Ditch the Template

How to Write Incident Reports They Want To Read

Laura Nolan @lauralifts

stanza

Laura Nolan

- Former Google and Slack SRE
- Currently Principal SWE at Stanza Systems
- Contributor to O'Reilly *Site Reliability Engineering* book and various other books on software operations and reliability, plus a lot of articles and talks
- Member of USENIX Association Board of Directors
- SREcon Steering Committee
- Avid reader (and writer) of incident reviews

slack engineering

Search

Slack's Incident on 2-22-22

Double Trouble with Datastores

Laura Nolan Senior Staff Engineer

O'REILLY
O'REILLY
Site Reliability Engineering
ated and edited by
N. Blank-Edelman
usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION
O'REILLY
Compliments of
NGINX
97 Things Every SRE Should Know
Collective Wisdom from the Experts
Edited by Emil Stolarsky & Jaime Woo

Elegant software for production management.

Reliability, Scalability, Velocity: pick three.

Enter your email

GET UPDATES

TL;DR

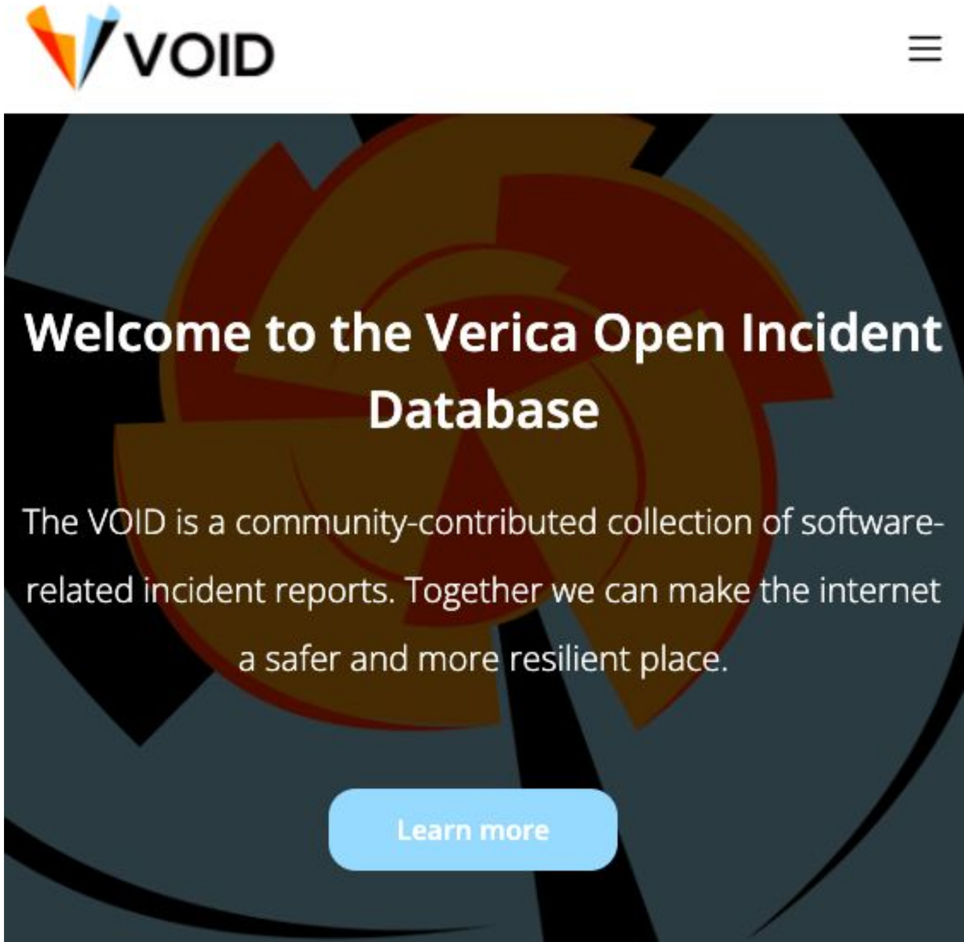
- Engaging incident reports are valuable
- Focus on narrative, not metadata
- Support your readers
- Be visual
- Don’t be afraid of analysis
- Pay attention to style

The value of written incident reports

- Share knowledge and context, including between teams
- Helps your organization understand and adapt if needs be
- Encourages thoughtful reflection
- Long-term store of knowledge
- Onboarding

