Monitoring Entities in an Uncertain World: Entity Resolution and Referential Integrity

Steven Minton, InferLink Corporation
Sofus Macskassy, Fetch Technologies
Peter LaMonica, Air Force Research Laboratories
Kane See, InferLink Corporation
Craig Knoblock, USC/Information Sciences Inst.
Greg Barish, Fetch Technologies
Matthew Michelson, Fetch Technologies
Ray Liuzzi, Raymond Technologies
Many “Steve Minton”s

- Steve Minton, InferLink
- Steven Minton, Stanford University
- Stephen Minton, Brain Surgeon
- Steve Minton, Fetch Technologies
- Steven Minton, convicted felon
- Steven Minton, Jonosboro High School
- Steven Minton, JAIR
Outline

- Application domain: Arms trafficking
- Entity Intelligence Portal (ENTEL)
- Entity resolution process
- Mistakes: Maintaining referential integrity
Web Monitoring

AIJ
AAAI
Grants.gov
IEEE Intelligent Systems
JAIR
ICML

Web Monitoring System
Web Monitoring

NASA
US Forest Service
Twitter
National Interagency Fire Center
InciWeb.org

Web Monitoring System
Web Monitoring

Airliners.net
Banned Airlines
ATWonline
Air Cargo News
Aviation Safety Network
Twitter
Aviation Week

Web Monitoring System
Victor Bout
Arrested and Extradited in 2010

Charged with conspiracy to support a terrorist organization, money laundering, ....,
Some Companies Linked with Bout

- Air France
- Cessna
- Centrafricain Airlines
- Great Lake Business Company
- Air Leone
- Inter Transavia
- MaxAir
- Air Mero
- Balkh Airlines
- Air Bas
- Pietersberg Aviation Services Systems
- Ishtar Airlines
- Air Foyle
- Air Cess
- Galaxy Air
- Omega Aircompany
- Irbis Air
- Mega Airlines
- WING AIR
- Aerocom
- Norse Air Charter
- Click Airways
- Anikay (Anikai) Airlines
- Santa Cruz Imperial
- JetLine International
- San Air General Trading
- African Express
- Flying Dolphin
# One Thread: Air Cess to Air Bas

<table>
<thead>
<tr>
<th>Registration</th>
<th>Aircraft Type</th>
<th>Construction Nbr</th>
<th>Previous Reg.</th>
<th>Sighting</th>
<th>Markings</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN-75002</td>
<td>Ilyushin 18E</td>
<td>185008603</td>
<td>3C-KKR</td>
<td>SHJ 11May03 SHJ 04Nov03</td>
<td>no markings a/w, n/t</td>
</tr>
<tr>
<td>UN-75003</td>
<td>Ilyushin 18V</td>
<td>184006903</td>
<td>3C-KKJ</td>
<td>SHJ 12Oct03</td>
<td>blue tail, no m/s</td>
</tr>
<tr>
<td>UN-75004</td>
<td>Ilyushin 18D</td>
<td>186009202</td>
<td>3C-KKK</td>
<td>SHJ 14Sep02 SHJ 04Nov03 SHJ 28Dec03</td>
<td>green cheatline and blue tail No t/t, blue tail all white</td>
</tr>
<tr>
<td>UN-75005</td>
<td>Ilyushin 18D</td>
<td>187010204</td>
<td>3C-KKL</td>
<td>SHJ 04Nov03</td>
<td></td>
</tr>
<tr>
<td>UN-11007</td>
<td>Antonov 12B</td>
<td>9346509</td>
<td>3C-OOZ</td>
<td>SHJ Oct02 SHJ 11May03 DXB 12Oct03</td>
<td>No m/s all white c/s no titles</td>
</tr>
</tbody>
</table>

[From Ruudleeuw.com]
An Online Data Source
This AN-12B 3C-OOZ (c/n 9346509) still wears the colourful scheme of Ural Airlines, but it seems likely that it is now operated by someone else. I took this photo April 26th, 2000 at Sharjah and as it was still around in 2001, one can assume that Sharjah has become its homebase. They probably have a reason for not advertising the name of the company on the aircraft, pity....
One More Source

AviationSafetyNetwork

an exclusive service of Flight Safety Foundation

www.flightsafty.org

Accident description

Status: 31 MAR 2005
Type: Antonov 12B
Operator: PFS Air Freight
Registration: UN-11007
C/n / msn: 9346509
First flight: 1969
Engines: 4 Ivchenko AI-20M
Crew: Fatalities: 0 / Occupants: 8
Passengers: Fatalities: 0 / Occupants: 0
Total: Fatalities: 0 / Occupants: 8
Airplane damage: Written off
Airplane fate: Written off (damaged beyond repair)
Location: Mukalla-Al-Rayyan International Airport (MKX) (Yemen)
Phase: Takeoff (TOF)
Nature: Cargo
Departure airport: Mukalla Airport (MKX) (MKX), Yemen
Destination airport: Dubai Airport (DXB) (DXB/OMDB), United Arab Emirates
Narrative:
The Antonov had made a refueling stop at Mukalla and was bound for Dubai, carrying 7 tons of fish. The takeoff from Mukalla had to be aborted, but the plane could not be stopped on the runway. It overran by 400 meters and caught fire. The fire was quickly contained. Possible hull-loss.
Fetch Agent Platform™ (web harvesting)

