Challenges for Big Data Processing: Dealing with the Vs Features

Petre DINI

IARIA

petre@iaria.org

Plan

- Big/small/linked/open data
- Data collection/filtering/storing
- Data retrieval/selection/interpretation
- Domains

- Case Study
- To come

Where to start from?

- **For existing systems; mostly, non-standard approaches**
- **- forensic**
- **- facial recognition**
- **- atmospheric models**
- **- universe models**
- **- health records**
- **- human behavioral models**
- **- etc.**

- **For new systems; they must be standardized**
- **- syntax**
- **- semantic**
- **- taxonomy**

Data/Support types

- **Data representation**
  - **A la SQL**
  - **NoSQL**
  - **Pictures**
  - **Voice**

- **Support digital (memory)**
- **Support digital (tapes)**
  - **Pros/cons**
  - **(Google story!)**

Data correlation

- **Hierarchical data**
- **Distributed/Isolated, Linked data**
- **Web / Deep Web [:) Web sommerso!]**

  Invisible Web, Dark Web, Hidden Web [not indexed]

- **Data features**

  Primary [P]

  Secondary [S]

  D :: = {[P] [S] [context]}

BIG | the Vs | 3v, 5v, 7v, 10v, …. ?

- **Volume** (length of a records, # of records) (entity-relationship databases)(datasets) || BIG vs. HUGE
- **Variety** (types: strings, pictures, voice, etc.) (structured, non-structured)
- **Veracity** (precision and accuracy of data)
- **Velocity** (of change)
- **Value** (as a business/service) IMPACT
- **Volatility** (temporary; quick action)
- **Vasting resources**
  (storage, computation, transfer)
- **Viability** (are data still useful?)
- **Visibility** (open, hidden, ..)
- **Validity**
  (are there still valid/updated data?)
  (in context validity)
  (e-government datasets)

- incomplete
- redundant
- inconsistent
- noisy

quality of data

filling missing values with estimated values calculated for complete records of the same dataset

Last News, on New Book on Veracity of Data

- **from: From: Brent Beckley beckley@morganclaypool.com**
- **Wed, 13 Apr 2016 09:52:14 -0500**

- **Veracity of Data: *From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*
by Laure Berti-Equille, Javier Borge-Holthoefer *(Qatar Computing Research Institute)*
ISBN: 9781627057714 | PDF ISBN: 9781627057721
Copyright Â© 2016 | 155 Pages | Publication Date: December 23, 2015
Retail Store (print & individual copies)
Digital Library (subscribing institutions)**
- **On the Web, a massive amount of user-generated content is available through various channels (e.g., texts, tweets, Web tables, databases, multimedia-sharing platforms, etc.). Conflicting information, rumors, erroneous and fake content can be easily spread across multiple sources, making it hard to distinguish between what is true and what is not. This book gives an overview of fundamental issues and recent contributions for ascertaining the veracity of data in the era of Big Data.**
- **The text is organized into six chapters, focusing on structured data extracted from texts. Chapter 1 introduces the problem of ascertaining the veracity of data in a multi-source and evolving context. Issues related to information extraction are presented in Chapter 2. Current truth discovery computation algorithms are presented in details in Chapter 3. It is followed by practical techniques for evaluating data source reputation and authoritativeness in Chapter 4. The theoretical foundations and various approaches for modeling diffusion phenomenon of misinformation spreading in networked systems are studied in Chapter 5. Finally, truth discovery computation from extracted data in a dynamic context of misinformation propagation raises interesting challenges that are explored in Chapter 6. This text is intended for a seminar course at the graduate level. It is also to serve as a useful resource for researchers and practitioners who are interested! in the study of fact-checking, truth discovery, or rumor spreading.**

BIG

- **Complexity, Security, Risks to privacy**
- **Complex links, (fuzzy links, context-based links)**
- **Mixed Formal/Informal features**

⇔ defined fields (syntax) / text-like information

Note:

- **90% of world’s data was generated in the last two years**
- **> 90% is unstructured**
- **+ Web ad Cloud offer new possibilities for discovery**