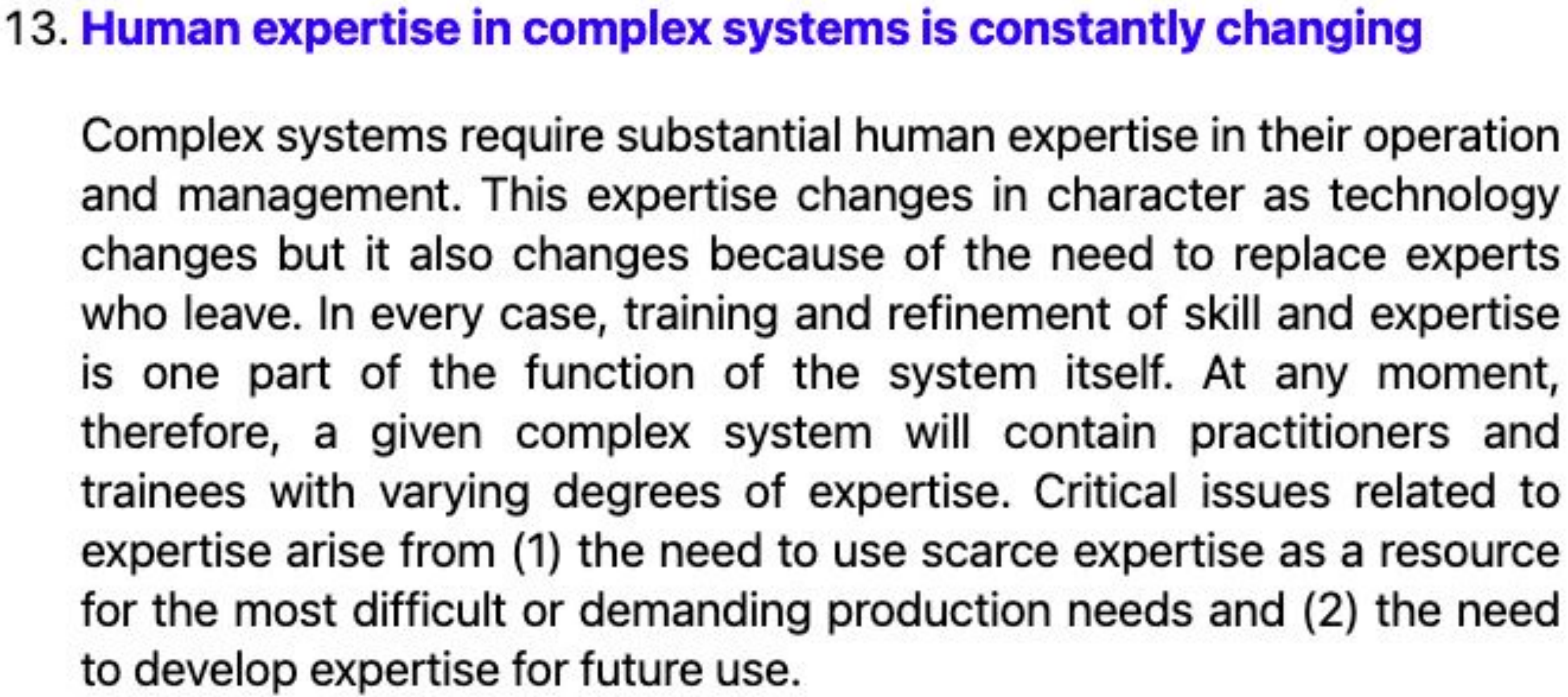
The value of written incident reports

- Share knowledge, help to lift the entire industry
- Particularly valuable for OSS and SaaS sharp edges
- Can feed into research, product improvements
- Your customers will appreciate transparency



Welcome to the Verica Open Incident Database

The VOID is a community-contributed collection of software-related incident reports. Together we can make the internet a safer and more resilient place.



13. **Human expertise in complex systems is constantly changing**

Complex systems require substantial human expertise in their operation and management. This expertise changes in character as technology changes but it also changes because of the need to replace experts who leave. In every case, training and refinement of skill and expertise is one part of the function of the system itself. At any moment, therefore, a given complex system will contain practitioners and trainees with varying degrees of expertise. Critical issues related to expertise arise from (1) the need to use scarce expertise as a resource for the most difficult or demanding production needs and (2) the need to develop expertise for future use.

