Analytics-Konto-Löschung droht

Personenbezogene Daten („PII“) entdecken und vermeiden

METRIKA
WEB ANALYTICS | BERATUNG | SEMINARE

Eine große Anzahl Inhaber von Google-Analytics-Konten bewegen sich in der latenten Gefahr, morgen keine Webanalyse-Daten mehr zu haben. Doch die wenigsten wissen es. Wenn du mithilfe der Software personenbezogene Daten, sogenannte PII (Personally Identifiable Information), übermittelst, kann das Google-Team dein Konto sperren oder mindestens die Daten löschen.

In der täglichen Praxis erlebe ich es immer wieder: „Nein, bei uns werden in der Webanalyse keine personenbezogenen Daten erhoben. Wir haben doch die IP-Adresse für Google Analytics anonymisiert.“

Teils ist es Unwissen, teils wird so etwas sogar wider besseres Wissen gesagt. Doch welche Gefahr seitens Google damit einhergeht, sehen viele Unternehmen nicht.

Doch mal im Ernst: Wie schnell sind personenbezogene Daten an Google Analytics übermittelt? Es muss ja nur jemand auf die Idee kommen, an eine URL einen Parameter wie etwa „www.domain.de?e-mail=MEIN_NAME@MEINE_DOMAIN“ zu hängen oder eine URL mit dem Namen des Kontoinhabers zu besuchen. Und schon sind wir mit unseren Daten in der Bredouille.

Doch von vorne …

Was sind überhaupt personenbezogene Daten in Analytics?

In Deutschland haben wir einen wichtigen Teil der personenbezogenen Daten schon eliminiert, indem alle, die mit Google Analytics tracken, das letzte Oktett der IP-Adresse anonymisieren müssen.

Aus einer echten IP wie 217.158.24.201 wird so in Google Analytics 217.158.24.0 und was dazu führt, dass Nutzer nicht mehr eindeutig anhand dieser Adresse identifiziert werden können. So viel zum Datenschutz gegenüber den Datenschutzbehörden.

Doch es gibt natürlich wesentlich mehr personenbezogene Daten, die an wesentlich mehr Punkten in der Webanalyse auftreten können.

Unter anderem betrifft das folgende Daten:

- URL-Bestandteile (z. B. www.domain.de/mein-konto/VORNAME-NACHNAME/)
- Ereigniswerte (Kategorie, Aktion, Label)
- Kampagnenparameter (die UTM-Parameter „source“, „medium“, „term“, „campaign“, „content“ zur Kennzeichnung von externem Kampagnen-Traffic)
- Site-Search-Suchbegriffe und -Kategorie
- Benutzerdefinierte Dimensionen
- User-ID

Es gibt leider in den „Google Analytics Bedingungen“ keine exakte Definition dessen, was Google als PII ansieht. Doch sie definieren dennoch die Spielregeln und sind hier zu finden: https://www.google.com/analytics/terms/de.html

Dort liest du unter Punkt 1 „Definitionen“:

„‚Personenbezogene Daten‘ bezeichnet jede Information, die sich auf die persönlichen oder sachlichen Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (die betroffene Person) bezieht.“

Damit dürften im Kern E-Mail- und Wohnadressen, persönliche Mobiltelefonnummern, Kreditkartennummern, Sozialversicherungsnummer usw. gemeint sein. Im „schlimmsten“ Fall können vermutlich sogar Postleitzahlen als personenbezogene Daten gewertet werden, wenn bspw. ein Ein-Mann-Unternehmen eine eigene Postleitzahl hat.

Es gibt also tatsächlich einen Grund, sich die Googleschen Spielregeln genauer anzuschauen. Ernsthaft.

Google definiert unter 7.1 noch etwas genauer:
„Sie werden keine Informationen an Google übermitteln, anhand derer Google eine Person identifizieren könnte, noch werden Sie eine solche Übermittlung Dritten erlauben oder diese dabei unterstützen.“

Damit ist auch klar, dass du nicht unbedingt „selbst schuld“ sein musst. Auch wenn ein Drittanbieter (z. B. ein Tracking-Plugin im Shopsystem oder in Wordpress) entsprechende Daten an Google Analytics übermittelt, hat Google alle Möglichkeiten in der Hand.

Was passiert, falls PII erhoben werden?

Doch was passiert dann?