Source A
Source B
Source C
Source D

Fact Extraction (entities, facts, relations from unstructured text)

OpenCalais
Semantex

Entitybase™ (entity resolution)

Facts
Entity IDs

Web GUI

Analytics Engine and Entity-Resolved Content Store

Text
Social Network
WatchLists
<table>
<thead>
<tr>
<th>Entity Name</th>
<th>Content Description</th>
<th>Content Type</th>
<th>Content Source</th>
<th>Date Publish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volga-Dnepr Airlines</td>
<td>Will Cargo Profiling Take Hold?</td>
<td>article</td>
<td>AirCargoNewsCom</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Volga-Dnepr Airlines</td>
<td>Big Lift To Subcontinent</td>
<td>article</td>
<td>AirCargoNewsCom</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Wings Air</td>
<td>Lion Air, Virgin Blue Add 33 ATR-72</td>
<td>article</td>
<td>DefenseAerospaceNews</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>SonAir</td>
<td>Heli-Expo: Sikorsky Sells 12 Helicop...</td>
<td>article</td>
<td>DefenseAerospaceNews</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>SonAir</td>
<td>Cargo Handlers Get No Respect</td>
<td>article</td>
<td>AirCargoNewsDot</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>SonAir</td>
<td>Prosperous Bosphorus IATA Cargo Week</td>
<td>article</td>
<td>AirCargoNewsDot</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Volga-Dnepr Airlines</td>
<td>Russian investigators have determined that the Tupolev...</td>
<td>article</td>
<td>FlightGlobal</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Philippine Airlines</td>
<td>Philippine carriers and Iran Air feature on latest EU blac...</td>
<td>article</td>
<td>FlightGlobal</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Air Koryo</td>
<td>Philippine carriers and Iran Air feature on latest EU blac...</td>
<td>article</td>
<td>FlightGlobal</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>CitiLink</td>
<td>Air Canada, Garuda, Thai and Air Austral could sublease...</td>
<td>article</td>
<td>FlightGlobal</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Volga-Dnepr Airlines</td>
<td>Hearts &amp; Flowers In SWISS SFO</td>
<td>article</td>
<td>AirCargoNewsDot</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Kartika Airlines</td>
<td>Farnborough International Airshow News</td>
<td>article</td>
<td>atiworldline</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>Kartika Airlines</td>
<td>Kartika Airlines Orders 30 Sukhoi Su...</td>
<td>article</td>
<td>DefenseAerospaceNews</td>
<td>2011-07-27</td>
</tr>
<tr>
<td>CitiLink</td>
<td>Airbus Wins 58 New Orders, Options a...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippine Airlines</td>
<td>AirAsia to launch Philippines-based affiliate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mandala Airlines</td>
<td>Indonesia's Mandala Airlines suspends operations, seeks new investors</td>
<td>article</td>
<td>atiworldline</td>
<td>2011-07-27</td>
</tr>
</tbody>
</table>
DB Record for Plane [Cubana] spotted at Airport [Recife - Guararapes]

Original link: www.airliners.net
airlinersphoto
Published: 2010-12-08T23:18:02-08:00
Source: airlinersphoto

Airline: Cubana
Airport: Recife - Guararapes
Aircraft: Ilyushin Il-62M
Location: Brazil

CU-T1284 (cn 4053732) CU-T1284. Very rare plane! Taking off to Cuba in a cloudy morning. Beautiful plane... I just love it! Not the best of quality, but a beautiful and special photo to me. Check out the condensation over the wings. [Nikon D5000 - 70/300 f/4:5.6 - ISO 1600]
# PT Merpati Nusantara Airlines

