FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

SEMANTiCS Vienna 2018

Taming the logs – Vocabularies for semantic security analysis

Elmar Kiesling
Andreas Ekelhart, Kabul Kurniawan

This work was sponsored by the Austrian Science Fund (FWF) and netidee SCIENCE under grant P30437-N31, as well as the COMET K1 program by the Austrian Research Promotion Agency.

Just last week…

The Guardian

BA customers' credit card details 'probably already for sale'

The credit card details of 380,000 British Airways customers could already be on sale on the internet after the airline suffered a "malicious" data breach, experts have warned.

The online theft saw details stolen including name, email address and credit card information, including the CVV code. BA has said that its encryption was not breached but that the hackers used other "very sophisticated" methods. Cybersecurity experts speculated that the inclusion of CVV

its website and app over a two-week period between 21 August and 5 September. The airline said it would compensate passengers for any losses, signalling the potential for large payouts, given the number of customers

https://www.theguardian.com/business/2018/sep/07/ba-british-airways-customers-hacked-credit-card-details-dark-web

INDEPENDENT

INDY/NEWS

BRITISH AIRWAYS HACKED: SCALE OF CUSTOMER DATA BREACH IS 'ASTOUNDING', SECURITY EXPERTS SAY

https://www.independent.co.uk/life-style/gadgets-and-tech/news/british-airways-hacked-customer-data-breach-astounding-ba-security-experts-a8527071.html

FT FINANCIAL TIMES

British Airways vows to compensate passengers after data breach

Customers express anger over hacking of their personal information

BA disclosed that hackers had stolen data relating to about 380,000 customers from its website © AP

https://www.ft.com/content/5eddd118-b27e-11e8-99ca-68cf89602132

Motivation

Security Challenges:

- Increasingly **sophisticated attacks**
- Indicators of compromise **hidden in high-volume log data**
- **Difficult to connect disparate clues** in scattered, heterogeneous logs

State-of-the-art approaches:

- *SIEM systems*: log aggregation, (often rudimentary) correlation, alerting
- *Intrusion detection/prevention (IDS/IPS) systems*
  usually based on statistical models or signatures

Not particularly good at:

- interpreting potential clues
- putting them into context
- establishing causal links between them
- automatically detecting unseen attacks

Consequences

- Tedious manual analyses
  → Lack of situational awareness
- False positives → "Alert fatigue"
- Long detection and response times
  when incidents do occur

The good news..

Systems generate vast amounts of log data..