Wenn du gegen die Regeln verstößt behält sich Google das Recht vor, dein Google-Analytics-Konto zu löschen bzw. die mit persönlichen Daten angereicherten Daten darin. Und zwar dort, wo sie aufgelaufen sind: auf Property-Ebene. Also unwiderruflich verloren für die Datenansichten.

Und: Es gibt kein „Ich habe das nicht gewusst“ oder „das ist nicht meine Schuld“. Fakt ist: Wenn personenbezogene Daten erstmal eingegangen sind, kann das zu Problemen führen.

So können PII aufgedeckt werden

Viele wissen nicht, dass Personally Identifiable Information bei ihnen in den Daten aufgelaufen sind und fallen oftmals aus allen Wolken, wenn sie es bemerken.

Doch du kannst auch proaktiv etwas unternehmen. Und es ist meist gar nicht so schwer, entsprechende Daten zu entdecken, wie man auf den ersten Blick meint.

Nehmen wir als Beispiel E-Mail-Adressen, die sich in Daten geschlichen haben. Eines der Hauptmerkmale einer E-Mail-Adresse ist das „@“-Zeichen. Und nach diesem kann jeder Kontoinhaber schon mal problemlos mittels Filterfeld suchen (siehe Abb. 1).

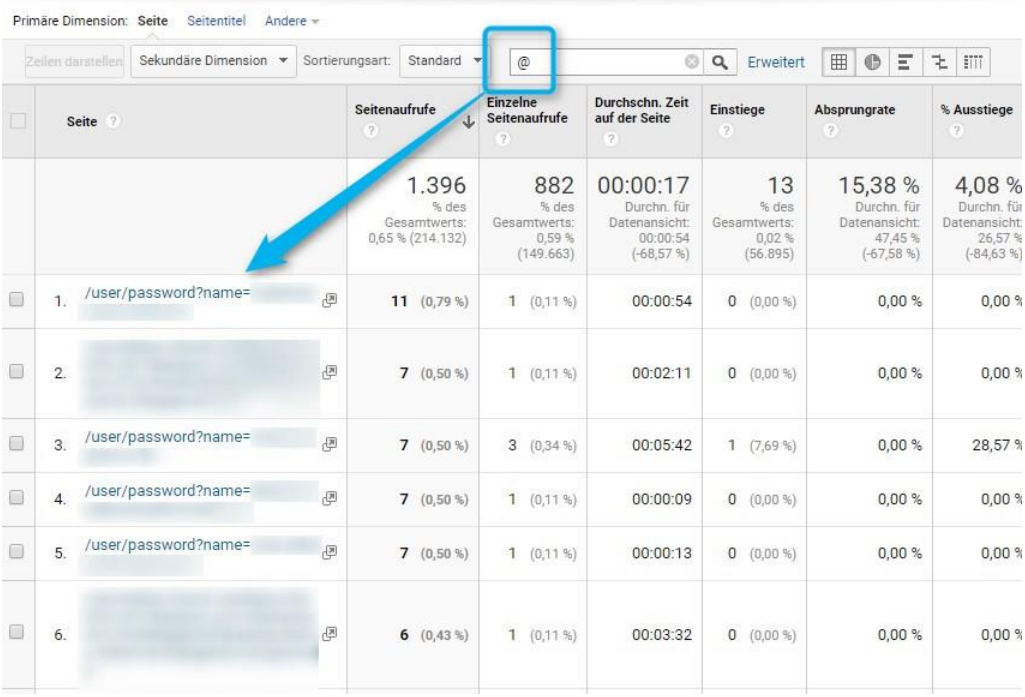

Abb. 1: So leicht kannst du überprüfen, ob in einer Dimension evtl. E-Mail-Adressen vorhanden sind.

Für so einen Fall lässt sich sogar eine benutzerdefinierte Benachrichtigung in der Verwaltung setzen, die dich schon am Tag danach darüber informiert, dass da „etwas mit einem @“ eingegangen ist (siehe Abb. 2).