**First seen:** 2011-07-08T12:35:33-07:00  
**Last seen:** 2009-07-06T07:00:00-07:00  
**Times seen:** 5

<table>
<thead>
<tr>
<th>Headline</th>
<th>Source</th>
<th>Type</th>
<th>Watchlist</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB record for Merpati Nusantara Airlines</td>
<td>aircarriersbanned</td>
<td>structured_input</td>
<td>0</td>
<td>2011-07-08</td>
</tr>
<tr>
<td>DB record for Merpati Nusantara Airlines</td>
<td>airlinesfeed</td>
<td>structured_input</td>
<td>0</td>
<td>2011-07-08</td>
</tr>
<tr>
<td><strong>DB Record for Plane [Merpati Nusantara Airlines]</strong></td>
<td>airlinersphoto</td>
<td>plane_spotting</td>
<td>1</td>
<td>2011-07-08</td>
</tr>
<tr>
<td><strong>Merpati abandons purchase deal with Xi'an aircraft</strong></td>
<td>CargoNewsAsia</td>
<td>article</td>
<td>2</td>
<td>2009-07-06</td>
</tr>
<tr>
<td>Name</td>
<td>Predicate</td>
<td>Value</td>
<td>Last Seen</td>
<td>Frequency</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>----------------</td>
<td>---------------------------------</td>
<td>------------</td>
<td>-----------</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Name</td>
<td>Merpati Nusantara Airlines</td>
<td>2011-07-08</td>
<td>2</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>UsesAircraft</td>
<td>Xian MA-60</td>
<td>2011-07-08</td>
<td>1</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Country</td>
<td>China</td>
<td>2011-07-08</td>
<td>1</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Callsign</td>
<td>MERPATI</td>
<td>2011-07-08</td>
<td>2</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>IATA</td>
<td>MZ</td>
<td>2011-07-08</td>
<td>2</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>ICAO</td>
<td>MNA</td>
<td>2011-07-08</td>
<td>2</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Country</td>
<td>Indonesia</td>
<td>2011-07-08</td>
<td>2</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Carrier</td>
<td>Merpati Nusantara Airlines</td>
<td>2011-07-08</td>
<td>1</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Name</td>
<td>PT Merpati Nusantara Airlines</td>
<td>2009-07-06</td>
<td>2</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Nationality</td>
<td>Indonesian</td>
<td>2009-07-06</td>
<td>1</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Name</td>
<td>Indonesian state-owned airline...</td>
<td>2009-07-06</td>
<td>1</td>
</tr>
<tr>
<td>PT Merpati Nusantara Airlines</td>
<td>Name</td>
<td>Merpati</td>
<td>2009-07-06</td>
<td>2</td>
</tr>
</tbody>
</table>
The Entity Resolution Problem

- Entity resolution: Link incoming records describing the same entity from multiple sources
  
  \[ \text{R. Landis, President, Fetch Technologies} \]
  
  \[ \text{Robert Landes, CEO, Fetch Software} \]
  
  \[ \text{R. Land, CEO, French Alliance Technologies} \]

- Many “common sense” issues, for instance:
  - Multiple formats for names, addresses, etc.
    - R.L. Landes vs. Robert Landes
  - Noisy, incorrect values
    - Landes vs. Landis
  - Multi-valued attributes
    - Landes can be both President and CEO
  - Aliases and Deception
Outspoken Qatar Airways boss Akbar Al Baker has responded angrily to a recent address by the Association of European Airlines' secretary general Ulrich Schulte-Strathaus, in which he called on the行业的 growing threat from the Gulf carriers.

"Outspoken Qatar Airways"
Cluster is a single entity

Composed of multiple data records

Confidence Threshold

New Record

E1
E2
E3
E4
E5
E6
E7
Estimating Probabilities

Robert Landes, CEO, Fetch Tech

R. Landis, President, Fetch Tech

Transformations
Initial: Robert → R.
Spelling: Landes → Landis
Title alias: CEO → President

Transformations
Spelling: Land → Landis
Spelling: French → Fetch
Robert Landes,
CEO, Fetch Tech

P(E₁ | D) = \frac{P(E₁) P(D | E₂)}{P(D)}

R. Landis,
President, Fetch Tech

P(E₂ | D) = \frac{P(E₂) P(D | E₂)}{P(D)}

R. Land,
President, French Tech
Estimating Probabilities

\[
P(E_{\text{new}}) \frac{P(D|E_{\text{new}})}{P(D)}
\]
Clustering
Clustering

New Record
Clustering
Correcting mistakes: Merging and splitting clusters

- **Merge example:**
  - Air Cess and Air Bas aircraft

- **Split example:**
  - George H. W. Bush and George W. Bush
Merging Two Clusters

EntityBase
Referential Integrity

EntityBase

E3
E5
E6
E4
E10

E2?
Refer-by-Description
Refer-by-Description
Refer-by-Description

EntityBase

Analytics

WatchList

Kartiga Air (D9)
Merpati Airlines (D11)
Air Cess (D138)
....
Refer-by-Identifier

EntityBase

Publish Merges/Splits

“Social” Network

E1
E2
E3
E4
E5
E6
E10
E15
E34
E9
E91
E200
Two approaches:

- **Refer-by-Description**
  - Indirect reference: Point to a cluster member
  - Advantage: Easy, no synchronization necessary
  - ...But limits information that client can cache

- **Refer-by-Identifier**
  - Direct reference: Cluster ID
  - Advantage: Client can cache arbitrary information
  - ...But client must synch with EntityBase and maintain consistency
Why It Matters

- **Vision:** Entity Resolution in a decentralized world
  - E.g., the Semantic Web (Glaser, Jaffri & Millard, 2009)
Summary

- Entity resolution can be hard: “AI Complete”
  - Arms trafficking domain
- Entity merges and splits will occur
- Entity resolution clients must be designed to deal with this
- Two strategies: Refer-by-Description and Refer-by-Identifier
- System status: Being evaluated by AF personnel