splunk>

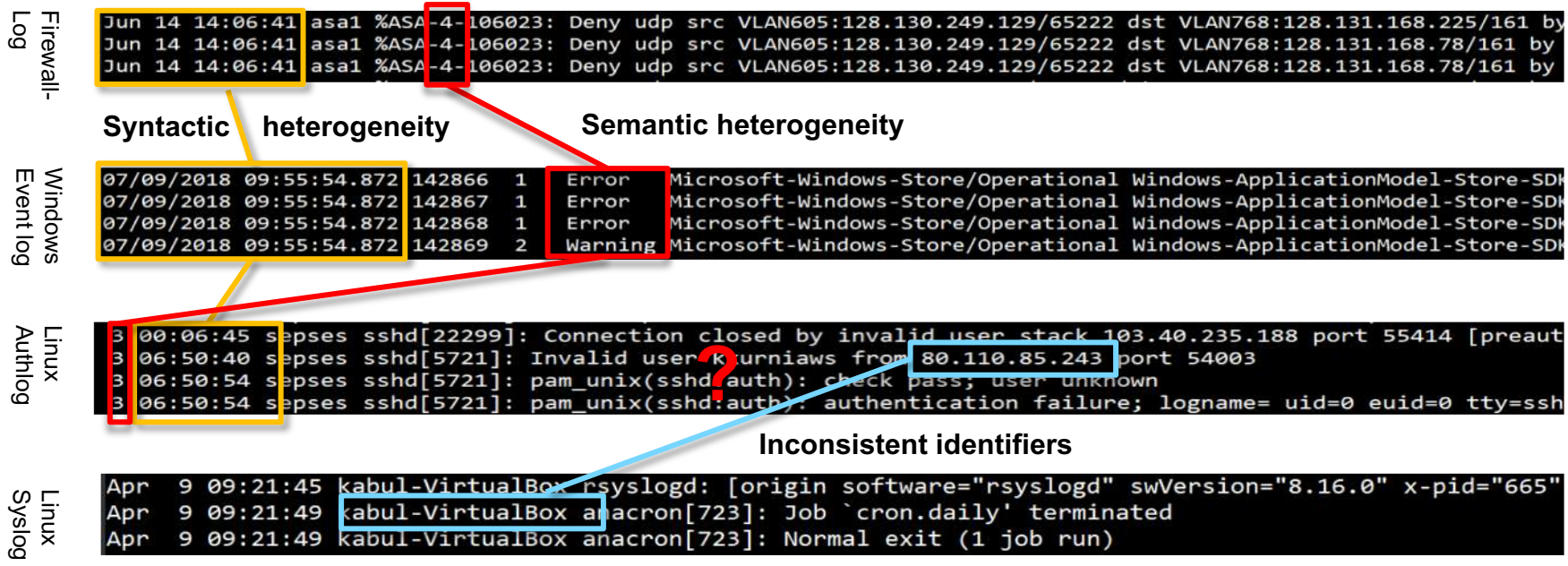
logstash kibana elasticsearch beats

rsyslog

logstash

.. and there are tools to collect it.

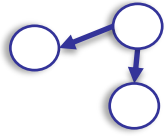
The bad news..



- Logs are weakly structured, use varying formats, and inconsistent terminology
- Interpretation is difficult, requires background knowledge, and is highly context-dependent
- Connecting clues from disparate, heterogeneous logs is hard
- There is just too much of it for human experts to analyze …

Semantics to the rescue..

Characteristics of semantic technologies:



Graph-based

flexible schemas

flexible querying

context-rich representation

Explicit Semantics

terminological clarity

“machine-readability”