– Richard I. Cook MD, 'How Complex Systems Fail'
https://how.complexsystems.fail/

The value of incident reviews is in the **LEARNING**, not in the process

Title (incident #)

Date

Authors

Status

Summary

https://github.com/dastergon/postmortem-templates/blob/master/templates/postmortem-template-srebook.md

Impact

Root Causes

Trigger

Resolution

Detection

Action Items

Lessons Learned

What went well

What went wrong

Where we got lucky

Timeline

Supporting information

Title (incident #)

Date

Authors

Status

Summary

Impact

Root Causes

Trigger

Resolution

Detection

Action Items

Lessons Learned

What went well

What went wrong

Where we got lucky

Timeline

Supporting information

Jurassic Park as an IR

- **Incident Number:** JP-001
- **Impact:** 5 human deaths, facilities on island a total loss, significant reputational damage.
- **Summary:** Dinosaurs escaped from enclosures. Some deaths resulted.
- **Resolution:** Costa Rican military bombed the island, killing all dinosaurs.
- **Trigger:** Operator disabled electric fences.
- **Root Cause:** Internal processes failed to detect corrupt employee who engaged in bribery.
- **How We Got Lucky:** Several staff and visitors escaped via helicopter.

A STEVEN SPIELBERG FILM
JURASSIC PARK
UNIVERSAL PICTURES PRESENTS AN AMBLIN ENTERTAINMENT PRODUCTION SAM NEILL LAURA DERN JEFF GOLDBLUM AND RICHARD ATTENBOROUGH "JURASSIC PARK"
LIVE ACTION DINOSAURS BY STAN WINSTON VISUAL EFFECTS SUPERVISOR DENNIS MUREN DINOSAUR SUPERVISOR PHIL TIPPETT SPECIAL DINOSAUR EFFECTS MICHAEL LANTIERI MUSIC BY JOHN WILLIAMS FILM EDITED BY MICHAEL KAHN, A.C.E.
ORIGINAL SOUNDTRACK ON MCA CDs AND CASSETTES PRODUCTION DESIGNER RICK CARTER DIRECTOR OF PHOTOGRAPHY DEAN CUNDEY, A.S.C. BASED ON THE NOVEL BY MICHAEL CRICHTON SCREENPLAY BY DAVID KOEPP AND MICHAEL CRICHTON SPECIAL VISUAL EFFECTS BY INDUSTRIAL LIGHT & MAGIC
AMBLIN ENTERTAINMENT PRODUCED BY KATHLEEN KENNEDY AND GERALD R. MOLEN COMING SOON DIRECTED BY STEVEN SPIELBERG A UNIVERSAL PICTURE
UNIVERSAL AN MCA COMPANY

Firefox Outage, January 13 2022

IR by Christian Holler, https://mzl.la/3PUPbMw

A GCP config change enabling HTTP/3 triggered a bug in Firefox Rust HTTP/3 handling code that caused an infinite loop, blocking all network access.

- *“we quickly discovered that the client was hanging inside a network request to one of the Firefox internal services. However, at this point we neither had an explanation for why this would trigger just now, nor what the scope of the problem was.”*
- *“Although we couldn’t see it, we suspected that there had been some kind of “invisible” change rolled out by one of our cloud providers that somehow modified load balancer behavior.”*

GitLab: The Consul Outage That Never Happened

IR by Devin Sylva, https://bit.ly/gl-consul-ir

GitLab discovered that the self-signed certs that their Consul cluster was using to communicate with itself had expired and couldn’t be replaced because the CA key had been lost.

Effectively, we were in the middle of an outage that had already started, but hadn't yet gotten to the point of taking down the site.

We all held our breath, and watched the database for signs of distress. Six minutes is a long time to think: "It's 4am in Europe, so they won't notice" and "It's dinner time on the US west coast, maybe they won't notice".

Supporting the Reader

- Write your IR for any engineer to be able to read
- Explain jargon and system names
- Explain ***why*** things are how they are
- Weave concise explanations into the narrative as needed
- Link out to more detailed documentation where appropriate

Sentry: Transaction ID Wraparound in Postgres

By David Cramer, https://bit.ly/sentry-pg-ir

The first time a transaction manipulates rows in the database (typically via an INSERT, UPDATE, or DELETE statement) the database's XID counter is incremented. This counter is used to determine row visibility. The best human-readable description we've found is Heroku's topic on the MVCC and concurrency.

Eventually, that database is going to reach a situation where the XID counter has reached its maximum value. [...] Since there are no longer any unique XID values available to identify new transactions, the database must halt to prevent XID re-use — otherwise, transactions that occurred in the past could suddenly appear to be in the future, leading to all kinds of strange and undesirable behavior.

