SW02: DO NOT LET LARGE DATASETS GO TO WASTE!

Workshop Agenda and Introduction

Yvonne Barnard
y.barnard@mail.ertico.com
ERTICO – ITS Europe

Helsinki 17/06/2014 14:00 - 15:30
FOT DATASETS

• Since 2008, the EU has supported large-scale Field Operational Tests (FOTS) of vehicle information technology. Thousands of drivers tested prototypes and products just entering the markets.

• Data has already been analysed in the projects that collected it, but there is much potential for reusing it in new studies.

• Sharing and re-using collected datasets will yield further research results, enhance cooperation between organisations and generally lead to more efficient use of resources.
DO NOT LET LARGE DATASETS GO TO WASTE!

What data do Field Operational Test projects have?

What will happen with the data?

What are obstacles and enablers for data sharing?
FOT-NET DATA

FOT-Net Data is a 3-year support action project with main objectives to:

- Support efficient sharing and re-use of FOT datasets
- Develop and promote a framework for sharing data
- Build a detailed catalogue of available data and tools
- Operate an international networking platform for FOT activities

January 2014 – December 2016
Budget € 1.8 m, EU funding € 1.4 m
Consortium: VTT, ERTICO, SAFER, IKA, CTAG, UNIVLEEDS, CEESAR, DAIMLER and 19 associated partners
AGENDA

- Sami Koskinen (VTT): FOT-Net Data
- Clement Val (CEESAR): euroFOT
- Satu Innamaa (VTT): DRIVE C2X
- Jonas Bärgman (SAFER): SHRP2
- Helena Gellerman (SAFER): UDRIVE

- Exercises on data sharing and re-using
EXERCISE 1

In a FOT data is gathered about speed, acceleration, braking behaviour, use of driver support systems, including video data and situational data.

What research questions could we answer using these data?
EXERCISE 2

• We want to answer research questions about driving behaviour in bad weather conditions

• What data would we like to have, and do the projects presented are able to provide them?
For more information and cooperation opportunities

www.fot-net.eu
info@fot-net.eu
SW02: DO NOT LET LARGE DATASETS GO TO WASTE!

Data Sharing Framework of FOT-Net Data

Sami Koskinen
sami.koskinen@vtt.fi
VTT

Helsinki 17/06/2014 14:00 - 15:30
MOTIVATION

• Sharing data will yield further research results, support education and e.g. contribute to market introduction of vehicle ICT
• Publicly funded research projects will be requested to share more of their data in the future but the practices need to be established

BENEFITS WHEN SHARING
– References to the original work and supplementing results
– New collaboration options
– Can be part of financing contract

BENEFITS WHEN RE-USING
– Access large sets of quality data
– Financial and time savings
– Post-processed indexes and shared tools for a faster start
DATA SHARING FRAMEWORK

• FOT-Net develops a data sharing framework and guidelines tailored for FOTs, addressing e.g.
  – Legal topics such as test user consent forms, participants’ privacy and topics to include in data sharing agreements
  – Documentation of key information from FOT/NDS execution and collected datasets, ensuring that the datasets can be reused
  – Financial aspects such as funding options, data release costs and arranging support for new analysts
  – Recommendations for data protection
• Framework to be agreed with and adopted by stakeholders – interviews during the congress
DATA CATALOGUE

• The new project phase will collect and add information on available FOT datasets and tools
  – Contacts data owners and publishes a data catalogue
  – Collects information on tools and data post-processing methods
  – Makes more datasets available: financial support available for high-priority datasets e.g. to improve documentation
• The catalogue supports researchers in quickly finding promising data; to include searchable key information, data samples, information on agreements and contact persons
• Initial launch around turn of the year
For more information and cooperation opportunities

www.fot-net.eu
info@fot-net.eu
SW02: DO NOT LET LARGE DATASETS GO TO WASTE!

euroFOT

Clément VAL
clement.val@ceesar.fr
CEESAR, France

Helsinki 17/06/2014 14:00 - 15:30
EUROFOT

- Comprehensive assessment of 8 ADAS with naturalistic driving / FOT approach
- Finished in June 2012
- eurofot-ip.eu
Around 1000 Vehicles

- 35 CAN + Video + Extra Sensors
- 150 CAN + Video
- 275 CAN only
- 460 Data Loggers in total

1500 questionnaires to 300 vehicles with LDW
50+ FEA trucks auto-logging
## INSTRUMENTED VEHICLES

