Basic Tasks

Version: 1.1

Updated: 19 January 2016

Author: Nigel Boor

Contents

Aims	5
The Basics	6
The structure of Google Analytics	6
Main menu	7
Reporting	9
Setting the date range	9
Basic reports (Overview reports)	9
Audience	9
Acquisition	9
Behaviour	10
Conversions	10
Most commonly used	10
Hits	10
Visits	10
Other indicators you may need to find	10
Filtered reports	11
Quick filters	11
Segments (Available across your Google account)	16
Output: Exporting or sending your reports	21
Save as	21
Email to	22
Saving reports: Setting up regular reports using the 'Customisation' menu	24
Specials: Niche reports	27
Dashboards: A collection of report widgets	30
Shared	30
All GCC accounts have standard dashboards	30
Private	30

Error! Bookmark not defined.	Add any personally useful widgets
31	Appendix 1: RegEx
31	Starts with
Error! Bookmark not defined.	Ends with
Error! Bookmark not defined.	"OR"
31	Special Characters
Error! Bookmark not defined.	Testing
	CM Tricks to make life simpler
	Using the multi edit grid to extract
rom a list of Articles35	Using Xenu's Link Sleuth to collect
Error! Bookmark not defined.	Account and Property info

Aims

After reading this document, you should be able to produce basic reports of traffic and visitors to the web site and identify the sources of that traffic.

If you want in-depth knowledge, go to the Analytics Help Center or Analytics Academy.



You may see this bomb scattered around the text.

Do yourself a favour and read the highlighted text alongside. These tips are there to stop you hitting some of the landmines Google Analytics leaves.

The Basics

To use Google Analytics, you need a Google account. This is the "One account. All of Google" you will have signed up for if you have Gmail. Use your personal account for most of your work on Google Analytics, only switch to the main GCC account for admin tasks and always check with SysAdmin¹ before you do.

You also use your Google Account for other apps such as:

- You Tube
- Gmail
- Google Drive
- Google Photos
- Signing on to your Android phone
- etc.

Start Google Analytics by going to https://www.google.com/analytics and hitting the "Sign In" link – blue and top right currently.



Google love moving things around and adding stuff temporarily.

The "sign in" link is currently blue and located top right of the Google Analytics home screen but it may move or change colour at Google's whim!

The structure of Google Analytics

Analytics has three levels

Account

An account is your access point for Analytics, and the top-most level of organization. Normally we would only have one but GCC run several accounts, for various historical and logical reasons.

Our main account is www.gloucestershire.gov.uk.

We also have access to other Google Analytics accounts such as Achieve Forms. We cannot manage these just create reports from them.

Property

A property is normally a website, mobile application, or device (e.g. a kiosk or pointof-sale device.) An account can contain one or more properties.

We have several properties within the main www.gloucestershire.gov.uk account; the

¹ Current SysAdmin is Nigel Boor but check with the Communications Manager, Creative and Consultation if you are unsure.

normal one we use is http://www.gloucestershire.gov.uk.

We have several other properties set up to collect data for external sub-sites, eg:

- o http://www.gscb.org.uk
- o http://staffnet.gloscc.gov.uk

View

A View is a defined view of data from a property and your access point for reports.

The normal View we use is www.gloucestershire.gov.uk.

We apply some basic filters at this level, the most common being "LC all" to switch all URLs to lower case and eliminate the tracking of "Default.aspx" and "default.aspx" as different addresses.

Another filter only available at this level is the IP address filter. We have a filter to exclude all GCC network computers so you can see traffic from the public only and we apply this to the "External Traffic Only" View.

Main menu

When you open Google Analytics, it will switch you to the first View within the first Property of the first Account on your list and you will need to switch to the View you want to report on.

You can always see who you are logged in as and which View you are using by checking the top right of the screen.



Use this drop down to see the views tree and find the correct View.

- www.gloucestershire.gov.uk
 - http://www.gloucestershire.gov.uk
 - www.gloucestershire.gov.uk



Unless you understand what the other views do, always use...

www.gloucestershire.gov.uk > http://www.gloucestershire.gov.uk > www.gloucestershire.gov.uk

You will see "Home", "Reporting", "Customisation" and "Admin" on the main menu. Google Analytics starts in the "Reporting" area and you will do most of your work here.



Do not go into the GCC "Admin" menu unless speak to SysAdmin first.

You only need to use "Admin" to check the tracking code for the Property against the site you are reporting on and to manage scheduled reports.

We will come back to "Customisation" later when we discuss "Setting up regular reports using the 'Customisation' menu".

