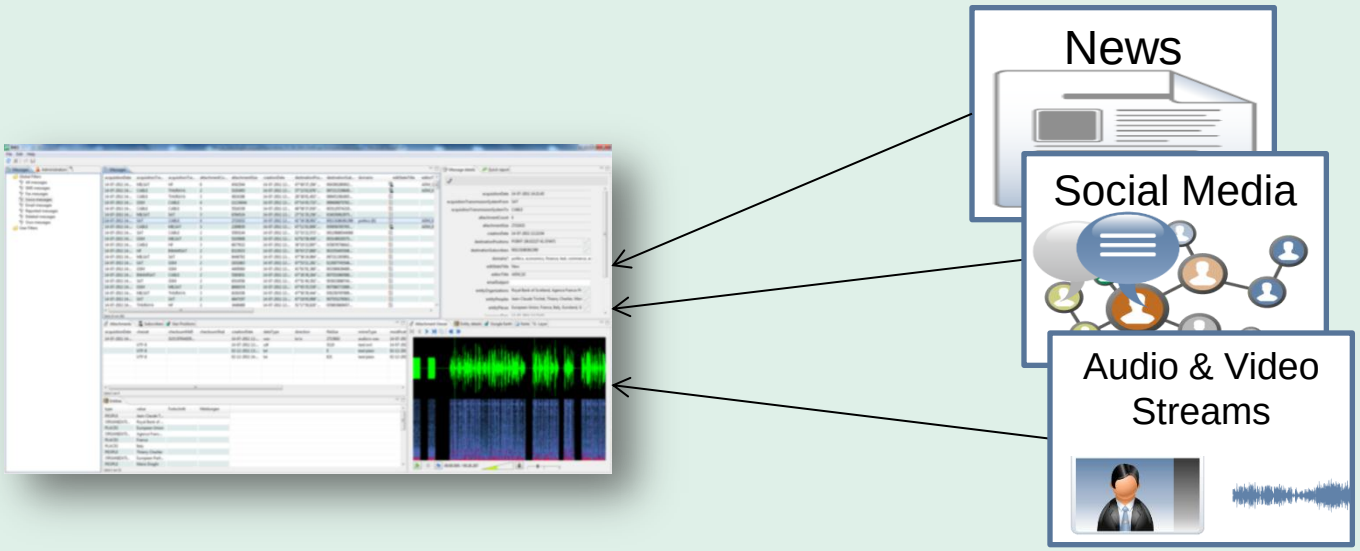


Detecting Structures in Massive Data – Using the Example of News Channel Intelligence Gathering