⇒ New technologies:

For extracting/transforming/loading (ETL) and processing

For cleaning and organizing unstructured data in big-data applications; e.g., Hadoop

Mixed media support

Tools

- **MapReduce**
- ***MapReduce* is a programming model and an associated implementation for processing and generating large data sets**
- **Hadoop**
- ***Hive Hadoop Component is used for completely structured Data; Hive Hadoop Component is mainly used for creating reports***
- ***Pig Hadoop Component is used for semi structured data. Pig Hadoop Component is mainly used for programming.***
- **OLAP – reporting tool**
- **OLAP (online analytical processing) is a function of business intelligence software that enables a user to easily and selectively extract and view data from different points of view**

Source

- Sensors
- System [any] reports
- Neural/body systems
- Atmospheric measurements [short, medium, long terms]
- Universe observations [long term]
- Health measurements [small + big,…]
- Social measurements [migrations, resources, etc.]

- collections
  - Raw
  - Partially processed

Retrieval

- Clustering
- Partitioning
- Summarizing
- Fusion
- Compressing

- ? selecting the right features

- Datasets [selected, validated, ….]

Bumps

- Noise
- Probabilistic data
- Fuzzy-data sets
- Incompleteness

- Time-sensitive
- Time-free

- Timestamps
- Hierarchical timestamps
- Timestamps: [source][storage][processor][console]

    ? Clock synchronization; No-clock entities

Storage

- **Distributed**

- **Access**
  - **Internet Neutrality**
  - **Accessibility**

- **Transparency Degree**
  - **Open [e-government]**
    - **OpenData Government**
    - **OpenData Forum**
  - **Private [financial, health]**

Status

- **Yesterday**
- **Today**
- **Tomorrow**

- **#1: big data exist**
- **#2: big data was dealt with**

- **? → classical "hype" case**

Applications i [SMALL data]

- **Using Patient Data for Personalized Cancer Treatments**
  - **- improve health outcomes**
  - **- support development of new therapies**
- **Small Data**
  - **Seeking personalized data-derived insights from analysis of our digital traces**
  - **Personal devices Internet services for self-tracking**
    - **Fitbit**
    - **Patients like me**
    - **http://quatifiedself.com**
  - **Digital traces accumulated by social networks, search engines, mobile operators, online games, e-commerce**

Applications ii [SMALL data]

- ? → regulatory challenges /FDA, HIPAA, privacy policies

  Health Insurance Portability and Accountability Act

- Open mHealth http://openmhealth.org
- http://smalldata.tech.cornell.edu

Application i [BIG data, bug traces]

- Large-scale bug traces
- Testing network devices before releasing them
- Binary/Linear Downsizing -> Downsizing Ratio
- Reproducing failures to facilitate the debugging process, real-world traffic needs to be captured and later replayed
- ➔ high volume [peak-hour, at Beta Site, 20Gbytes, 30 minutes]
- Remove data redundancy in large a trace
- Note
- Linear Downsizing: rollback-and-reply; whenever a failure is triggered, the failure would be logged and the sequential traces triggering the failure are regarded as a whole and divided into equal-sized pieces of traces from the beginning based on a predefined size, rollback size.
- Binary Downsizing: BD locates the sequential traces triggering the failure by recursively splitting the traces on halves and replying the smaller ones in turn, until the failure is missed….

Applications ii [BIG Data]

- Government Sector
- Ref: Communications ACM, 03/2014, vol. 57, no. 03
- BIG Data initiatives
- Japan: ITS [Intelligent Traffic System], Info-plosion, MEXT/NSF [Education..]/NSF
- UK: HSC [Horizon Scanning Center]
- Singapore: RAHS [Risk Assessment and Horizon Scanning]
- Korea: KOBIC, MFAFF, MOPAS
- EU: DOME [The Netherlands, Switzerland, UK, + 17 countries] + IBM /supercomputing center to handle a data set in excess of one exabyte per day derived from SKA radio telescope
  - Exascale computing, transport and storage
  - Analyze all raw data collected daily (observable universe)
  - (One exaflops is a thousand petaflops or a quintillion, 1018, floating point operations per second.) At a supercomputing conference in 2009, Computerworld projected exascale implementation by 2018
- US: Genome Data on AWS [Amazon Web Services], CDC, NSF/NIH: BIGDATA, US Michigan [Statewide Data Warehouse]