GoCardless: Outage 25 October 2020

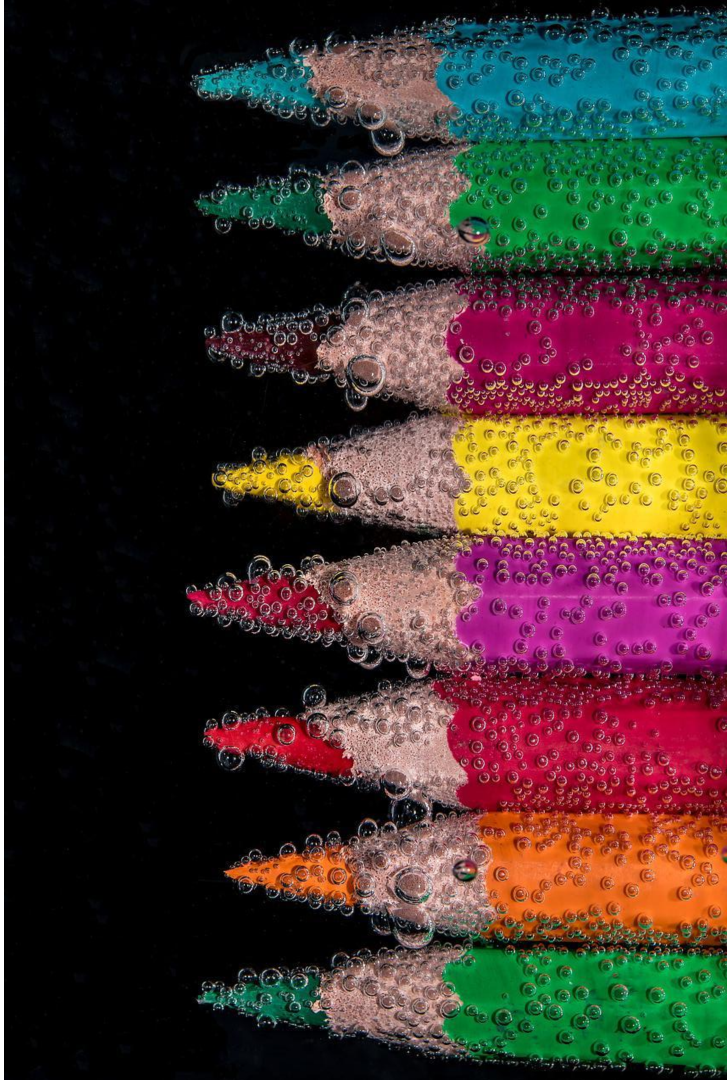
By Ben Wheatley, https://bit.ly/gocardless-ir

At GoCardless we aim to keep velocity high by ensuring that this secret data can be managed by the engineers building these services, rather than relying on an operational team to make changes. To this end, we've implemented a system based on Hashicorp's Vault product.

We run an instance of Vault that all engineers can authenticate with, using their Google identity, and have the permissions to write secrets into, but not read the secret data out of. We additionally have policies configured in Vault that ensure our applications, authenticated via their Kubernetes service account token, can read secret data belonging to that application only.