Niels Beuck niels.beuck@plath.de

PLATH

Scenario

News channel intelligence gathering

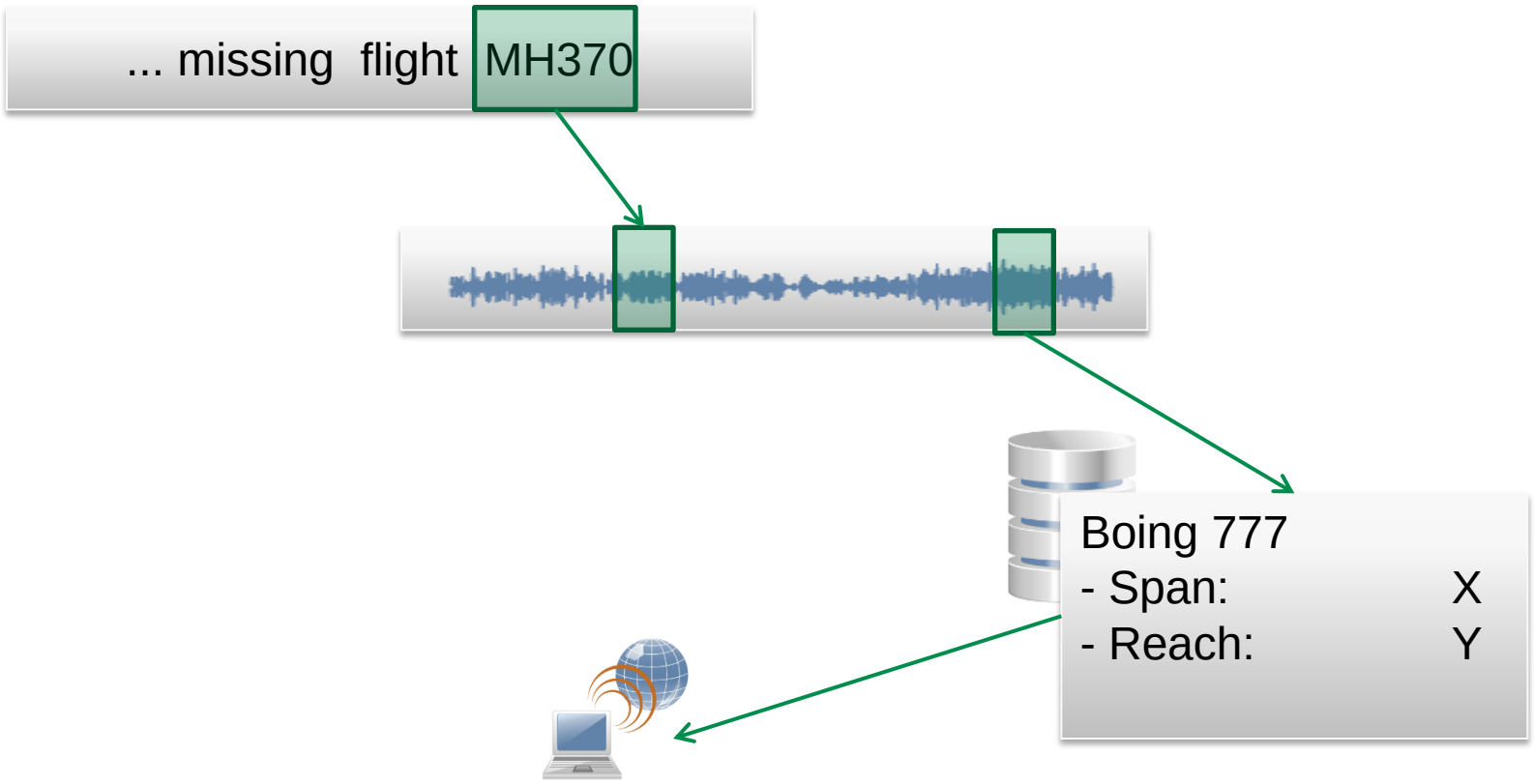
- Filtering and triggering news on customer specific profiles as used by, e.g., top executives and news agencies
  - current example topic: search for flight MH370
- Based on different channels like
  - online news
  - or social media
- We want to
  - get an overview over the situation
  - find relevant information
  - initiate additional investigation
  - generate an article or report

Improve Situation Awareness from Multiple Data Streams

The task is defined by topics, not by the source

PLATH

Finding Relevant Information via Data Links



Open Data

Where does the data come from?

- Event streams
  - e.g. news feeds
  - current information driving the search
  - filtered, e.g. by topic
- Archived data
  - relevant when linked by
- LOD knowledge bases
  - background information

PLATH

Linked Data

Where do the links come from?

- explicit links
- implicit links
  - via entities
    - identify entities by semantic analysis
  - to an ontology
    - align data to an ontology
    - link to knowledge bases using that ontology
  - by topic
    - automatically identify the topic by the content
  - by origin
    - link information from the same origin, regardless of the channel

PLATH

Value Added to the Events

- Relevance:

  Relevant information (Semantic Search and Browsing)

- Normalization:

  Semantic information interoperability and –integration

- Relations:

  Semantic correlation, Mining, Analysis, Discovery, Early Warning

PLATH

Agenda

- Motivation: Scenario and Goals
- Challenges & Architectural Requirements
- Experimental Platform: IMES
- Open Issues

PLATH

Challenges

- Large amount data
  - → scalable architecture
  - → powerful search and filter tools
- Heterogeneous data sources
  - → make data comparable
  - → link information from multiple sources
    - to background knowledge (e.g., spatial)
    - to each other
- Unstructured data
  - → make data searchable
    - e.g. transcription of video and audio data
- Redundancy
  - → detect and filter redundant data

PLATH

Working with the Data

Problem:

„How should I know what I am looking for before I see what I found?„

- We can’t decide locally what is relevant without considering the big picture
- But: the big picture emerges by analyzing relevant data

- -> The notion of relevance changes as the big picture emerges

PLATH

Divide-and-Conquer Approach?