Applications iii [BIG data]

- **Local Governments**
- **2011, Syracuse (NY) + IBM → Smarter City**

  Bid data to help predict and prevent vacant residential properties
- **Michigan's Department of Information → data warehouse**

  To provide a single source of information about citizens of Michigan to multiple government agencies and organizations to help provide better services
- **Facts**

  ? E-Coli story

  ? Driver License story

Case study

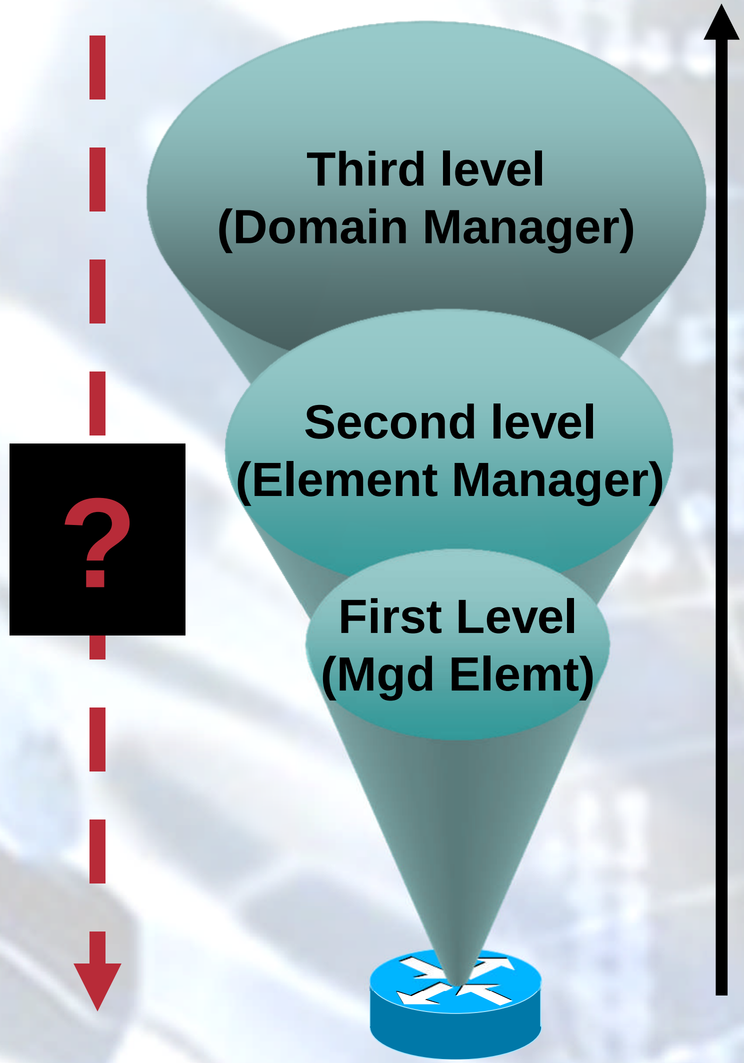