Be Visual

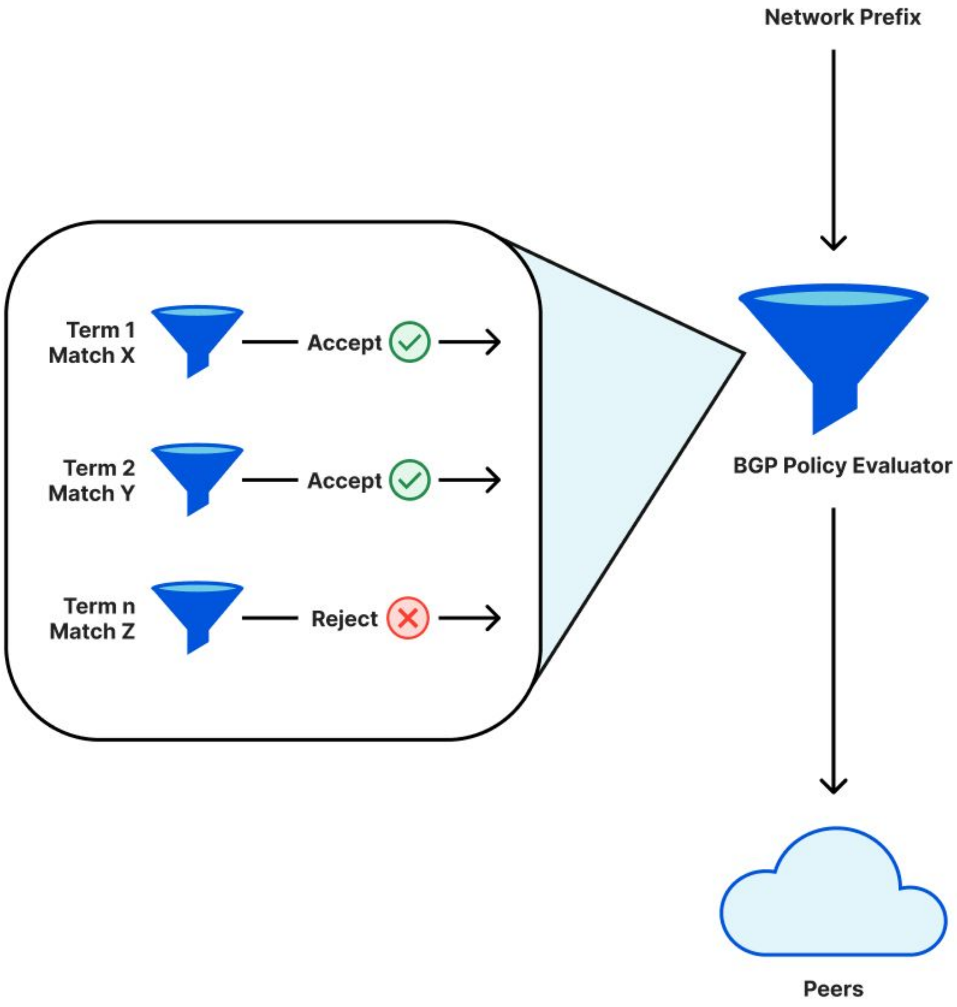
- A picture is worth a thousand words
- They reinforce the text and help people understand
- Visuals add texture to a long report and help to keep reader interest
- Don’t limit yourself to architecture diagrams: graphs, sequence diagrams, timelines, screenshots, topologies - are all great - be creative!