- Single-Channel/Multi-Level
  - Independent flows from sources via validation to higher level analysis
- Nearly impossible to find correlation across channels
  - Separated data silos without links
- Pushing data via reports only for topics which *seem* relevant
  - Local decisions with no eye for the Big Picture

→ Severe drawbacks!

PLATH

Explorative Topic-oriented Approach

- Multi-Channel
  - Merging Data from multiple sources
  - Enables link traversal
- Explorative
  - Iteratively refining filter and search queries
  - Access to the individual news events needed, not just aggregated data

PLATH

Data Source Hopping

- Relevance of information becomes apparent dynamically
  - -> new sources become relevant and need to be integrated,
  - others become irrelevant
- In the MH370 example, we might want to add
  - other news channels
  - flight schedule data
  - or a technical airplane database

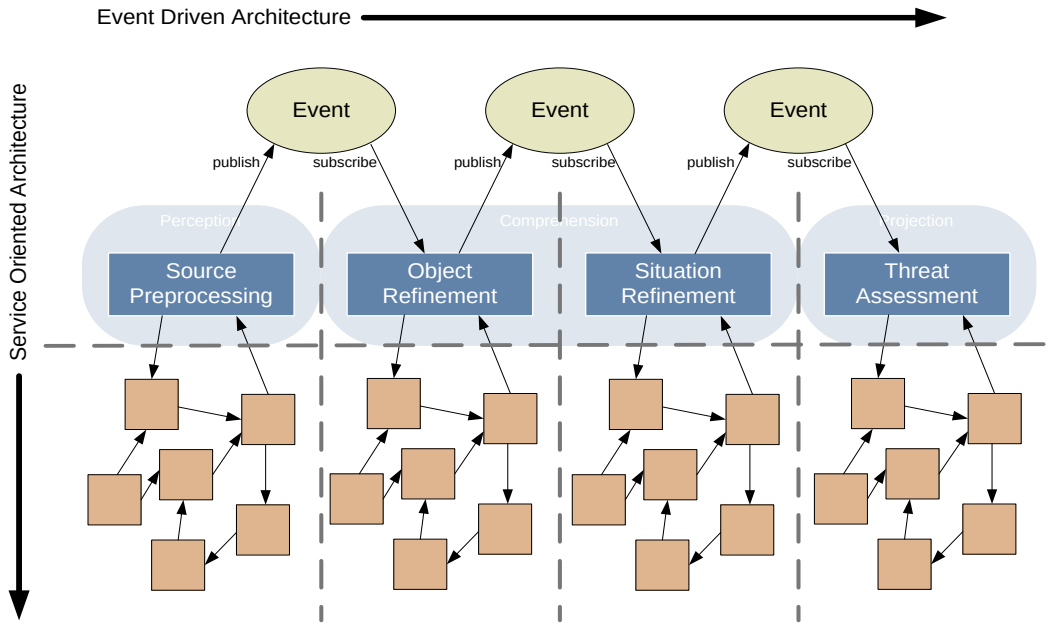
PLATH

Key Requirements for an Architecture for Data-Stream Analysis

- Flexibility to incorporate new data streams easily
  - track different channels like online news or social media
  - Add new sources or services on the fly
- Flexibility with respect to processing steps
  - Channel-specific pre-processing
    - Transcription of audio
    - Extraction of metadata
    - Removing markup
  - Geo locating the event
  - Semantic analysis
    - Topic detection
    - Entity recognition
- Scalability to mass data