Positioning

Issues

- Event definition
- Event transport
- Event processing
- Business-driven events

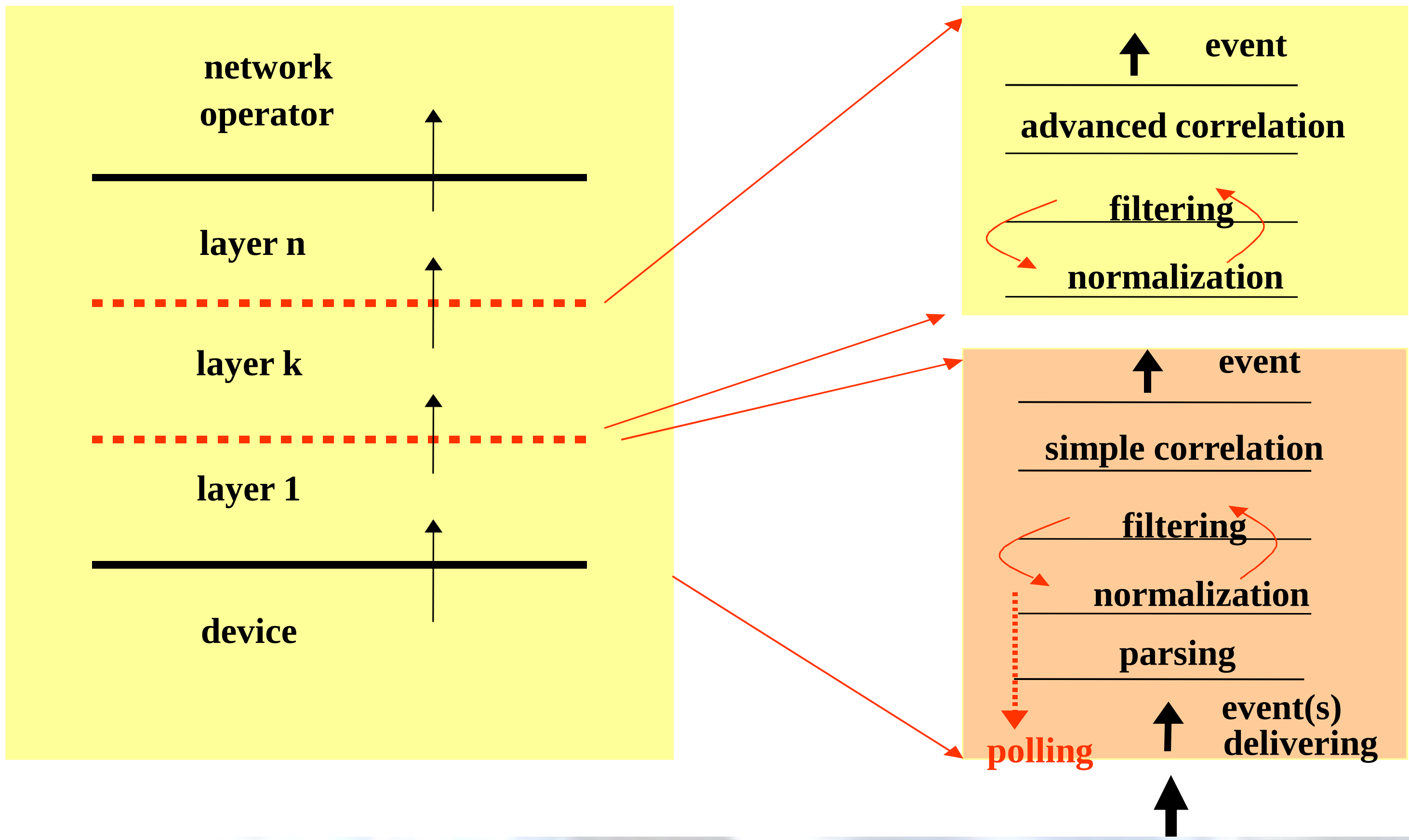
Bottom-up vs. Top-down

- Domain Manager enriches with domain information
- EMS enriches with multi-device information
- Notification Engine collects OS notifications



- Event Information + Device Information + Domain Information
- Event Information + Device Information
- Event Information

A Layered Processing View



Log Files, Syslog, MIBs, RMON, etc.