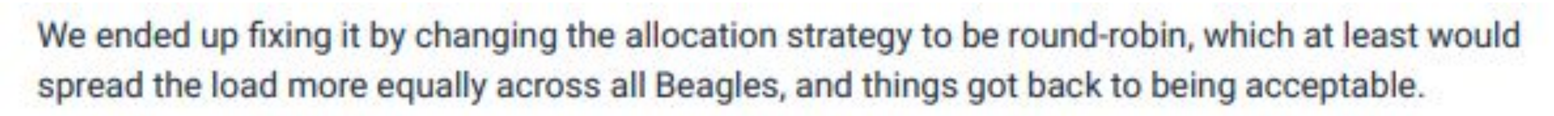
Cloudflare Outage on June 21 2022

By Tom Strickx and Jeremy Hartman

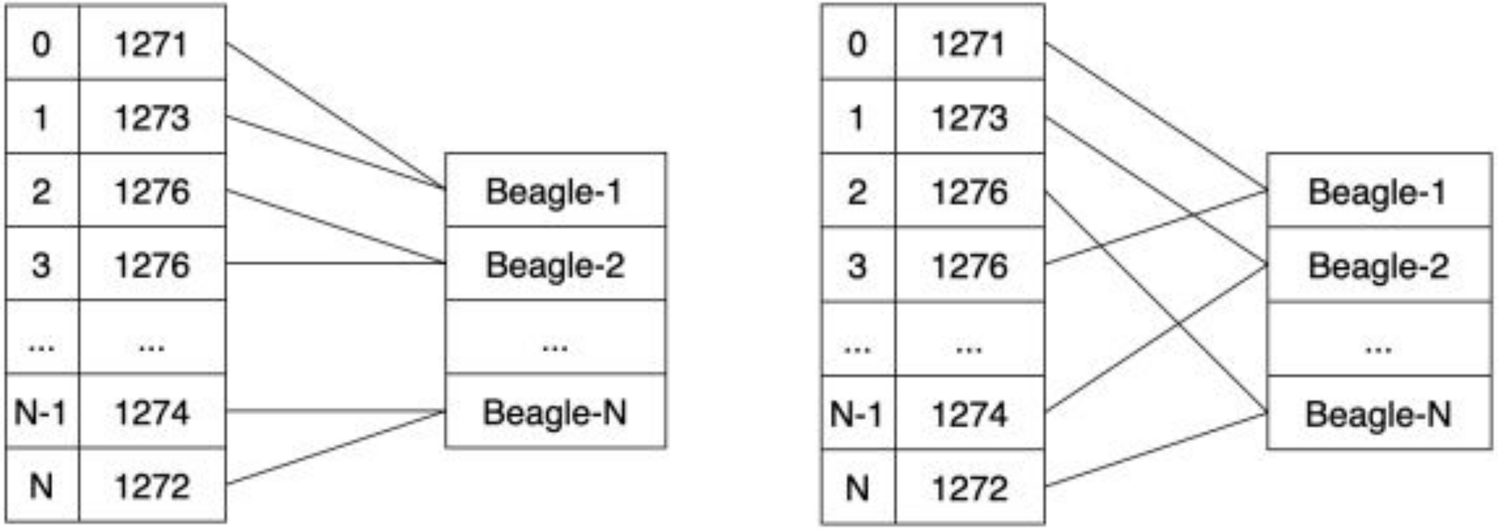
https://bit.ly/cf-bgp-ir



Honeycomb, The 20 Fires of September



We ended up fixing it by changing the allocation strategy to be round-robin, which at least would spread the load more equally across all Beagles, and things got back to being acceptable.



0	1271
1	1273
2	1276
3	1276
...	...
N-1	1274
N	1272

Beagle-1
Beagle-2
...
Beagle-N

0	1271
1	1273
2	1276
3	1276
...	...
N-1	1274
N	1272

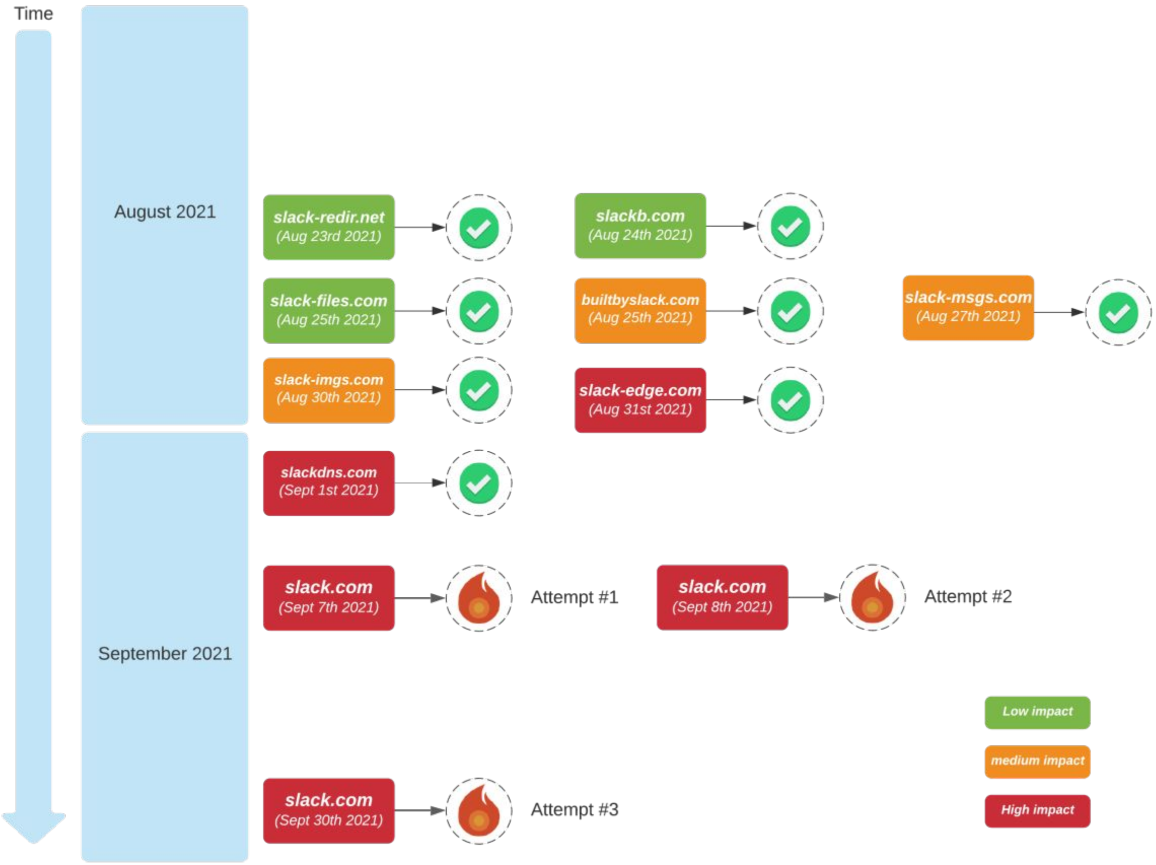
Beagle-1
Beagle-2
...
Beagle-N

https://bit.ly/honeycomb-fires

Slack: The Case of the Recursive Resolvers

By Rafael Elvira

https://bit.ly/slack-dnssec-ir



Time

August 2021

slack-redir.net (Aug 23rd 2021)