PLATH

Technical Enablers: Events and Services



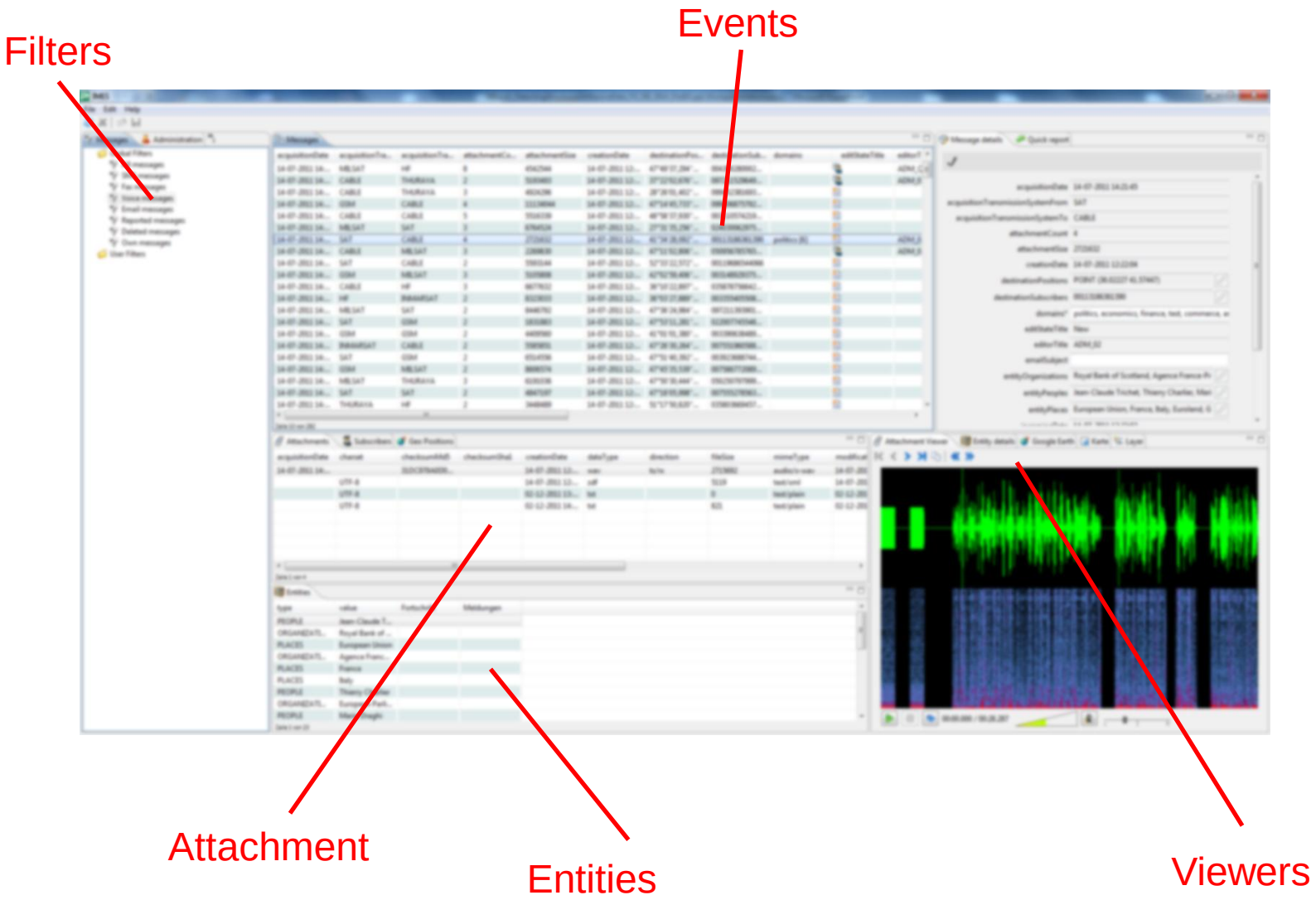
- Information gathering from data streams
- Event aggregation: Intelligent merging to Common Stream of Information (CSI)
  - Provide tracking by event locating
- Combine event-driven architecture and service orientation
  - work event-driven
  - realize logic as a service (e.g. to encapsulate private knowledge, to easily add new features)
- CSI and KB act as blackboards for information sharing

IMES – An Experimentation Platform

- IMES - Intelligent Merging and Evaluation Suite
  - is a shell for adaptable analysis solutions
  - supports
    - source registry
    - authorization and authentication
    - access to knowledge bases and CSI
  - Common Components
    - Content viewers, Maps, Networks,