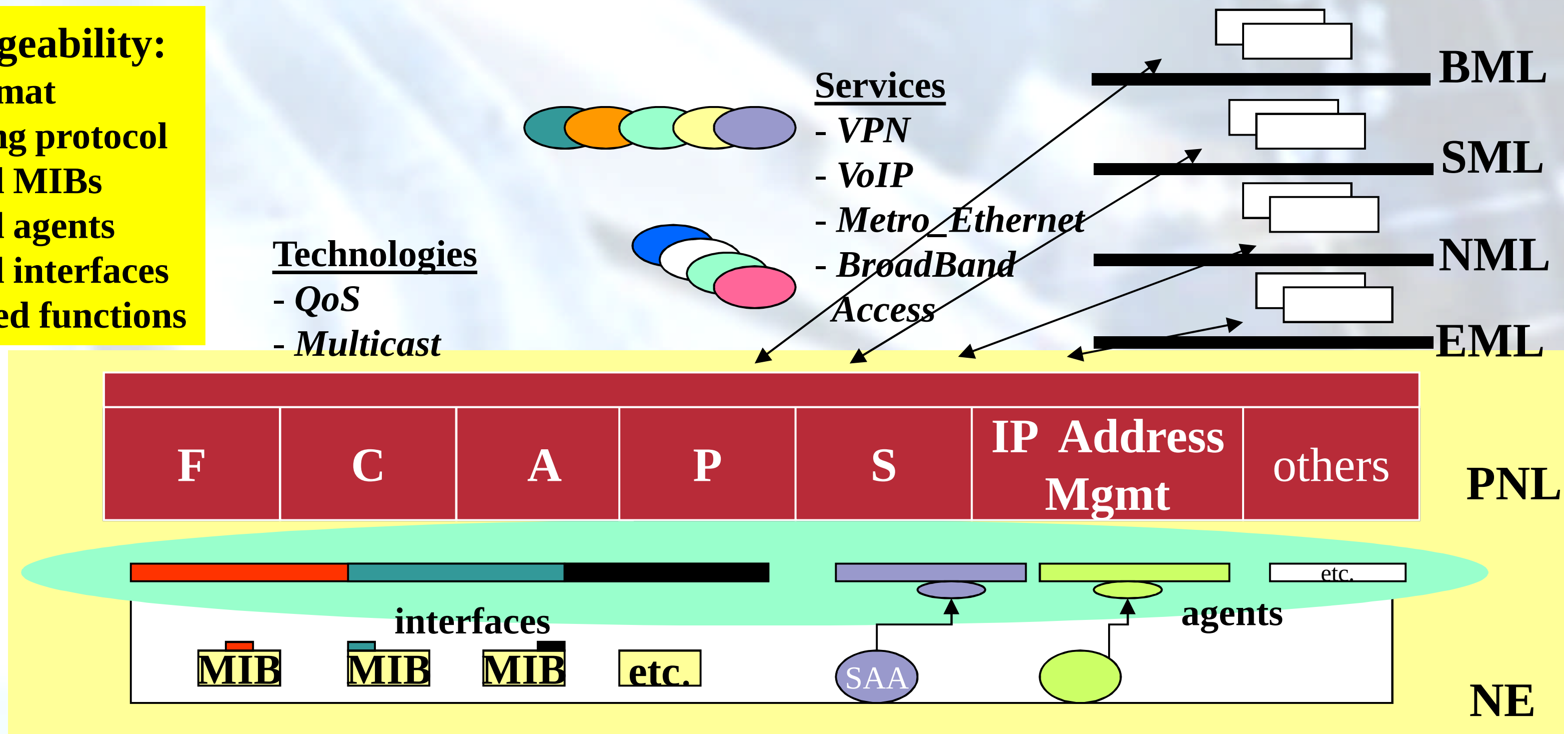
Challenging Issues

Too Many

Syntax Issues

- Various formats
- Myriad of conversions needed
- Lack of syntax control

NE Manageability:
- ? data format
- ? conveying protocol
- ? required MIBs
- ? required agents
- ? required interfaces
- ? embedded functions



Syslog Message “Body” Format in the IOS

CONSOLE

* Sep 20 01:12:31: %SYS-5-CONFIG_I: Configured from console by vty1 (144.254.9.79)

Timestamp	IOS Component	Severity	Mnemonic	Message-text

Timestamp from the server

SERVER

Sep 20 01:07:00 router.cisco.com 571: Sep 20 01:12:31: %SYS-5-CONFIG_I: Configured from console by vty1 (144.254.9.79)