<table>
<thead>
<tr>
<th>VMC</th>
<th>Operation site</th>
<th>CAN</th>
<th>Video</th>
<th>Extra sensors</th>
<th>Eye tracker</th>
<th>GPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>French VMC (CEESAR)</td>
<td>CEESAR (Renault cars)</td>
<td>30+5</td>
<td>●</td>
<td>● (5)</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>German VMC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operation Centre 1 (IKA)</td>
<td>FORD</td>
<td>98+2</td>
<td>-</td>
<td>● (2)</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>MAN</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>VW/AUDI</td>
<td>32</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td>Operation Centre 2 (IZVW)</td>
<td>BMW</td>
<td>15</td>
<td>●</td>
<td>●</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>DAG</td>
<td>15</td>
<td>●</td>
<td>●</td>
<td>-</td>
<td>●</td>
</tr>
<tr>
<td>Swedish VMC (CHALMERS)</td>
<td>VOLVO</td>
<td>30+5</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>VCC</td>
<td>100</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Italian VMC (POLI)</td>
<td>CRF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Questionnaires only*
VEHICLE DATA

- **CEESAR**: CAN + GPS + Extra Sensors + Video + Used
- **FORD**
- **VW/AUDI**
- **BMW**
- **DAG**
- **VCC**

- **VOLVO**: CAN + GPS + Extra Sensors + Video + Used
- **MAN**
VEHICLE DATA

Acquisition
- Raw Data

Preprocessing
- Synchronized/
  Harmonized data

External data integration
- Enriched data

Detection / Segmentation
- Events
- Situations

Characterization / Annotation
- Attributes
  Performance Indicators

Sorting / Grouping
- Aggregated data

shared within project
FACTS 1/2

• VERY rich data exists. In huge quantities and great quality.
• Not all collected data could be used within project.
• All sorts of data exist. It requires tools, know how…
• No harmonized/common database: 11 parties with different data, access rights, tools, database structure...
• Only aggregated data was shared.
FACTS 2/2

• Using data for really different purpose requires going back to continuous data → need for ‘big data’ expertise, understanding of the dataset, tools…
• Video, GPS, questionnaires are personal data (privacy issue)
• CAN data is proprietary (benchmarking, reengineering issue).
• Geographical databases proprietary as well.
DATA SHARING IN EUROFOT

- A process was set-up during the project:
  - Sharing limited to subset of data
  - No personal data (GPS, Video,…)
  - Required a thorough project proposal and acceptance from data owner

- Result (as of September 2013):
  - IKA: 3 parties asked and some analysis was carried out
  - CHALMERS: 1 request was recorded.
  - CEESAR: 4 projects done and still ongoing. 2 other requests didn’t conclude (cost, privacy)
  - Likely, data still being used at some OEMs
WHAT ARE OBSTACLES AND ENABLERS FOR SHARING DATA OF EUROFOT?

• Obstacles
  – Complexity (technical, legal…)
  – Cost

• Enablers
  – Some datasets still alive
  – Flexible tools have been developed and can be used
  – Some of the ‘data experts’ are still around
More information on euroFOT: eurofot-ip.eu

For more information and cooperation opportunities

www.fot-net.eu
info@fot-net.eu
DRIVE C2X

- DRIVE C2X has carried out a comprehensive assessment of cooperative systems through extensive European Field Operational Tests
  - Created and harmonised a European-wide testing environment for cooperative systems
  - Final event July 16-17 in Berlin
DATA IN DRIVE C2X

• Type of data collected
  – Logged data (CAN + GPS)
  – Questionnaire answers, focus groups
  – Backseat observation data, front view video, driver video
  – Metadata on tests, test sites, vehicles, participants

• Data sharing within DRIVE C2X
  – 7 test sites including simTD, SCORE@F, SISCOGA
  – Data collected at central data server at Fraunhofer FOKUS, validated, processed and delivered in a harmonised format to the analysts
WHAT WILL HAPPEN WITH THE DATA FROM DRIVE C2X?

• No formal agreements on data sharing or reuse possibilities after the end of DRIVE C2X
• Data owned by partners who set up the tests
• DRIVE C2X partners likely to re-use the data
• FOT Net Data supports data re-use
  – Decisions case by case
WHAT ARE OBSTACLES AND ENABLERS FOR SHARING DATA OF DRIVE C2X?

