SRE: Toil Management Strategy

Toil management is the practice of taking toil and systematically converting that work from operations work to SRE work. It is a formal method of taking the mundane and making it interesting and impactful. This helps ensure the operations work is capped at 50% and remedies any issues that would lead to that ceiling being breached.

Maintaining a healthy sustainable SRE team is crucial as the google SRE books says, “Toil becomes toxic when experienced in large quantities.” The google SRE book mentions some effects like career stagnation, low morale, confusion, slow progress, attrition just to name a few.

It is a simple strategy: continuous improvement process iterating on identifying the highest ROI opportunities, acting on them, and recording them.

Identification: *“Toil is the kind of work tied to running a production service that tends to be manual, repetitive, automatable, tactical, devoid of enduring value, and that scales linearly as a service grows” - Google SRE Book.* Based off that definition there is more than enough toil to go around. The highest value opportunities need to be identified in quantifiable and qualitative ways to ensure we are focusing on the right work. First dividing toil into 2 types: alert/incident toil (come from alerts) and support toil (come from people) give a more holistic perspective. By aggregating incidents, alert data chat data, we can begin to understand the pain points and identify where the toil is coming from. However, the problems we cannot see, we must identify them by talking to customers, developers, and engineers. Performing both activities allow for the ability to answer: What do developers need help with? What do engineers hate doing? What do customers want?...

Act: Investigate the feasibility and risk around performing an activity and weigh it against the benefits. Sift the activities into one of the following 4 buckets: automate, RCA, monitoring, or process improvements. If there is an issue either automate it away, fix it, change the monitoring sensitivity to remove the noise, or change the way we do things, so it does not happen again. Sometimes this means creating a chatbot solution to help users get answers to their questions, creating selfheal and self-service automation for standard issues, using anomaly detection and predictive analytics to remove alert noise, changing the way servers are provisioned to prevent some issue, anticipating an issue, or just fixing a bug.

Record: Track the value of removing the toil and incidents prevented. Keep a hot sheet tracking active issues and a ledger of completed ones. Finally translate the ledger from increased efficiency and impact prevented to saved engineer minutes and reduced costs demonstrating greater customer/developer/engineer satisfaction and better cost efficiency.