slackb.com (Aug 24th 2021)

slack-files.com (Aug 25th 2021)

builtbyslack.com (Aug 25th 2021)

slack-msgs.com (Aug 27th 2021)

slack-imgs.com (Aug 30th 2021)

slack-edge.com (Aug 31st 2021)

September 2021

slackdns.com (Sept 1st 2021)

slack.com (Sept 7th 2021) Attempt #1

slack.com (Sept 8th 2021) Attempt #2

slack.com (Sept 30th 2021) Attempt #3

Low impact

medium impact

High impact

Analysis

- If an IR is a story, the analysis is the moral of the story
- Sharing analysis and lessons learned is the most satisfying way to wrap up an IR
- It creates a feeling of resolution

GitLab: The Consul Outage That Never Happened

IR by Devin Sylva, https://bit.ly/gl-consul-ir

Every once in a while we get into a situation that all of the fancy management tools just can't fix.

In these types of situations there is no shortcut around thinking things through methodically. In this case, there were no tools or technologies that could solve the problem. Even in this new world of infrastructure as code, site reliability engineering, and cloud automation, there is still room for old fashioned system administrator tricks. There is just no substitute for understanding how everything works.

Honeycomb, The 20 Fires of September

If we operate too far from the edge, we lose sight of it, stop knowing where it is, and can’t anticipate when corrective work should be emphasized. But if we operate too close to it, then we are constantly stuck in high-stake risky situations and firefighting.

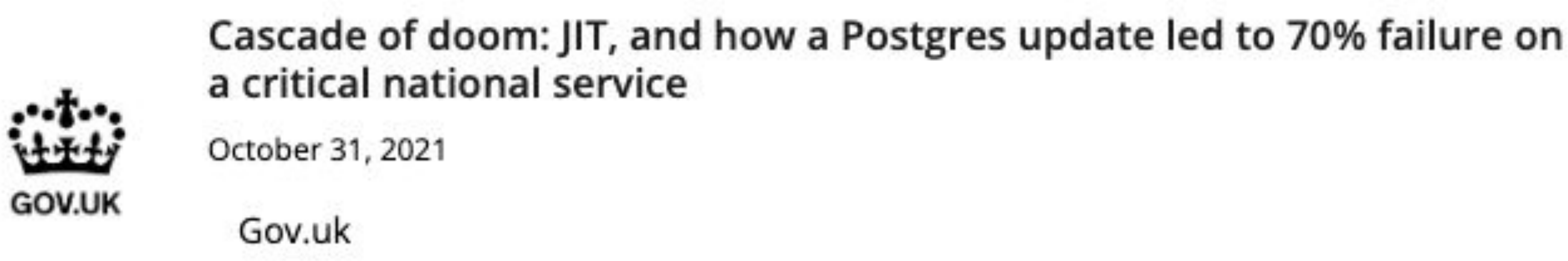
This gets exhausting and we lose the capacity, both in terms of time and cognitive space, to be able study, tweak, and adjust behavior of the system. This points towards a dynamic, tricky balance to strike between being too close to the boundary and too far from it, seeking some sort of Goldilocks operational zone.

https://bit.ly/honeycomb-fires

Craft

Titles

A title helps your readers remember your IR so they can refer to it later



Cascade of doom: JIT, and how a Postgres update led to 70% failure on a critical national service

October 31, 2021

Gov.uk

Use simple and clear language

“It turns out that the TLS certificate was expired. This is normally a simple fix. Someone would go to the Certificate Authority (CA) and request a renewal – or if that fails, generate a new certificate to be signed by the same CA. That certificate would replace the expired copy and the service would be restarted. All of the connections should reestablish using the new certificate and just like with any other rolling configuration change, it should be transparent to all users.