integration

reasoning

Decentralization

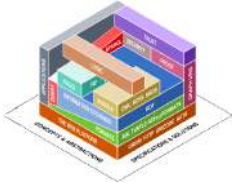
linking

federation

alignment

sharing

reconciliation



Standards-based

SPARQL

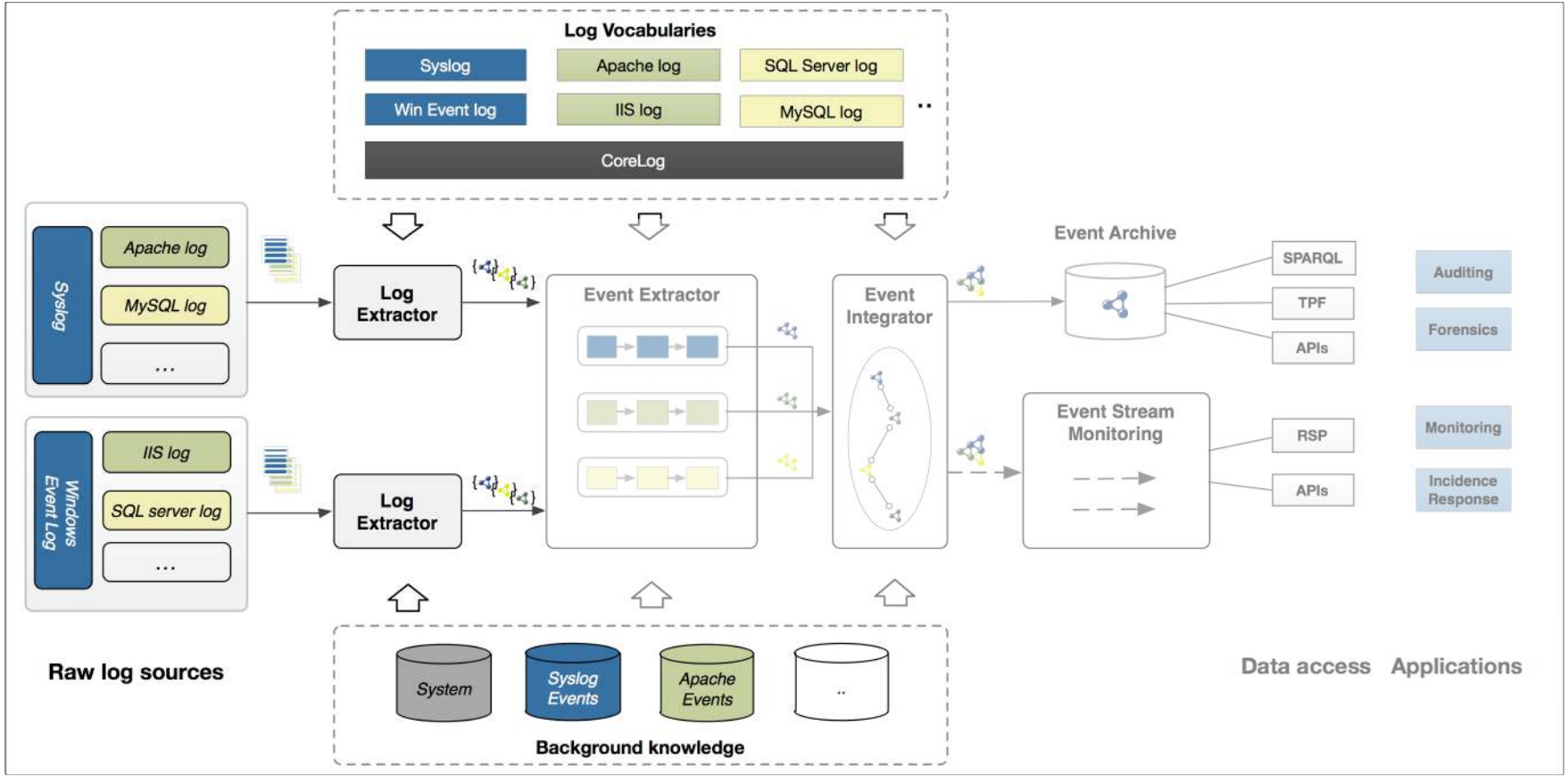
JSON-LD

RDF

Semantic foundation for security analytics

(monitoring, threat intelligence and detection, forensics, incident response etc.)

Semantic security log analysis: Architecture

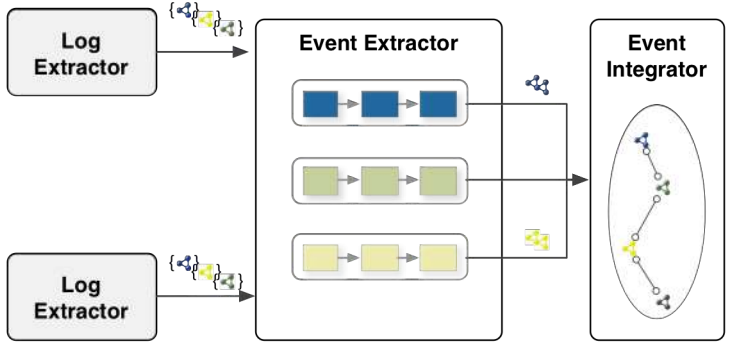


Vocabulary design

- **Log** level:
  - Core vocabulary: basic logging concepts
  - Source-specific extensions
  - Largely literal representation for efficient transformation
  - Use existing vocabularies where directly possible

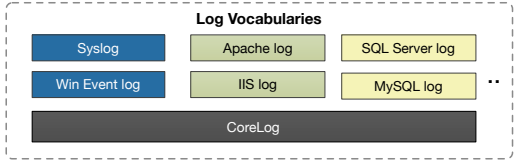
- **Event** level:
  - Enriched and extracted events
  - 1..n log entries → single log event
  - Vocabularies security and process knowledge