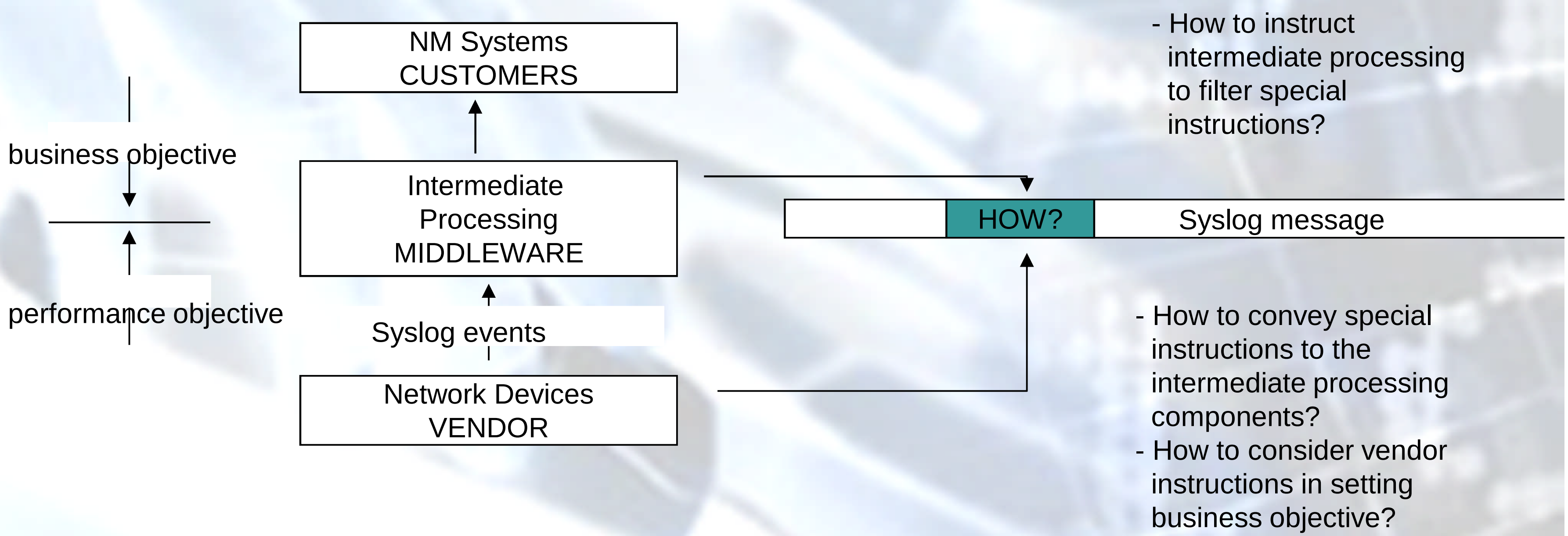
Router

Timestamp from the router

- **NTP is needed!**
- **Header:level can be different than Body:severity**

Semantic Issues

- **Naming**
- **Context-defined**
- **Smart events**



XML Tagging is Not Enough

% *versus* <XML>
: <<<a><b>><c>>
: (((a)(b)c))

1. <a> <b> <c>
? ? ?
2. <<a> --- r1 -- <b>> -- r2 -- <c>
? ?

e.g.,
<a> -- Interface (? OID)
<b> -- Port (? OID)
<c> -- Severity

?
Tag table (??)
Tag List:
<name><semantics>

?
Tag relationships

?
Naming service required

- Despite the problems caused by its use:
- – The messages don't have a standardized definition
- – Priority is geared toward UNIX problems
- – Priority is not used consistently
- – Not reliable
- – Not secure
- some key features, (i) ease of use for developers, (ii) familiarity, and (iii) ubiquity makes it a workable solution.

Timestamps issues

- Format
- Clock-free event sources
- Sources-destination timestamps
- Delay tolerant networks
- Localizing processing
  - Local synchronization
  - Wide synchronization
- Reliable timestamps

Example: Syslog

[field1] % [field2] % [severity] % [priority]%[mnemonic] %[free form field]

Well identified fields
[timestamps]
[facility]
[severity]
[priority]
[mnemonic]

Free form field (the richest in semantic)
[..English plain text..]

