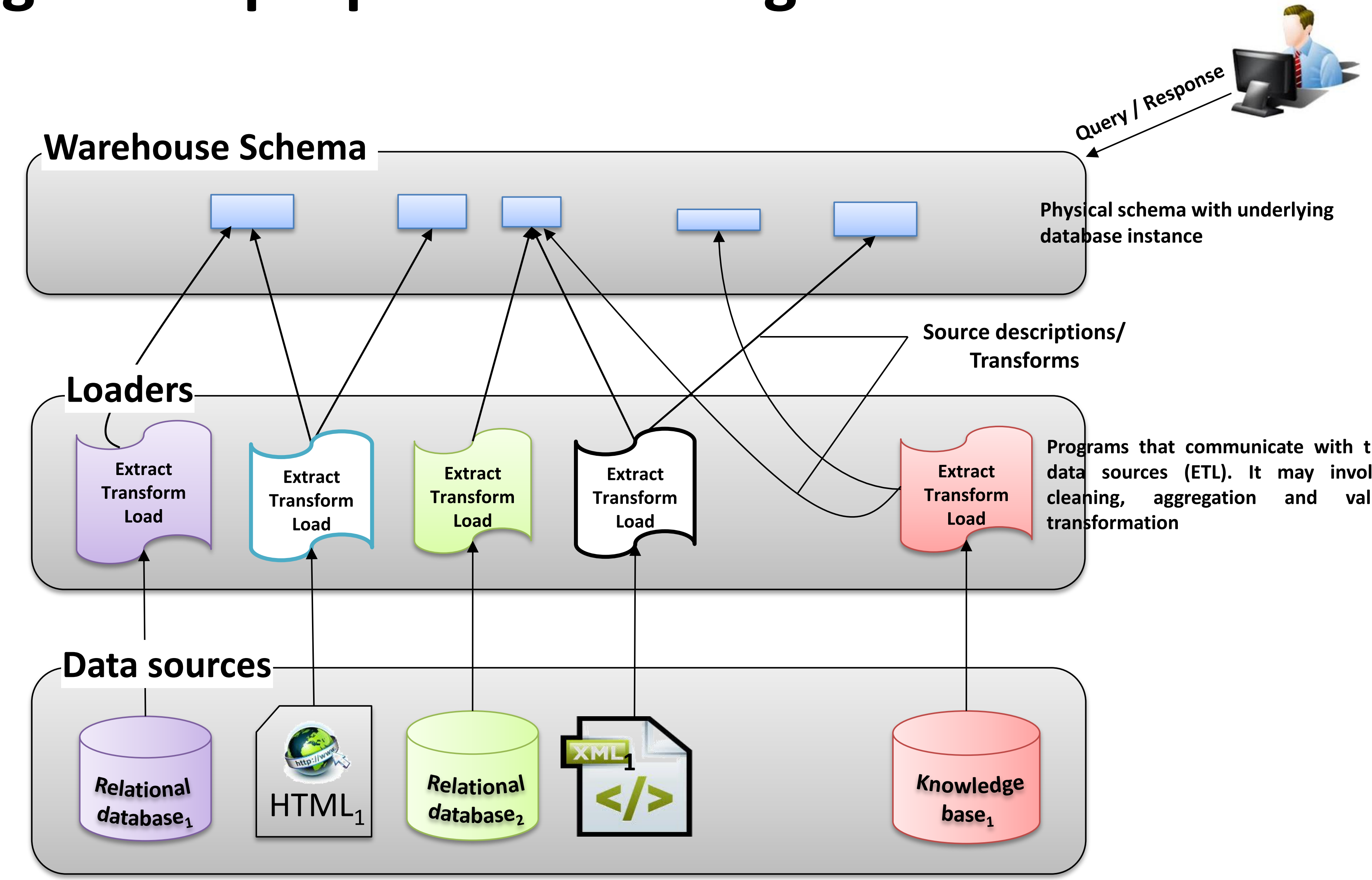
FullTimeEmp(ssn, empID, firstName, middleName, lastName)
Hire(empID, hireDate, recruiter)
TemEmplofees(ssn, hireStart, hireEnd, name, hourlyRate)

Semantic heterogeneity

Employee Database [UVA]

Emp(ID, firstNameMiddleInitial, lastName, salary)
Hire(ID, hireDate, recruiter)