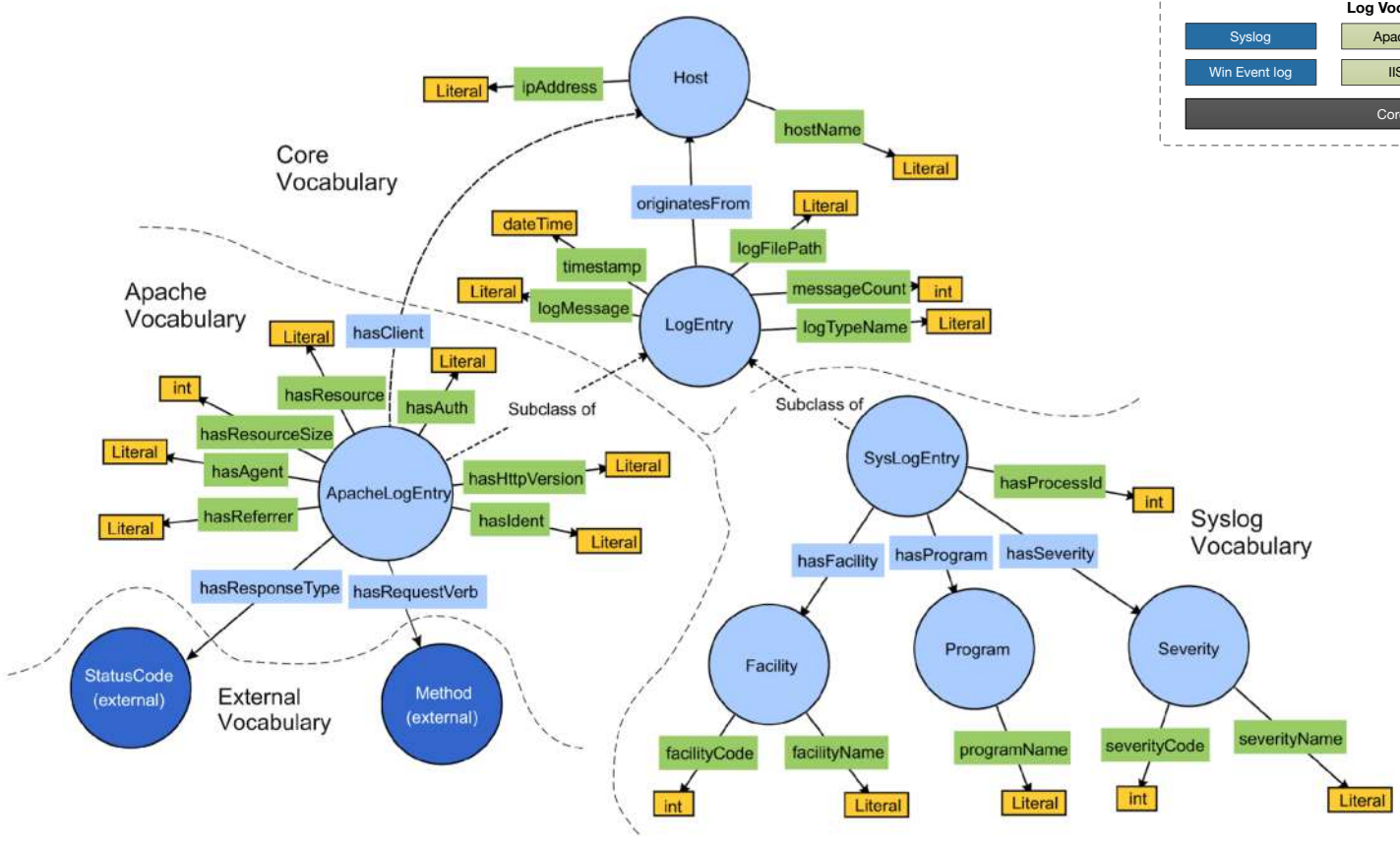
- **High-level event** level
  - events from multiple sources
  - patterns of related events



+ System vocabularies
+ Threat pattern vocabularies
..

Log Vocabularies

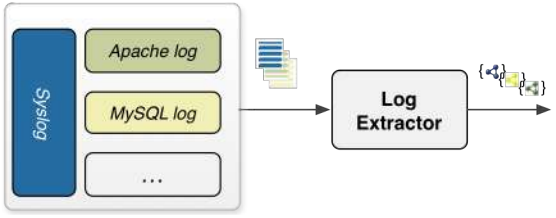




Log Extraction

Design goals:

- Preliminary harmonization, uniform representation
- Initial lifting for many sources in a scalable manner
- Large-scale, high-throughput log processing
- (near) real-time extraction
- Support for stream processing
- “Extension points” for subsequent enrichment, alignment, entity reconciliation and linking with background knowledge



Approach:

- Independent extraction for each log source
- Add semantics to raw JSON stream through JSON-LD `@context` injection
- Add nodes for subsequent linking (e.g., anonymous assets)