• Obstacles
  – No agreement on data sharing, no funding for maintaining or sharing data
  – Data owned by several organisations

• Enablers
  – Data already validated
  – As data sharing already took place in DRIVE C2X project, good metadata exists in English
  – Tools exist for processing of raw data coming in different formats from different DRIVE C2X test sites
EXAMPLE OF DATA THAT COULD BE RE-USED

• Driver behaviour data in baseline conditions (no cooperative functions in use)
• User acceptance data
• …
More information on DRIVE C2X:
www.drive-c2x.eu

For more information and cooperation opportunities
www.fot-net.eu
info@fot-net.eu
SESSION SW02: DO NOT LET LARGE DATASETS GO TO WASTE!

SHRP2 Data

Jonas Bärgman
jonas.bargman@chalmers.se
Chalmers University of Technology/SAFER

Helsinki 17/06/2014 14:00 - 15:30
SHRP 2 NATURALISTIC DRIVING STUDY & ROADWAY INFORMATION DATABASES

Data from 3,147 volunteer drivers in six sites
- Passenger cars, vans
- SUVs, pickups

New data collected
12,500 centerline miles consistent across six sites

Acquired data (DOTs, others)
- 200,000 centerline miles
- Roadway, weather, traffic ..
Largest naturalistic driving study ever undertaken
- 3,147 drivers, all age/gender groups.
- 3,958 data years; 5 M trip files; 50 M vehicle miles
- 3 years of data collection
  - Most participants 1 to 2 years
- Vehicle types: All light vehicles
  - Passenger Cars
  - Minivans
  - SUVs
  - Pickup Trucks
- Six data collection sites
SHRP2 NDS DATA OVERVIEW

• **Size:** The largest NDS data set collected
  – 2 petabytes = 2 million 1 GB flash drives (1.2 PB video, 0.8 PB sensor)
  – “Give me the whole raw data file” isn’t possible or sensible

• **Complexity:** different data types
  – Categorical data constant over a trip: driver age, vehicle type
  – Sampled data: collected at original resolution (once a trip up to 640 Hz during a crash): speed, acceleration, GPS position, radar, vehicle network information
  – Video data from 4 cameras; must be coded
    • Automated reduction: lane tracker
    • Manual reduction: all other items for specific analyses
SHRP2 RID DATA OVERVIEW

• **Size**: the size is manageable
  – 50-60 GB without video, 6-8 TB with video

• **Complexity**: 4 different data sources
  – ESRI: baseline data for entire country
  – State roadway inventory data: from 6 study States; data vary by State
  – Mobile van data: very detailed, about 12,500 centerline miles; 43,195 intersections, 518,570 MUTCD signs; includes forward video
  – Supplemental data: from 6 study States, data vary by State

• **NDS and RID are linked**: match trips with road segments; match road segments with trips
WHAT WILL HAPPEN WITH THE DATA FROM SHRP2?
DATA AVAILABILITY

- **NDS data collection completed**
  - April 2014: 4.7 million trips through quality control
  - Quality control rest of trips; add radar, cell phone data by December 31, 2014

- **RID data collection completed**
  - Complete quality control, add supplementary data by October 31, 2014

- **Trip summary files**
  - April 2014: 1,143,033 trips, not all variables
  - December 31, 2014: complete

- **Crash, near-crash, and baseline event and epoch files**
  - April 2014: 113 crashes and near-crashes
  - December 31, 2014: all crashes, near-crashes, baseline

- **Linking NDS and RID data complete December 31, 2014**
WHAT WILL HAPPEN WITH THE DATA FROM SHRP2?
2015 AND BEYOND: PHASED APPROACH

Phase 1: up to 5 years, begins January 2015

- Test approaches to management, security, support, etc.
- Make data widely available, provide user support
  - General: website, help desk, documentation, tools
  - Individual: consultation, customized data, PII data access
  - Information: training, user groups, symposia
- $25 million available to TRB from SHRP 2 implementation funds
- Preparation for Phase 1 will continue throughout 2014
WHAT ARE OBSTACLES AND ENABLERS FOR SHARING SHRP2 DATA?

Enablers

- Re-use of the data present from the start of the project
- Funding ($25 million) available
- Methods for data sharing tested/used in project
- Proposed structure for hosting the data
  - Single primary data center
  - Remote access centers
- Website for initial easy access to data
WHAT ARE OBSTACLES AND ENABLERS FOR SHARING SHRP2 DATA?