Basic Architecture of a Physical general-purpose data integration

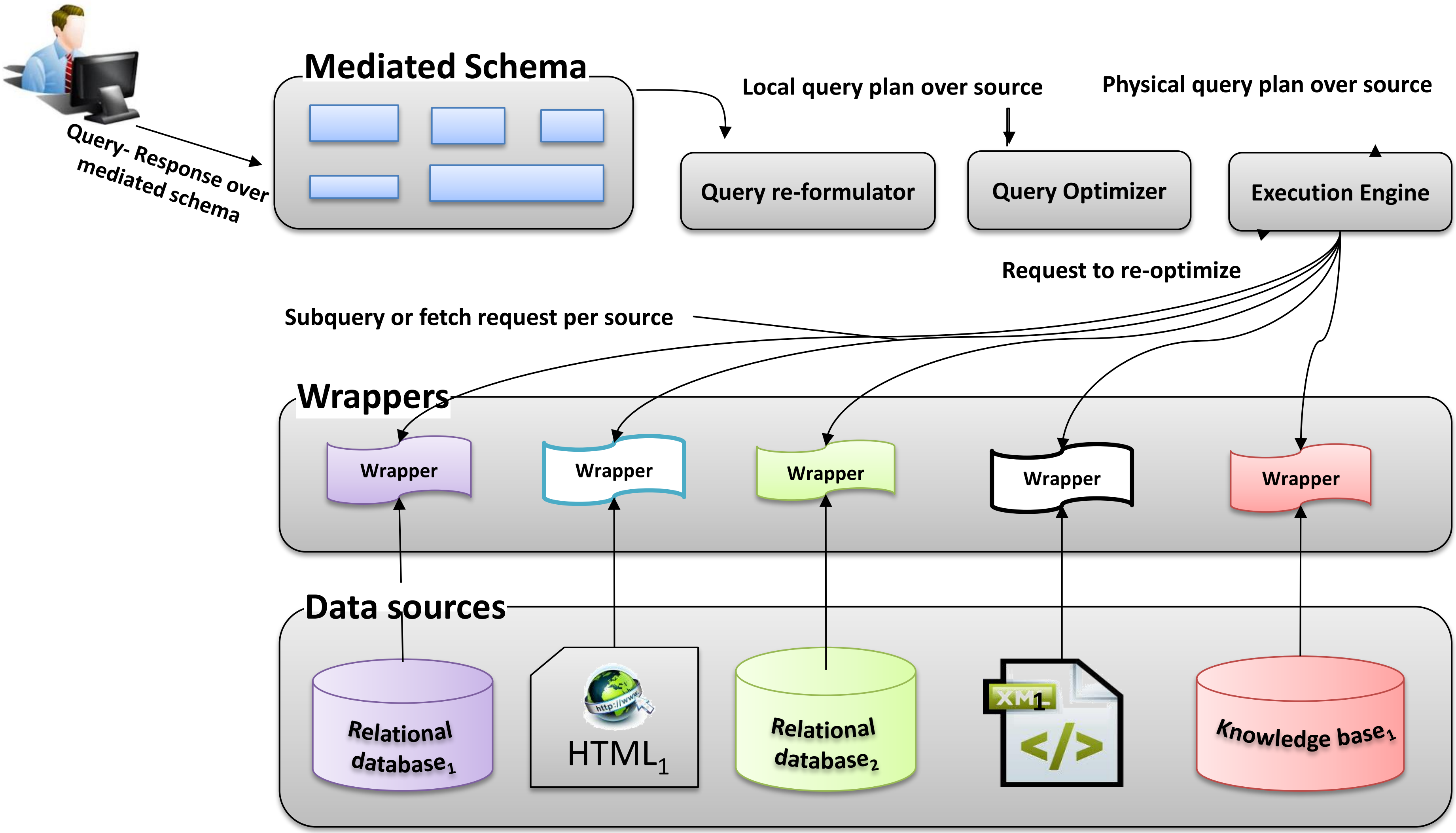


Physical schema with underlying database instance

Source descriptions/ Transforms

Programs that communicate with the data sources (ETL). It may involve cleaning, aggregation and value transformation

Basic Architecture of a Virtual general-purpose data integration



Tasks toward data Integration?