Example: Extracted log message (JSON-LD)

```
Apr 9 09:37:47 kabul-VirtualBox systemd[1]: Mounted Huge Pages File System.
```


```
{
        "@context":"http://sepses.ifs.tuwien.ac.at/contexts/syslog.jsonld"
        "logMessage":"Apr 9 09:37:47 kabul-VirtualBox systemd [1]: Mounted Huge Pages File System.",
        "timestamp":"2018-04-09T07:37:47.000Z",
        "hasProcessId":"1",
        "hasSeverity":{
                "severityName":"notice",
        "severityCode":"5"
        },
        "@type":"http://purl.org/sepses/vocab/log/sysLog#SysLogEntry",
        "hasLogType":"http://example.org/system#syslog",
        "@id":"http://example.org/logEntry#logEntry-ca1c3894-114f-432d-befd-abca46258e85",
        "hasProgram":{
                "programName":"systemd"
        },
        "logFilePath":"/var/log/syslog",
        "hasFacility":{
                "facilityCode":"1",
                "facilityName":"user-level"
        },
        "input":{
                "type":"log"
        },
        "originatesFrom":{
                "hostName":"kabul-VirtualBox"
        }
}
```



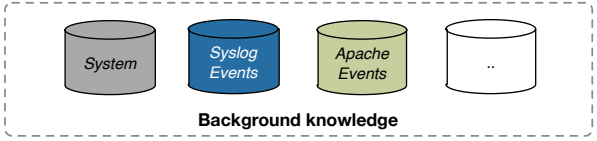
```
{
  "@context": {
              "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
              "xsd": "http://www.w3.org/2001/XMLSchema#",
              "syslog": "http://purl.org/sepses/vocab/log/sysLog#",
              "scl": "http://purl.org/sepses/vocab/log/coreLog#",

                 "logFilePath": {
                        "@id": "scl:logFilePath",
                        "@type": "rdfs:Literal"
                 },
                 "timestamp": {
                        "@id":"scl:timestamp",
                        "@type": "xsd:dateTime"
                 },
                 "logMessage": {
                        "@id":"scl:logMessage",
                        "@type": "rdfs:Literal"
                 },

              …
              }

}
```

Background Knowledge



- **System knowledge:**
  Capture organization-specific concepts, assets, policies..
  e.g., users, network components etc.

- **Event knowledge:**
  Event definitions and associated extraction patterns

- **Threat knowledge:**
  Threat intelligence (indicators of compromise, common attack patterns,..)

Challenges:

- Knowledge evolves dynamically
  e.g., IPs, devices, projects etc.

- Historic context is important!

```
@prefix : <http://example.org/sepses/bg-knowledge#> .
@prefix cl:<http://purl.org/sepses/vocab/log/coreLog#> .
@prefix sys_bg:<http://purl.org/sepses/vocab/bg/system#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .


:bg0a894ed5-746c-4f3e-90e5-efc22cdbcca9 a cl:host,
        cl:hostName "kabul-VirtualBox"^^xsd:String ;
        cl:IPAddress "192.168.0.164"^^xsd:String ;
        sys_bg:hostType "DatabaseServer"^^xsd:String .

:bg93f2287a-07e2-4757-874b-990f37a26db9 a cl:host,
        cl:hostName "linux-Machine"^^xsd:String ;
        cl:IPAddress "192.145.0.124"^^xsd:String ;
        sys_bg:hostType "WebServer"^^xsd:String .

:bg93f2287a-07e2-4757-874b-990f37a26db9 a cl:host,
        cl:hostName "DESKTOP-MAC610T"^^xsd:String ;
        cl:IPAddress "192.164.1.151"^^xsd:String ;
        sys_bg:hostType "FileServer"^^xsd:String .
```