PLATH

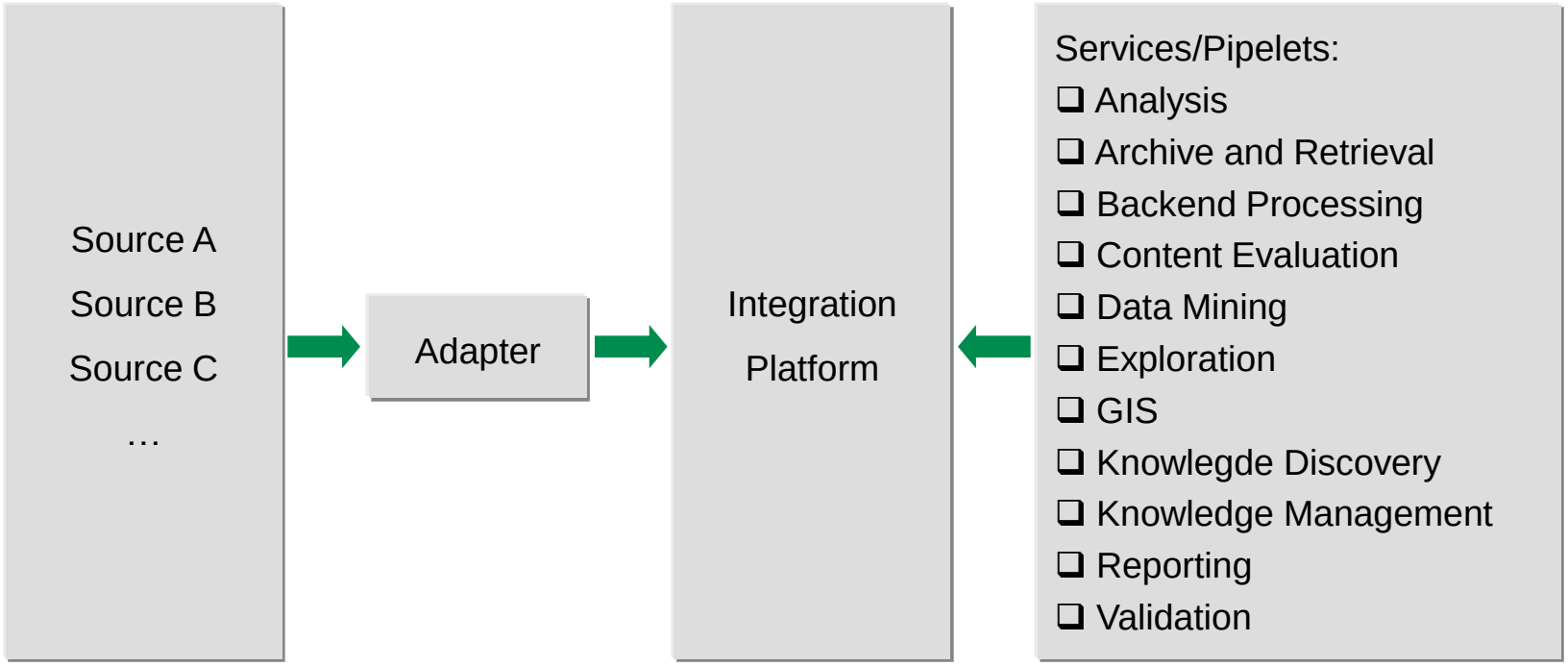
IMES-Client



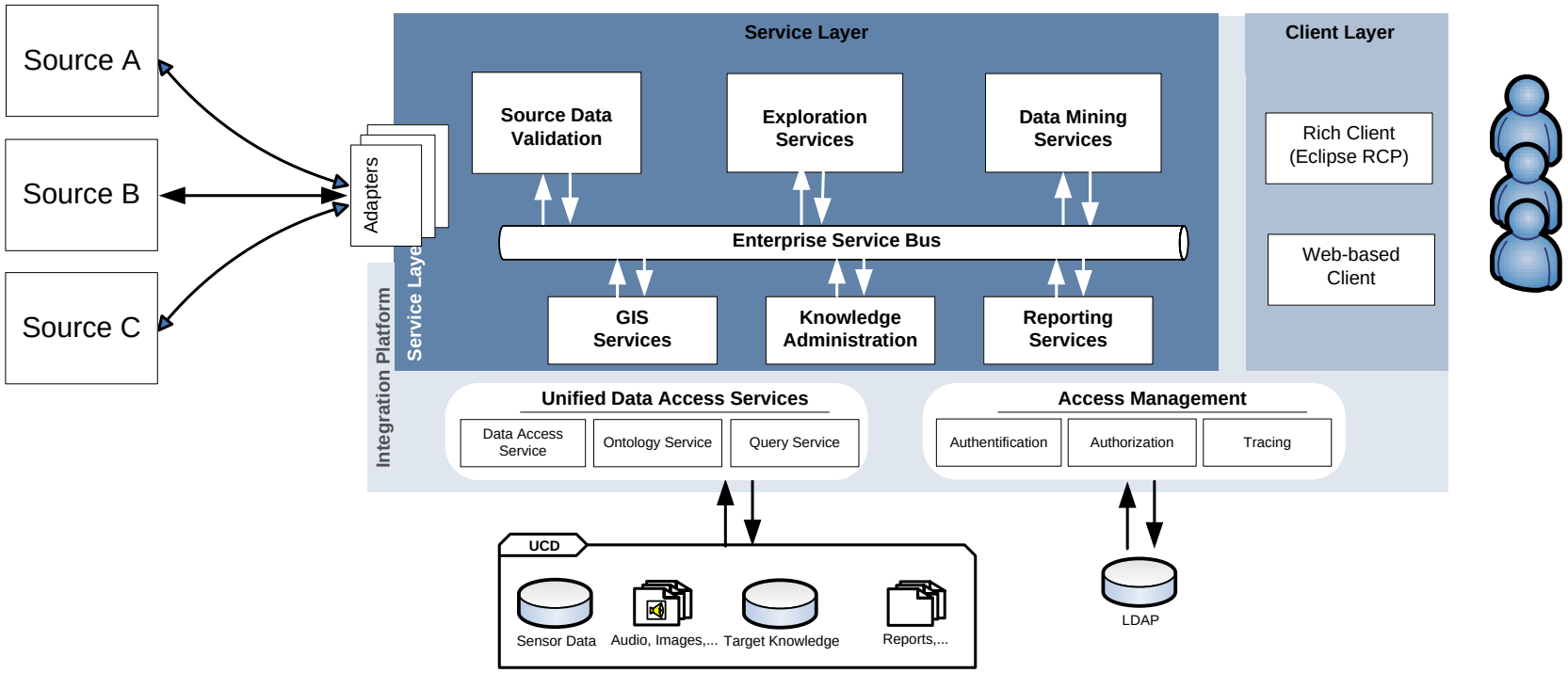
PLATH

Architecture Overview I

- Common platform with exchangeable functional units and modules
- Atomized single functions in applications and flexible packing
- Customer-individual configuration of software systems



Architecture Overview II



Configurability

- Modular and configurable architecture
  - Implementation of specific clients per project mainly by configuration
  - Configurable event sources
    - pipelining concept for automatic preprocessing
  - Different content viewers
    - Map or google earth for spatial data
    - Audio player
    - Browser
    - Graph viewer
  - Pluggable third party web-services
    - for data analysis and enrichment

PLATH

Integrated Third Party Services

- Semantic text analysis
  - Named Entity Recognition
  - Topic classification
  - -> Entities and topics are added to the event, usable in faceted browsing

- Speech analysis
  - Transcription service
    - -> makes audio content searchable
    - -> enables text analysis
  - Keyword spotting
    - -> alerting
  - Language recognition
    - -> for filtering

PLATH

Integration of New Data Sources

- As an event source
  - Pipeline configuration with camel in XML
    - Consisting of adapter and pre-processing step
- As a service
  - Manually triggered or in the pipeline of an event source
  - Example:
    - Annotate technical data to airplane models mentioned
    - Annotate flight schedule information to time points

PLATH

Elements of an Event

- Metadata
  - like time and source
- Attachments
  - like media files
- Annotation
  - tags
  - semantic annotation
- Links
  - to other events

PLATH

Blackboard

- Building and sharing the big picture
  - In the event table
    - Tag relevant events
    - Enrich them by traversing links to related information
  - In the knowledge base
    - Add new entities and relations into the knowledge base