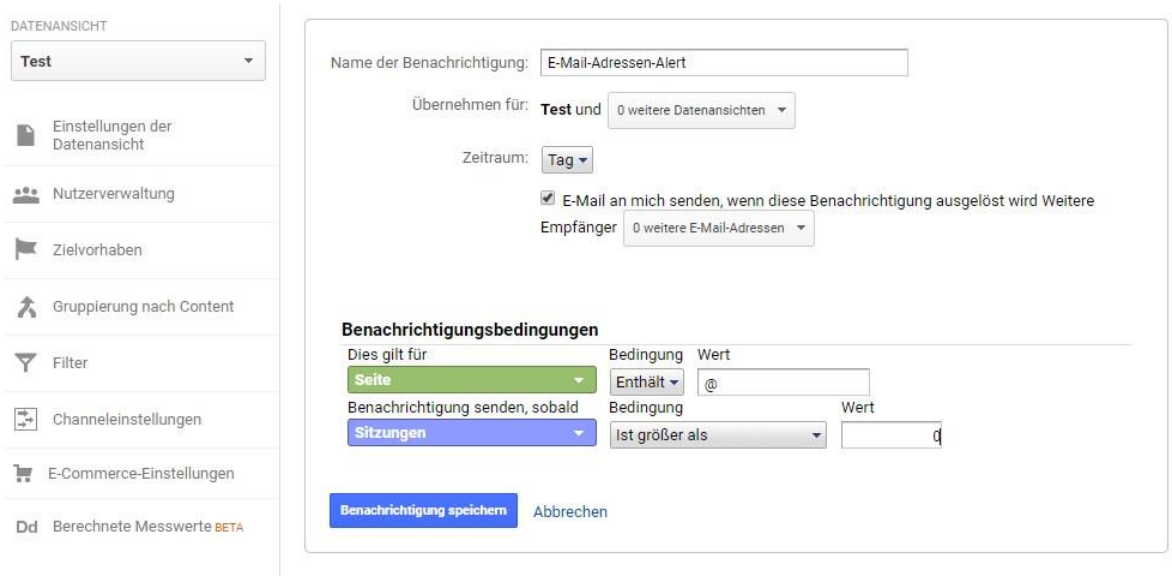

Abb. 2: Eine benutzerdefinierte Benachrichtigung kann bei der Aufdeckung helfen.

Auch benutzerdefinierte Segmente, die mit verschiedenen regulären oder einfachen Ausdrücken (z. B. „+49“ für Telefonnummern oder einfach „passwort“ oder „password“, um herauszufinden, ob irgendwo ein Parameter dieses Namens existiert) für eine genaue Suche nach personenbezogenen Daten bestückt werden, können dabei helfen, entsprechende Datenbestände aufzudecken.

Doch du hast wesentlich mehr Optionen, nach diesen PII zu suchen. Dazu gleich mehr.

Prio A: Personenbezogene Daten verhindern

Doch wie schon erwähnt: Wenn die Daten erstmal in Google Analytics aufgelaufen sind, ist das Problem schon da. Definitiv besser ist, wenn schon vor dem Absenden der Daten an Google Analytics nichts mehr von den kritischen Stellen übrig bleibt.

Darauf weist auch Google in seinen „Best Practices: Keine personenbezogenen Daten senden“ unter https://support.google.com/analytics/answer/6366371?hl=de hin.

Filtern der PII in der Datenansicht hilft nicht

Der Irrglaube, dem viele bislang unterliegen, ist, dass die Nutzung von Filterfunktionen in Google Analytics ausreicht, um das Problem einzudämmen.

Zur Erläuterung: Durch Filter können eingehende Daten so verändert werden, dass sie etwas anderes zeigen, als ursprünglich in den Daten zu finden war. URLs können umgeschrieben, Quellen anders zugeordnet werden usw.

Allerding genügt das nicht: Denn Filter werden auf Ebene einer Datenansicht angelegt und greifen erst, nachdem die personenbezogenen Daten längst Bestandteil der Datensammlung geworden sind. Die Datensammlung findet nämlich eine Ebene „höher“ statt, auf Property-Ebene. Das zeigt die Abb. 3.