Field separator
%

Issues
- Number of fields varies
- Value space of the fields is is not uniform/standardized
- Semantic of timestamps is not uniform/or not defined
- Mnemonic is not modeled

- The English text is only humanly readable/useful

- Automation is difficult due to the “natural language processing” needs

Things started to get fixed

- Syslog, SNMP/MIB: IETF
- Adaptive message format: IBM/Cisco
- Intrusion detection format: IETF
- Anomaly report format: OASIS
- Incident handling format: IETF

- NGN management : ITU-T [Focus group]

Still to answer…

- **Concepts such utility-based computing, autonomic computing, diagnosis-in-the-box, diagnosis out-of-box, adaptable applications, self-adaptable applications, and reflexive environments require a new approach of formalizing events, architecting event-based systems, and integrating such systems.**
- **Additionally, GRID systems bring into the landscape the concept of intermittent and partial behavior related to resource sharing that may require a special semantic on SLA/QoS violation events.**
- **Events related to traffic patterns and the dynamics of performance and availability changes in such environments requires particular metrics and processing, as well [accounting, outage].**
- **Another hot area quite poorly covered in terms of event-related requirements is MPLS OAM and all aspects related to MPLS VPN.**

ALLDATA, DATA ANALYTICS conference series

- **The First International Conference on Big Data, Small Data, Linked Data and Open Data**

- **ALLDATA 2015 | Barcelona**

http://www.iaria.org/conferences2015/ALLDATA15.html

- **ALLDATA 2016 | Lisbon**

http://www.iaria.org/conferences2016/ALLDATA16.html

- **ALLDATA 2017 | Venice**
- April 23-27, 2017
- **DATA ANALYTICS 2017**
- **Nov 12-16, Barcelona**

ALLDATA 2015 | 2016 | 2017

- **BIG DATA**
  Big data foundations; Big data understanding (knowledge discovery, learning, consumer intelligence); Big data semantics, search and mining; Big data processing and transformations; Big data handling, simulation, visualization, modeling tools, and algorithms; Managing big data (large-scale, integration, etc.); Unknown in large Data Graphs; Reasoning on Big data; Big data analytics for prediction; Applications of Big data (health, financial, social, weather forecasting, etc.; Business-driven Big data; Big data and cloud technologies; Technologies handling Big data; High performance computing on Big data; Big data persistence and preservation; Big data protection, integrity and privacy ; Big data toolkits
- **SMALL DATA**
  Social networking small data; Relationship between small data and big data; Statistics on Small data; Handling Small data sets; Predictive modeling methods for Small data sets; Small data sets versus Big Data sets; Small and incomplete data sets; Normality in Small data sets; Confidence intervals of small data sets; Causal discovery from Small data sets; Deep Web and Small data sets; Small datasets for benchmarking and testing; Validation and verification of regression in small data sets; Small data toolkits
- **LINKED DATA**
  RDF and Linked data; Deploying Linked data; Linked data and Big data; Linked data and Small data; Evolving the Web into a global data space via Linked data; Practical semantic Web via Linked data; Structured dynamics and Linked data sets; Quantifying the connectivity of a semantic Linked data; Query languages for Linked data; Access control and security for Linked data; Anomaly detection via Linked data; Semantics for Linked data; Enterprise internal data 'silos' and Linked data; Traditional knowledge base and Linked data; Knowledge management applications and Linked data
- **OPEN DATA**
  Open data structures and algorithms; Designing for Open data; Open data and Linked Open data; Open data government initiatives; Big Open data; Small Open data; Challenges in using Open data (maps, genomes, chemical compounds, medical data and practice, bioscience and biodiversity); Linked open data and Clouds; Private and public Open data; Culture for Open data or Open government data; Data access, analysis and manipulation of Open data; Open addressing and Open data; Specification languages for Open data; Legal aspects for Open data

Q&A

Thanks!