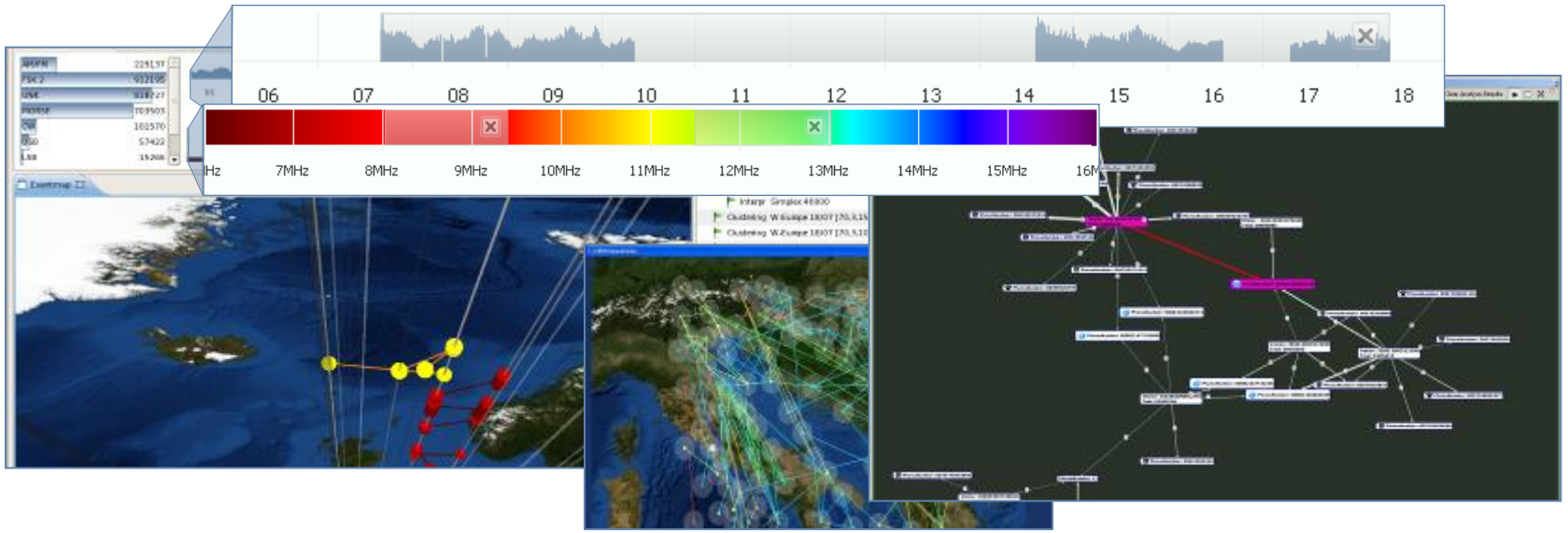
PLATH

Link Establishment

- Manually
  - via filters in the client
- In the pre-processing steps
  - Linked to and by semantic information
  - followed by comparison to other events and the knowledge base
- Advanced Link Establishment
  - Custom links between events
  - Constraint system for model definition
  - Finding new relevant events based on a given relevant event
  - Finding meta events consisting of several events
    - Even without identifying the individual events as relevant first

PLATH

Advanced Navigation
Visual Analytics, making sense of massive data streams



- Combine automation (e.g., data correlation) and human interpretation
- Use visual navigation
  - Spatial and temporal attributes to visualize events
  - Faceted browsing for data exploration

PLATH

Open Issues

- Representation of enrichment and linked data
  - Flexible enough to cover a broad range of data types
  - Restricted enough to be able to find links between data
- Quality of Analysis
  - depends highly on the type of content
  - error propagation (e.g. Transcription plus text analysis)

PLATH

Summary

- To establish a big picture from news data we need an explorative, multi-channel workflow.
- Data from different channels needs to be made comparable, links need to be established
- IMES is a configurable platform suitable for exploration and evaluation of different technologies
- There are open issues like data representation

PLATH

PLATH GmbH

Gotenstraße 18 ▪ 20097 Hamburg ▪ Germany
Tel.: +49 40 23734-0 ▪ Fax: +49 40 23734-173
Email: info@plath.de ▪ www.plath.de

PLATH