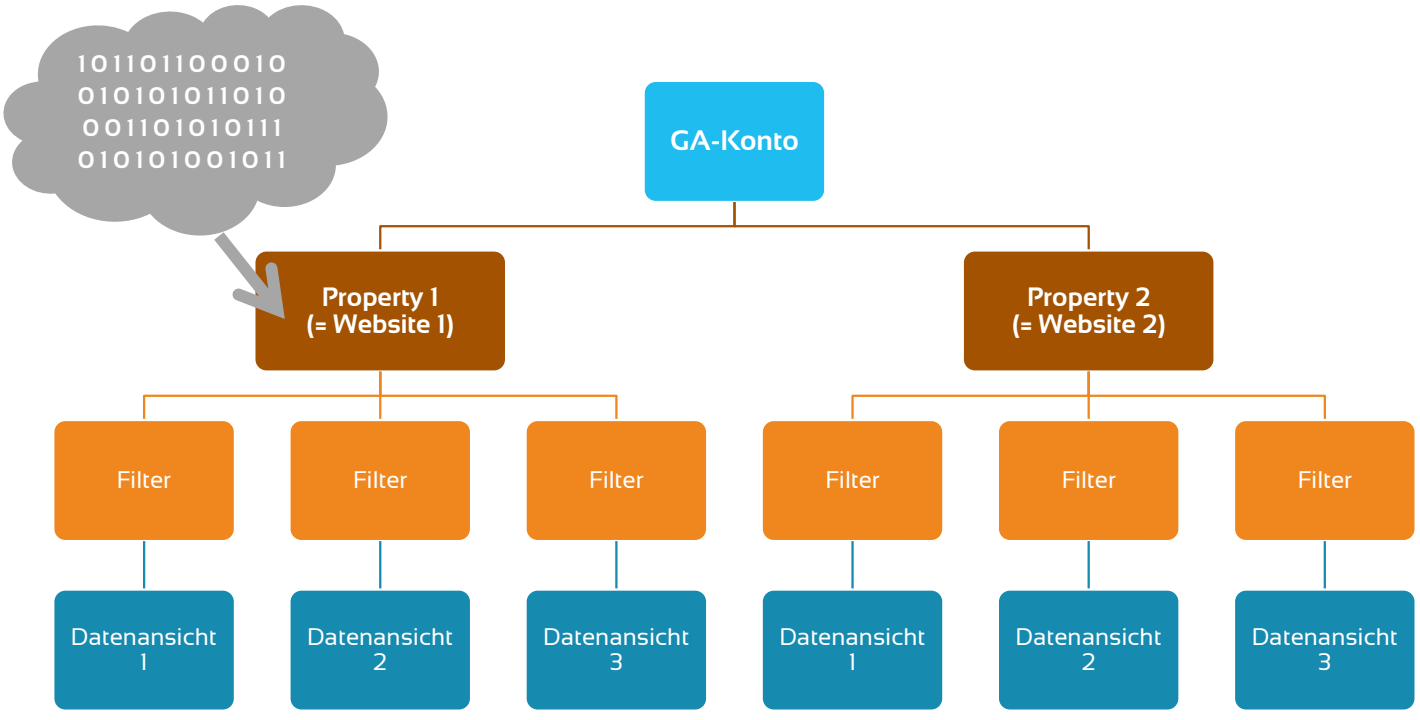

Abb. 3: Daten fließen auf Property-Ebene in das Analytics-Konto.

Das heißt: Auch wenn du die Daten in der Datenansicht nicht mehr siehst, weil Filter es verhindern, sind sie technisch trotzdem vorhanden. Und das darf eben nicht sein.

Und obschon es natürlich Möglichkeiten gibt, die Übergabe von Daten bei einer „hard coded“-Lösung zu verhindern, setzen die meisten Lösungen auf die Nutzung des Google Tag Managers. Denn dort lässt sich zentral die Möglichkeit schaffen, vor JEDER Übergabe an Analytics die entsprechenden Daten zu löschen oder zu ändern.

Möglichkeit 1: Der Google Tag Manager als Waffe gegen PII

Wenn du deine Analytics-Tags mithilfe des Google Tag Managers ausspielst, hast du im Grunde genommen alle Karten in der Hand, um personenbezogene Daten zum allergrößten Teil und sogar recht bequem auszufiltern.

So gibt es etwa die Möglichkeit, entweder die Daten komplett vor der Übergabe an Google Analytics zu löschen oder in etwas Unkritisches zu verändern. Ein beliebtes Beispiel sind etwa E-Mail-Adressen, die es immer mal wieder in die Datensammlung von GA schaffen.