Example system background knowledge

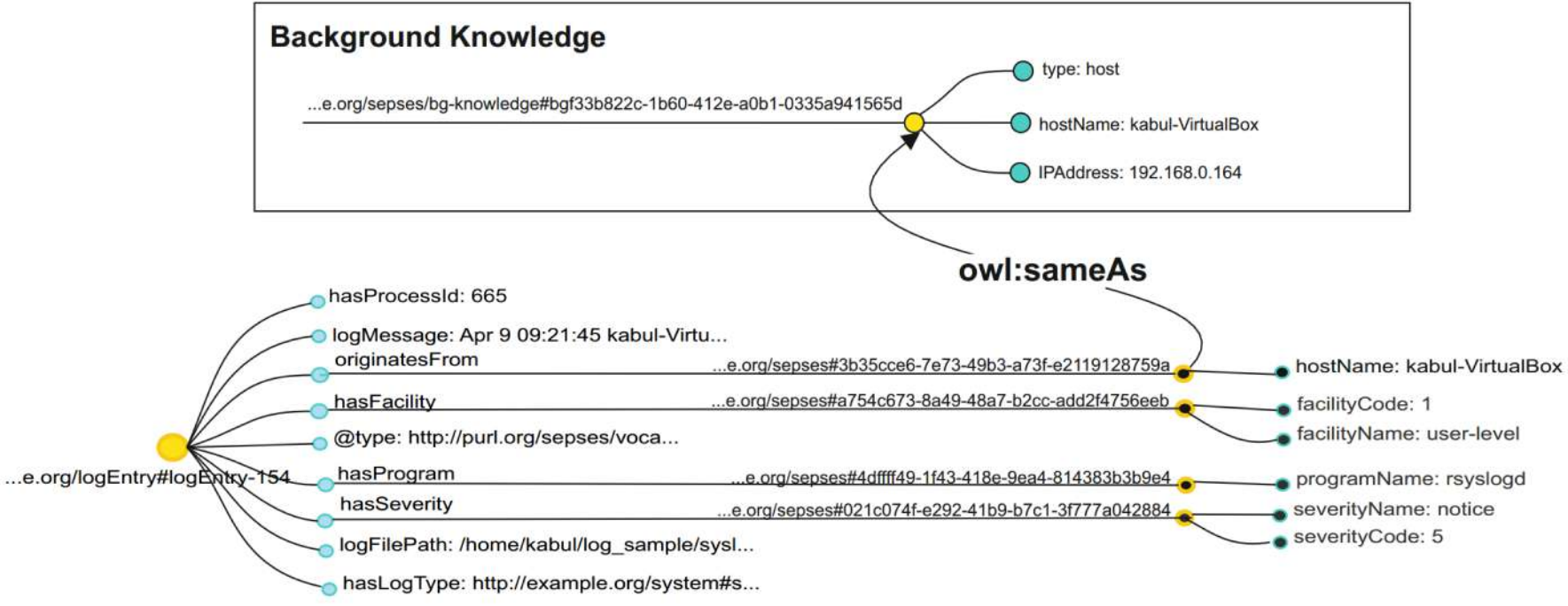
Linking to Background Knowledge

```
<LinkageRule linkType="owl:sameAs">
     <Compare id="levenshteinDistance1" required="false"weight="1" metric="levenshteinDistance"
threshold="0.0" indexing="true">
                <TransformInput id="lowerCase1" function="lowerCase">
                       <Input id="sourcePath1" path="/syslog:hostName"/>
                </TransformInput>
                <TransformInput id="lowerCase2" function="lowerCase">
                        <Input id="targetPath1" path="/bgk:hostName"/>
                </TransformInput>
               <Paramname="minChar" value="0"/><Paramname="maxChar" value="z"/>
     </Compare>
</LinkageRule>
```

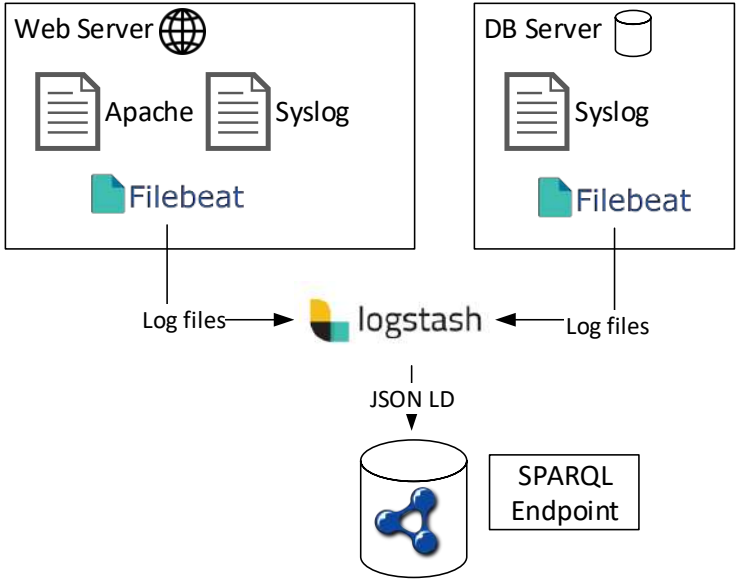
Example linking rule to establish equivalency based on host name