Obstacles

• Video crucial for validation. Personal Identity data such as video and GPS only accessible at primary data center and remote secure enclaves.
• Funding distribution and what it should cover not yet known.
• Data sharing agreements and IRB approvals will always be required, although this process could be delegated to multiple hosts.
WHAT ARE OBSTACLES AND ENABLERS FOR SHARING YOUR DATA?
SHRP2 WEBSITE

Content on website

- Study description and overview video
- Data access guide
- Data dictionaries
- Sample data
- Data
  - Descriptive data for whole data file
  - Categorical data on trips, from trip summaries; query capability
  - Event data from crashes, near-crashes, baseline
  - Viewer - forward video and time series data for crashes, near-crashes
  - NDS/RID linking tables (files)

https://insight.shrp2nds.us/
EXAMPLE OF DATA THAT COULD BE RE-USED
SHRP2 NDS
CRASH, NEAR-CRASH, BASELINE FILES

- Crashes: expect 700, varying severity
- Near-crashes: “almost” crash but for …; 7,000
  - Crash surrogates; how did driver avoid a crash
- Baseline: randomly selected across all vehicles; 30,000
  - Denominator for risk calculations; measure overall prevalence
- Event files for each – categorical data
  - Coded from last 6 seconds of data before precipitating event
  - Manual video reduction (e.g., driver distraction)
  - Manual eye-glance coding
- Epoch files for each – sensor and video data
  - 30-second data segments (20 before precipitating event, 10 after; only 20 for baseline)
  - Manual eye-glance coding
More information on SHRP2: 
https://insight.shrp2nds.us

For more information and cooperation opportunities 
www.fot-net.eu
info@fot-net.eu
SESSION SW02: DO NOT LET LARGE DATASETS GO TO WASTE!

UDRIVE Data

Helena Gellerman
helena.gellerman@chalmers.se
SAFER

Helsinki 17/06/2014 14:00 - 15:30
UDRIVE – EUROPEAN NDS

• 19 partners from 11 countries

• Oct 2012 => Sept 2016 (4 years)

• 7 test sites

• Analysis across sites

• Data available for post-project research
UDRIVE DATA

• ~500 000 hours (120 cars, 50 trucks, 40 PTW)
• 21 months data collection
• Data
  – 8 video views (cars+trucks, 5 för PTW)
  – CAN (speed, steering angle, etc.)
  – GPS
  – Accelerometers and angular rate sensors (3-axis)
  – Map data
• Post-processed data related to:
  – risk calculations (Safety Critical Events), normal driving, inattention/distraction, vulnerable road users (bicyclist, PTW and pedestrian), eco driving
WHAT WILL HAPPEN WITH THE DATA FROM UDRIVE?
REMOTE ACCESS TO DATA TESTED IN PROJECT

- 7 test sites
- Central Data Center (CDC) located at SAFER
- Analysis performed remotely on the common dataset
- Set-up of infrastructure for remote access including data protection
WHAT WILL HAPPEN WITH THE DATA FROM YOUR PROJECT?

POST-PROJECT ACCESS TO UDRIVE DATA

- Central Data Center (CDC) located at SAFER hosts all data.
- Partner can download data and become Partner Data Center (PDC). Their decision if data open for Third Parties.
- Third party can remotely access all data from CDC, except Personal Identity Data (PID).
- PID is accessed at the premises of either the CDC or the PDC.
- Third parties must apply to analyse data and also adhere to the Data Protection Concept.
- All data remains in original storage. Data for publications can be extracted based on consent from drivers.
- Post-project data only available if funding is solved.
WHAT ARE OBSTACLES AND ENABLERS FOR SHARING UDRIVE DATA?

Enablers

- Re-use of the data agreed in project documents
- Remote access method tested in project
- Application procedure including Data Protection Concept
- Website – discussion; FOT-Net Data “Data Catalogue” which is under development

Obstacles

- Funding for post-process data storage and access not yet solved
- Data protection of PID, but necessary
EXAMPLE OF UDRIVE DATA THAT COULD BE RE-USED

- Time series data and video (Car/truck/PTW data)
- Safety Critical Events (near-crash and crash events)
- Baseline epochs (annotations etc)
- Intersection/VRU related annotations
More information on UDRIVE: www.udrive.eu

For more information and cooperation opportunities www.fot-net.eu info@fot-net.eu
EXERCISE 1

In a FOT data is gathered about speed, acceleration, braking behaviour, use of driver support systems, including video data and situational data.

What research questions could we answer using these data?
EXERCISE 2

• We want to answer research questions about driving behaviour in bad weather conditions

• What data would we like to have, and do the projects presented are able to provide them?
Please fill in our questionnaire on data sharing

If you are available for an interview, contact us

For more information and cooperation opportunities

www.fot-net.eu
info@fot-net.eu