Hierzu gibt es verschiedene Möglichkeiten, entsprechende Modifikationen vorzunehmen. Die für mich besten stammen von Simo Ahava und Brian Clifton (Letzterer setzte seine Programmierung auf der von Simo Ahava auf). Das Konzept sieht dabei vor, die personenbezogenen Daten nicht einfach zu löschen, sondern zu überschreiben, um in den Analytics-Daten zumindest einen Hinweis auf das Vorhandensein von PII zu haben.

Unter https://brianclifton.com/blog/2017/09/07/remove-pii-from-google-analytics/ findet sich ein Beispiel für eine benutzerdefinierte JavaScript-Variable, die für diesen Beitrag nochmals stark angepasst wurde und jetzt etwa auch deutsche Postleitzahlen und Telefonnummern erkennt. Die im Code verwendete Domain „metrika\.de“ musst du natürlich deine eigene Domain ersetzen.

Dabei bitte beachten: Dem Punkt vor der Top-Level-Domain muss ein „\“ vorausgehen, da es sich hier um einen regulären Ausdruck handelt und der Punkt ansonsten als Steuerzeichen interpretiert wird.

```
function() {
  return function(model) {
    // Add the PII patterns into this array as objects
    var piiRegex = [{
      name: 'E-MAIL',
      regex: /.{4}@.{4}/g
    },{
      name: 'E-MAIL EIGENE DOMAIN',
            regex: /[^\/]{4}@(?=metrika\.de)[^\/]{4}/gi
          },{
      name: 'PASSWORT',
      regex: /((passwort=)|(password=)|(passwd=)|(pass=)).*/gi
    },{
      name: 'POSTLEITZAHL',
      regex: /((plz=)|(postleitzahl=))+(d[-]{0,1}){0,1}\d{5}/gi
    },{
      name: 'TELEFONNUMMER',
      regex: /(\+|0)+\d+(.?[\/\-]?.?\d+)*/g
    }];

    // Fetch reference to the original sendHitTask
    var originalSendTask = model.get('sendHitTask');

    var i, hitPayload, parts, val;

    // Overwrite sendHitTask with PII purger
    model.set('sendHitTask', function(sendModel) {
      hitPayload = sendModel.get('hitPayload').split('&');
      for (i = 0; i < hitPayload.length; i++) {
        parts = hitPayload[i].split('=');
        // Double-decode, to account for web server encode + analytics.js encode
        val = decodeURIComponent(decodeURIComponent(parts[1]));
        piiRegex.forEach(function(pii) {
          val = val.replace(pii.regex, '[REDIGIERT ' + pii.name + ']');
        });
        parts[1] = encodeURIComponent(val);
        hitPayload[i] = parts.join('=');
      }
      sendModel.set('hitPayload', hitPayload.join('&'), true);
      originalSendTask(sendModel);
    });
  };
}
```

Wer sich gerade in die Zeiten von alten C64-Magazinen mit mehrseitigen Code-Beispielen zum Abtippen zurückversetzt fühlt, kann einen Google Tag Manager Container mit der entsprechenden Custom-JavaScript-Variablen unter der Adresse am Ende dieses Beitrags downloaden.

Der Code definiert zum einen die typischen Schreibweisen für E-Mail-Adressen, Postleitzahlen, Telefonnummern oder identifiziert typische Parameter, die Passwörter oder Postleitzahlen einleiten, und überschreibt sie mit „REDIGIERT“. So ist in Google Analytics zwar die personenbezogene Info nicht mehr zu sehen, jedoch, dass etwas „gelöscht“ wurde. Das kann bei der Suche nach entsprechenden Quellen helfen.

Doch bitte: Dieser Code muss auf jeder Domain erst ausgiebig getestet werden, bevor er live geschaltet wird. Es muss zuvor sichergestellt werden, dass keine „normalen“ URLs betroffen sind.

Und: Im obigen Code werden nur einige Möglichkeiten von personenbezogenen Daten angesprochen. Sie sollen das Prinzip klar machen. Doch auch Klarnamen, Kreditkartennummern und viele andere können mit ein wenig Unterstützung eines Programmierers über ähnliche Methoden oftmals ausgeschlossen werden.