- Declarative Linking Patterns (Silk)
- Linking predicate: `sameAs`
- Challenges:
  - Inconsistent identifiers (e.g., IP, host name) → entity reconciliation
  - Unstable identifiers (e.g., DHCP leases, PIDs) → temporal reasoning (?)
  - Linking target may not exist in BGK → discovery mechanism

Example: Linking to Background Knowledge



Prototype Implementation



- Virtual machine with Apache web server (syslog and Apache log)
- Log acquisition: filebeat
- Custom Logstash configuration to produce JSON-LD log entries

Example Query and Results

[Q1] All log messages in a particular time window

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX scl: <http://purl.org/sepses/vocab/log/coreLog#>
PREFIX sysbg: <http://purl.org/sepses/bg/system#>

SELECT ?time ?logType ?hostName ?ipAddress ?hostType ?message
      WHERE {?logEntry a  scl:LogEntry;
                          scl:originatesFrom ?host;
                          scl:hasLogType ?logType;
                          scl:logMessage ?message;
                          scl:timestamp ?time.
             ?host        sysbg:hostType ?hostType;
                          scl:hostName ?hostName;
                          scl:ipAddress ?ipAddress.
FILTER(?time > "2018-04-09T07:29:00+00:00"^^xsd:dateTime &&
             ?time < "2018-04-09T07:34:00+00:00"^^xsd:dateTime)}
```

time (xsd:dateTime)	logType	hostName	ipAddress	hostType	message
2018-04-09T07:29:15+00:00	scl:syslog	kabul-VirtualBox	192.168.0.164	DatabaseServer	”org.debian.apt[683]: . . .”
2018-04-09T07:31:45+00:00	scl:apache	linux-Machine	192.145.0.124	WebServer	*”GET /presentations/ ”*
2018-04-09T07:31:45+00:00	scl:apache	linux-Machine	192.145.0.124	WebServer	*”GET /presentations/ . . . ”*
2018-04-09T07:31:45+00:00	scl:syslog	kabul-VirtualBox	192.168.0.164	DatabaseServer	*”systemd-tmpfiles[3572]: . . . ”*
2018-04-09T07:31:45+00:00	scl:syslog	kabul-VirtualBox	192.168.0.164	DatabaseServer	*”systemd[1]: Started . . . ”*

Example: reconciliation with external threat intelligence

CVE: Common Vulnerabilities and Exposures

- Semi-structured public vulnerabilities list published by MITRE
- XML-based
- Attributes:
  - CVE-ID
  - Date, Description, References, …

Common Platform Enumeration

- Vulnerable software and versions
- CVE-PDE mapping available

```
cpe:2.3:a:notepad_plus_plus:notepad\+\+:7.3.3:*:*:*:*:*:*:*
```

Query example:

- Match CVEs to affected assets based on background knowledge
- Link to relevant log stream
- ..

Example Query with background knowledge

[Q2] Vulnerability information from CVE linked to affected server

```
PREFIX cve_ex: <http://example.org/sepses/cve#>
PREFIX win: <http://purl.org/sepses/vocab/log/winEventLog#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX cve: <http://purl.org/sepses/vocab/cve#>