- **Input**: Structured data about Entity
- Data pre-processing
- Schema matching and Mapping
- Data Deduplication
- Data Disambiguation
- Data reconciliation
- Data Consistency
- Data Integrity
- **Output**: RDF

Data Pre-processing

Removing Unwanted Characters and Tokens

\FR?D?RIC JOLIOT-CURIE\" NATIONAL RESEARCH INSTITUTE FOR RADIOBIOLOGY AND RADIOHYGIENE"

→ \fr?d?ric joliot-curie\" national research institute for radiobiology and radiohygiene"

2.-0 LCA CONSULTANTS APS → 2.-0 lca consultants aps

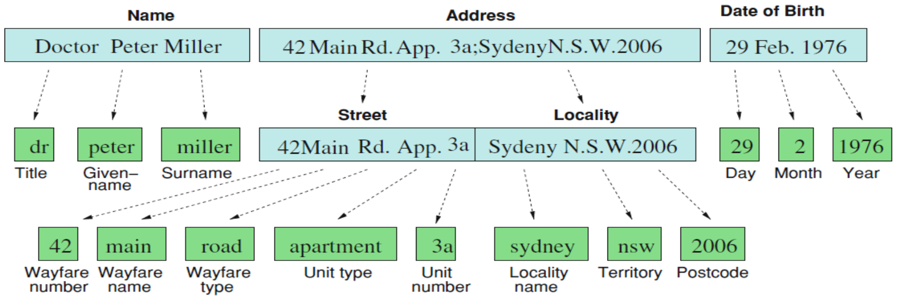
3d Scanners Ltd → 3d scanners | 3d scanners limited | 3d Ssanners limited

anti-discriminatory → antidiscriminatory | anti discriminatory

U.S.A. → usa | united states of america

résumé vs. resume | tuebingen vs. tübingen | pena vs. peña.

Standardisation , Tokenisation & Segmentation into Output Fields



2/3/91 → 2/3/91 | 2-3-91 | February 3rd, 1991

OR

3/2/91 → 3/2/91 | 3-2-91 | Mars 2sd, 1991

Verification

Schema Matching and Mapping

DVD-VENDOR

Movies(id, title, year)
Products(mid, releaseDate, releaseCompany, basePrice, rating, saleLocID)
Locatios(lid, name, classification, price)

AGGREGATOR]

Items(name, releaseInfo, classification, price

Schema Matching identifies **correspondenc**e between elements of two schemas

Movies.title ≈ **Items.**name
Movies.year ≈ **Items.**year
Products.rating ≈ **Items.**classification
Items.price ≈ **Products.**basePrice * (1 + **Location.**taxRate)

Schema Mapping describes how to convert a **source schema** into a **target schema**

Record Deduplication - Data Inconsistency
Data Integrity

Data Deduplication is a technique for storing only one copy of repeating data (identical data)

ID	Title	First	Last	AddressLine	City	Postcode	Telephne	DOB
1	Miss	Catrina	Trewin	123 Sample Road	Town	ABC 123	123456789	06/15/75
2	Miss	Catrina	Trewin	123 Sample Road	Town	ABC 123	**123456780**	06/15/75
3		Catrina	Trewin	123 Sample Road	Town	ABC 123	123456789	**06/15/76**
4	Miss	**C**	Trewin	123 Sample Road		ABC 123		06/15/75