Einbau in den GTM

Dieser Code muss nun im Google Tag Manager jedem Hit vorangeschaltet werden. Das geht so:

1. Den obigen Code in eine neue Variable vom Type „Benutzerdefiniertes JavaScript“ einbetten (siehe Abb. 4).

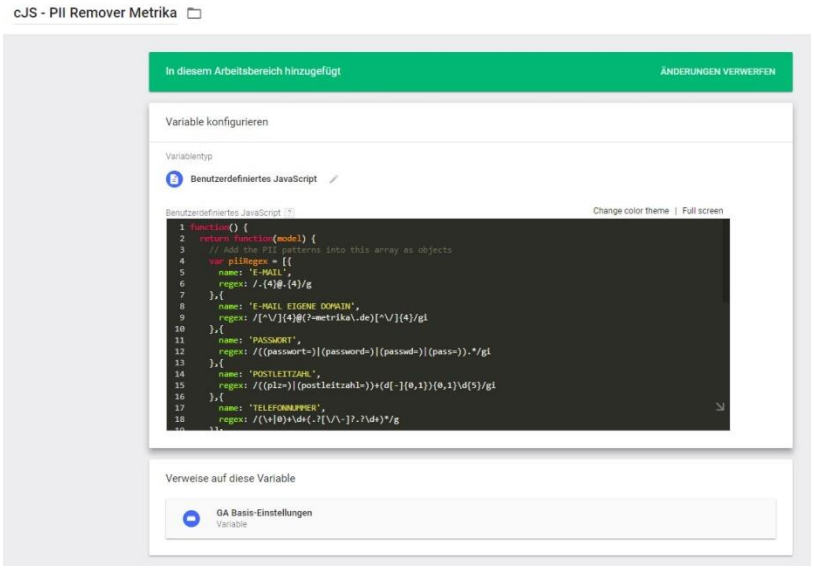

Abb. 4: Eine neue Variable anlegen (benutzerdefiniertes JavaScript)

2. Den Code fügst du idealerweise in der global verfügbaren Variable „Google Analytics Einstellungen“ in den festzulegenden Feldern als „customTask“ mit Verweis auf das vorhin erstellte Stückchen „benutzerdefiniertes JavaScript“ ein (siehe Abb. 5). Natürlich nur, sofern eine solche Variable schon für die Analytics-Tags genutzt wird. Die Nutzung hat allerdings den charmanten Vorteil, nicht jedes einzelne Analytics-

Tag bei globalen Änderungen anpassen zu müssen, sondern „nur“ die eine Variable.

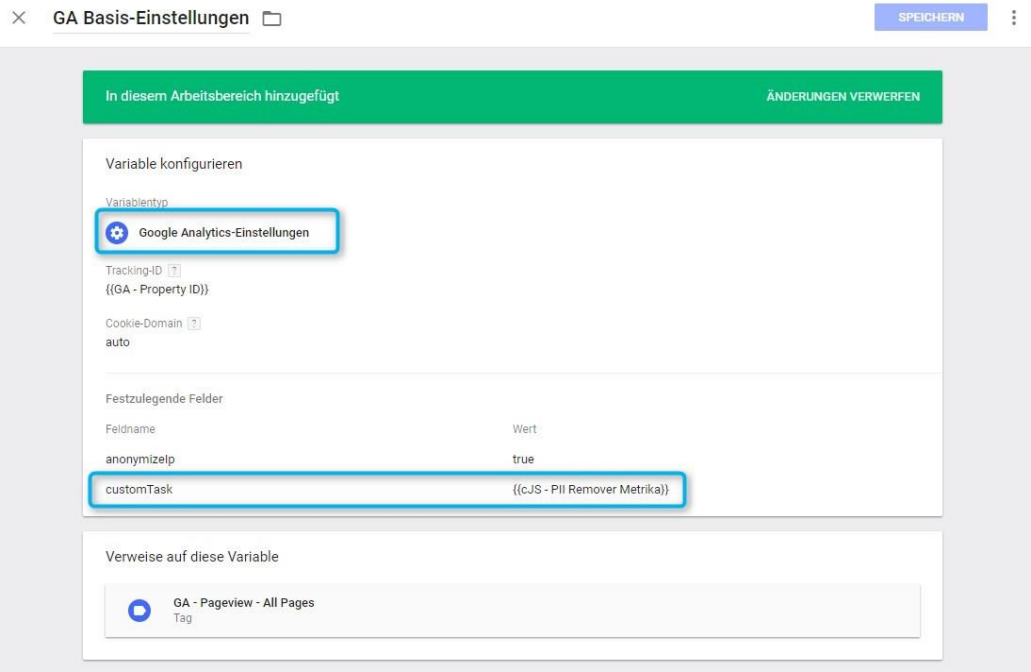

Abb. 5: ... z. B. in der Variablen „Google Analytics Einstellungen“ als customTask einbinden.