SELECT  ?host ?programName ?programVersion ?CVE ?CVEDesc
            WHERE {   ?logEntry a   win:WindowsHighLevelEvent ;
                                    win:sourceHost ?host ;
                                    win:programName ?programName ;
                                    cve:hasVulnerability ?CVE .
                        ?CVE        cve:CVEDesc ?CVEDesc ;
                                    cve:ProgramVersion ?programVersion
}
```

CVE-2017-8803 Detail

Current Description

Notepad++ 7.3.3 (32-bit) with Hex Editor Plugin v0.9.5 might allow user-assisted attackers to execute code via a crafted file, because of a "Data from Faulting Address controls Code Flow" issue. One threat model is a victim who obtains an untrusted crafted file from a remote location and issues several user-defined commands.

Source: MITRE
Description Last Modified: 07/05/2017
+View Analysis Description

host	programName	programVersion	CVE	CVEDesc
"DESKTOP-MAC610T"^^	"notepad++.exe"^^	"7.3.3 (32-bit)"^^	cve_ex:CVE-2017-8803	"might allow user-assisted attackers to execute code via a crafted file"^^
"DESKTOP-MAC610T"^^	"notepad++.exe"^^	"6.6.9"^^	cve_ex:CVE-2014-9456	"Buffer overflow in NotePad++ 6.6.9 allows remote attackers to have unspecified impact via a long Time attribute in an Event element in an XML file"^^

Conclusions

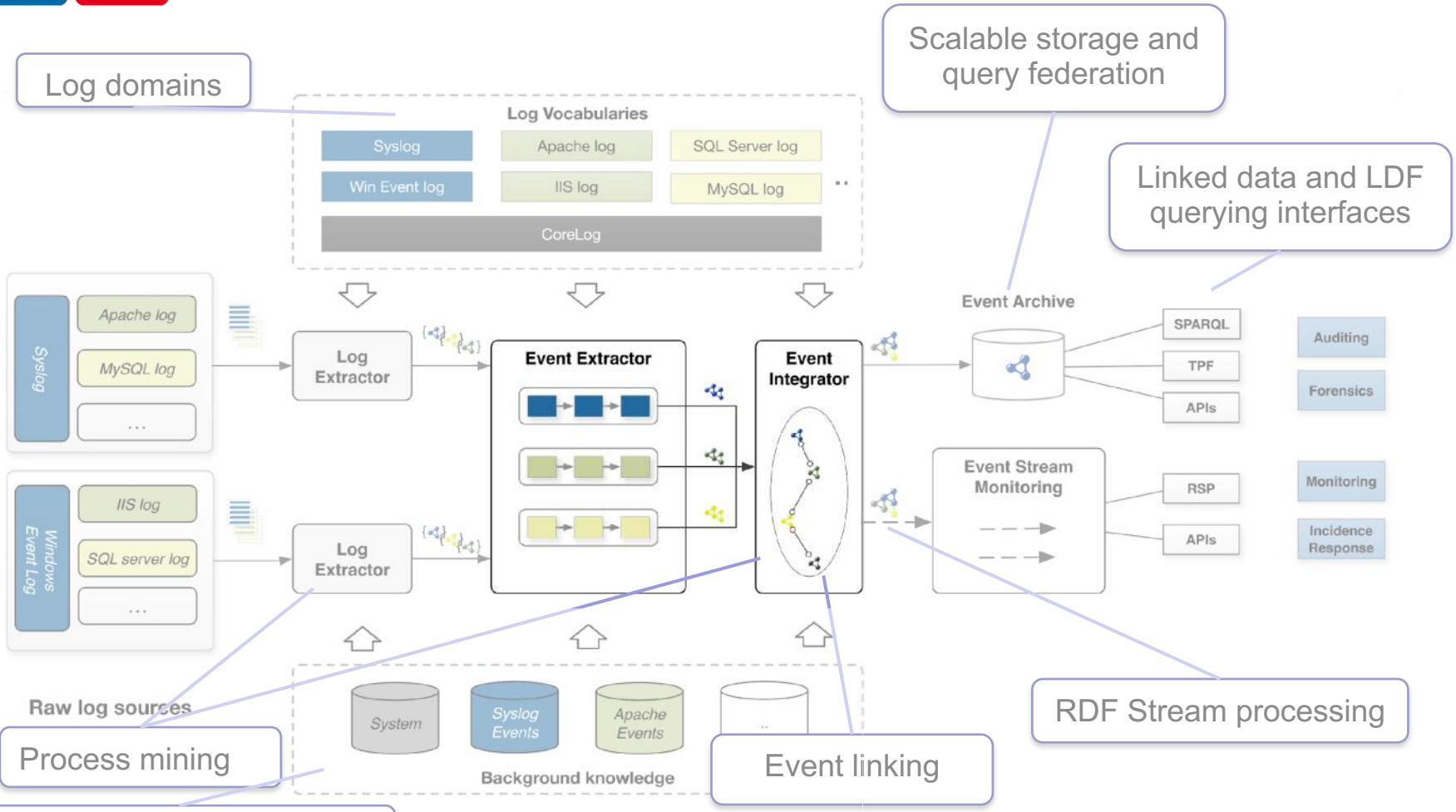
Semantic security logs create a foundation for tools that could..

- improve situational awareness
- facilitate better security decision-making
- create better understanding
- enable faster forensics
- make more effective incident response and containment possible

What we have accomplished so far..

- Architecture
- Domain-specific Vocabularies
- Log acquisition and knowledge extraction
- Alignment and entity reconciliation
- Linking to background knowledge
- Prototype

Current/Future Work



?

elmar.kiesling@tuwien.ac.at