Advantage

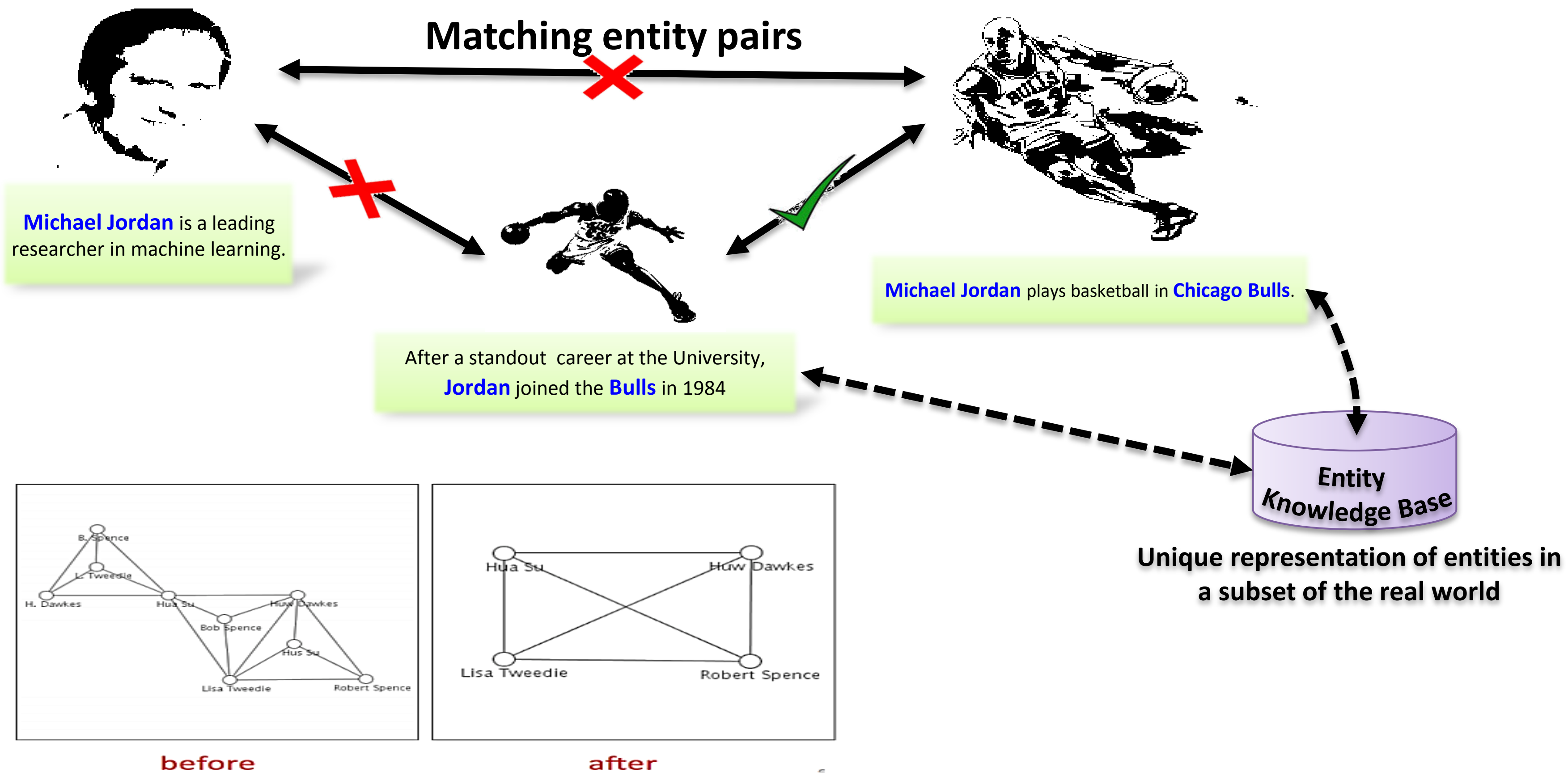
- Improve **storage** utilization
- Reduces the amount of bytes (data) that must be **transferred**
- Provides accuracy for **statistical analyses**

Drawbacks

- **Data Integrity** (potential loss of data)
- Computational **resource intensive**

Entity Disambiguation

Entity Disambiguation is the process of removing **uncertainty** or **confusion** introduced by a named entity suggesting multiple interpretations. In other words, Entity Disambiguation is the practice of **identifying and linking a confusing named entity** (textual form) **to its true and unique representation** (real world) **within a knowledge base** which, in turn, provides a single semantic interpretation.



Matching entity pairs

Michael Jordan is a leading researcher in machine learning.

After a standout career at the University, **Jordan** joined the **Bulls** in 1984

Michael Jordan plays basketball in **Chicago Bulls**.

Entity Knowledge Base

Unique representation of entities in a subset of the real world

before

after

Data Integration Tasks vs. Tools Overview

	Pre-processing	Consistency Check	S. Matching	D. Integrity	Deduplication	Disambiguation	Reconciliation	Prov.
Duke					🟧		🟧	
Karma			⬜					⬜
Open Refine	🟩	🟩		🟩	🟩	🟩	🟩	
D-Dupe					🟦			
FOXPSL					🟫	🟫	🟫	

Vision Automated Data Integration Platform for Researchers

Plan

- Create an Entity Knowledge-base
- Guidelines for Technical Harmonization
- Tools Implementation and Evaluation
- What to reuse? what to improve?
- RISIS Integration Tool
- Populate the Knowledge-base

Reference

- Anhai Doan, Alon Halevy and Zachry Ives. Principles of data integration. Morgan Kaufmann Publishers. 2012 Elsevier. ISBN: 978-12-416044-6
- Peter Christen. Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer-Verlag Berlin Heidelberg 2012. ISBN 978-3-642-31164-2
- Philip A. Bernstein, Jayant Madhavan and Erhard Rahm. Generic Schema Matching, Ten Years Later. PVLDB 4(11): 695-701 (2011)