Ansonsten muss dieser Schritt tatsächlich bei jedem einzelnen Google Analytics Tag durchgeführt werden.

Durch die Einbindung als customTask wird jeder einzelne Hit, der an Analytics geschickt wird, mithilfe der JavaScript-Variable entsprechend modifiziert.

In Google Analytics laufen die Daten dann so ein wie hier in Abb. 6.

Abb. 6: Die Daten werden so modifiziert, dass kein Personenbezug mehr möglich ist.

Möglichkeit 2: „Vorbeugen ist besser als heilen“

Jeder kennt den Spruch aus dieser Zwischenüberschrift. Auch hier findet er wunderbar Anwendung, nämlich wenn die an den Webanalyse-Prozessen beteiligten Personen schon frühzeitig sensibilisiert werden für das, was passieren kann.

Es sollten also Analysten, Marketer, Entwickler, Datenschutzbeauftragte und Co. zu Beginn eines Webprojektes auch zu diesem Punkt befragt und zurate gezogen werden. Alle Beteiligten sollten, nein, müssen wissen, was in Google Analytics in puncto „personenbezogene Daten“ erlaubt ist und was nicht. Das verhindert bspw. schon zu einem frühen Zeitpunkt, dass die Dinge aus dem Ruder laufen.

Dementsprechend können frühzeitig Weichen gestellt werden. Etwa, indem bei der Konzeption von Newslettern verhindert wird, dass Links zur Website mit personenbezogenen Daten (z. B. Login) ausgestattet werden, dass Kunden-Bereiche mindestens URL-seitig frei bleiben von Klarnamen und anderen PII und viele andere Dinge mehr.

Und du oder die Personen in Charge sollten den Mut und das Mandat haben, das Projekt im Zweifel an der Stelle zu stoppen, wenn nicht sichergestellt werden kann, dass keine PII übertragen werden.

Andernfalls droht der Datenverlust.

Rechtsberatung? Mitnichten. Ein Disclaimer.

Wenn du dir nicht sicher bist, was im Sinne von Google oder dem allgemeinen Datenschutzrecht „personenbezogene Daten“ sind, solltest du dich unbedingt mit deinem Rechtsbeistand zusammensetzen. Diese wichtige Arbeit kann dieser Artikel nicht überflüssig machen.

Unter dieser Adresse kann der Google Tag Manager Container mit dem „PII-Remover“ heruntergeladen werden:

https://go.metrika.de/PIIRemoverContainerP

Mein Fazit: PII sind kein unlösbares Problem

Das Wichtigste ist schon geschehen, wenn man sich diesen Beitrag durchgelesen hat: Sensibilisierung für das Thema. Und: Sensibilisierung für die Konsequenzen aus der Nicht-Beachtung. Also: Es geht darum, sich die eigenen Daten in Google Analytics nochmal unter dem Aspekt der personenbezogenen Daten anzusehen – und zu handeln. Der Tag Manager Container kann dabei ggf. helfen.

Wenn du auf dem Laufenden bleiben möchtest, melde dich doch einfach für unsere Data Stories an. Dort gibt es immer mal wieder Neuigkeiten aus der Webanalyse. Kein Spam, versprochen.

Der Autor: Maik Bruns

Der Ex-SEO ist seit Jahren der Webanalyse mit Google Analytics und dem Google Tag Manager stark verbunden.

Er schreibt und bloggt zu diesen Themen, wenn er nicht gerade unterwegs oder bei seiner Familie ist und ist in verschiedenen sozialen Netzwerken zu finden. Er hostet den Podcast "Die Sendung mit der Metrik" und spricht außerdem zu Webanalyse- und SEO-Themen auf Konferenzen und gibt Seminare/Workshops zum Thema Google Analytics und Google Tag Manager.

Nebenbei fotografiert er leidenschaftlich gerne und hat mit Freeletics einen Sportnachfolger für Volleyball gefunden.

Twitter Google+ LinkedIn XING