After looking everywhere, and asking everyone on the team, we got the definitive answer that the CA key we created a year ago for this self-signed certificate had been lost.”

https://bit.ly/gl-consul-ir

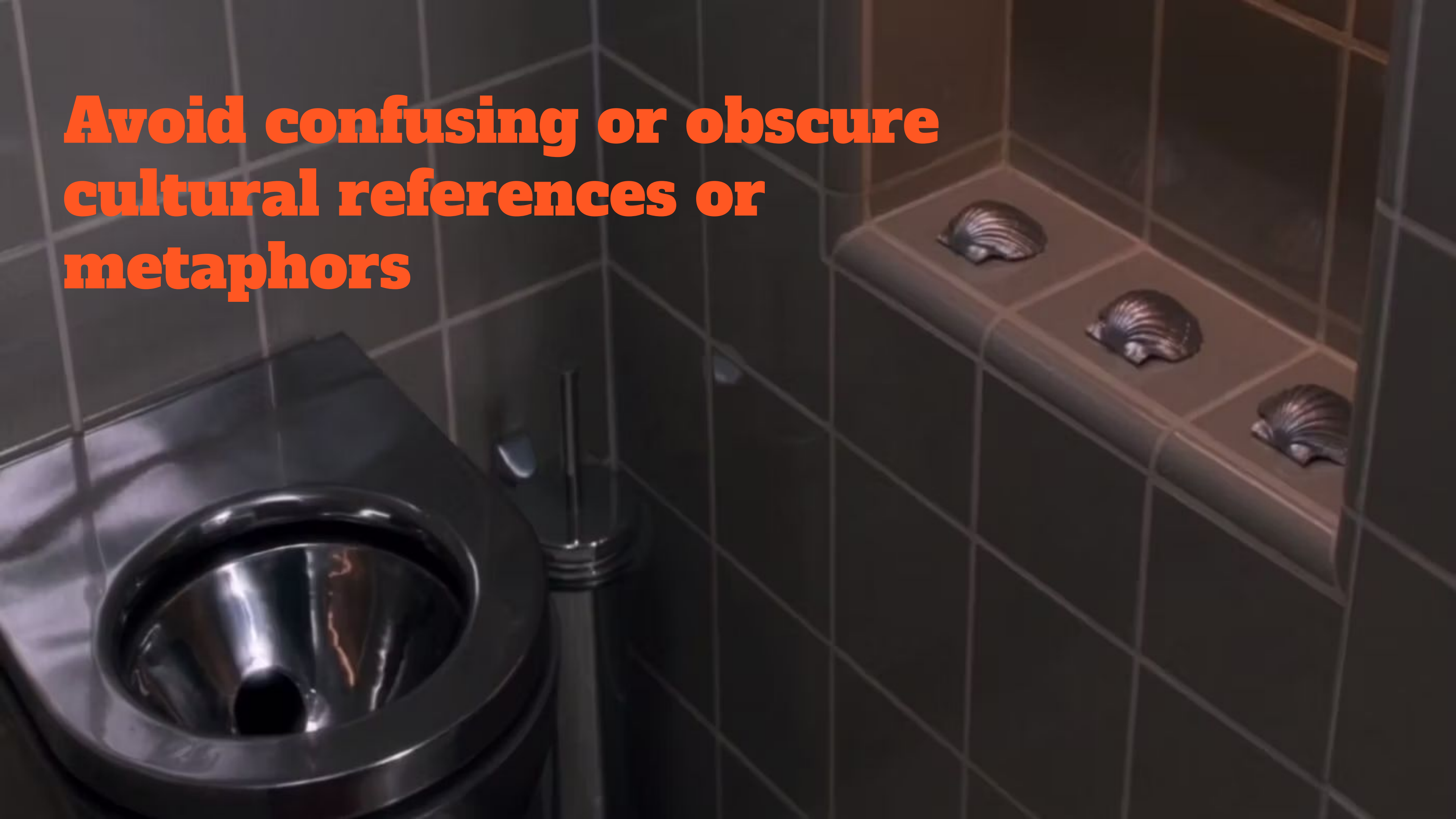
Don’t be too formal (or too informal)

Avoid Walls of Text: Use Headings

Use sentence rhythm

Use a consistent tense

Avoid confusing or obscure cultural references or metaphors



Your IR is not a sales pitch

Looking for nutritious choices? Reach for the Stars!
Looking for nutritious choices? Reach for the Stars!

TL;DR

- Written incident reports are how we collectively learn
- Telling stories is a very effective way to communicate information
- Try to write reports that don’t assume specialised knowledge: give concise explanations and signpost readers to more detailed information
- Use graphs, diagrams, screenshots, timelines, and other sorts of visuals liberally
- Share your analysis and takeaways
- Pay attention to style
- Don’t treat your IR as a sales pitch - be authentic, humble, and honest

FIN

Questions? @lauralifts