Reporting



The first part of this section describes how to set up one off reports. Once created, you can export or email them but they are not easy to find. If you may need to edit them, see "Saved reports" on page 24.

Remember to use

www.gloucestershire.gov.uk

- → http://www.gloucestershire.gov.uk
 - → www.gloucestershire.gov.uk

Setting the date range

Choose the date using the calendar that appears when you hit the date drop down. There are several pre-set date ranges or you can select a custom range by either typing in a valid date or selecting it from the calendar.

You also get an option to compare with a previous period producing two sets of metrics.

Basic reports (Overview reports)

When you open Google Analytics and select the View, Google will drop you at the "Audience Overview" report, which is a good place to find the basic session and demographic data.

Google Analytics groups reports by the kind of data they show:

Audience

Data for

- · Who they are
- Where they come from
- What they are using

Acquisition

Data showing how they got here.

- Advertising
- Search
- Social media
- Internal campaigns (email newsletters, press releases)

Behaviour

Data showing what they do and where they go when here

- Pages visited
- Page load time
- Site search terms

Conversions

This is an under-used report that can be useful in finding out whether we are meeting our aims. You can set up "Goals" which match the aims of the web site (eg "get to the archives site" or "find 'Report-it'") and track how many of our visitors end up completing these goals.

Most commonly used reports

People usually ask for two indicators: "hits" and "visits" or "visitors". Google uses different terminology but you can get the answer quite easily.

Hits

Use "Unique Page Views".

Found in the left menu under:

- Behaviour
 - Site content
 - All pages

Visits

Use "Sessions".

Found in the menu under:

- Audience
 - o Overview

Another indicator you may need to find: Source of visit

Use "Sessions" or percentages as appropriate.

Depending on the type of "Source":

For how they got here – ie search (Google calls it "Organic Search") or another route – use:

- Acquisition
 - Overview

This report is a "drill down" type, which means if you want to dig deeper into "organic search" you can click the row title and Google Analytics will show the search terms. If you click

"referral", Google Analytics will show which site referred the visitor. Social will show which network.

To see which type of kit they used to get here – (desktop, tablet or phone), use:

- Audience
 - o Mobile
 - Overview

Filtered reports

The reports so far show data for the entire site and its sub-sites. You will usually want to restrict the report to a sub-set of data so you only see traffic to the areas that interest you. Filters offer a way to do this.



It is possible to filter to see only traffic you need. Build and test your filters carefully. You should know exactly what should appear in the results and need to investigate if there are any odd or missing elements.

Direct filters

These are only available on "hit" type² reports; if they are, you will see a filter box above and to the right of the tabular data on the report. It looks like a search box and has an "advanced" link to the right.



Simple filters

Type an article ID or friendly URL into the filter box and usually you will see that item in the filtered report. Sometimes you will see more than one but if you are just after the stats for a single page this is a quick way of getting that.

If you need anything smarter, you will need advanced filters.

Advanced filters

Clear any filters in the simple box and hit the "Advanced" link to open the filter builder.

By default, you can only use the filter builder for multiple parameters with "AND".

E.g. you may want to see stats for an article on the /extra site with article ID = 108088. You can reference this in several ways so the simplest way to narrow it down is to ask for all

² During a session, the user interacts with your property. Each interaction is referred to as a hit. Example hits include page views, events, transactions, etc.

articles where the "Page" (which is what Google Analytics calls the URL) starts with "/extra" AND the "Page" contains 108088.



The above filter will produce a report with several rows showing some of the alternative ways the visitors may have found the article. If some of them are invalid, you could use the "Exclude option to remove them until you get just the rows you want.

This can be a tedious process and is limited by the "AND" combination. Remember that articles can have "Friendly URLs" so you need to check for those too, but this condition would need an "OR" option not available in a simple filter.

Regular Expressions are far more flexible and powerful but can be complicated until you get to know the syntax.

Advanced filters with Regular Expressions: RegEx³

This simplest way to show you Regular Expression is to build one.



Regular Expressions – known as RegEx – are very powerful but you need to know what you are doing. Follow the section below carefully, read the Google Analytics RegEx help and use their RegEx tester to help track errors.

The brief is to "show all traffic to the School Admissions area, including the documents downloaded". Seems simple but "it's a little bit more complicated than that⁴". Fortunately, there is a relatively simple way to build the query. Once you master this method, it should not take long to construct a RegEx, even if it seems complex at first.

There are three quite distinct steps:

1) Find the Article IDs and Friendly URLs

iCM offers a simple way to do this through the multiple edit grid

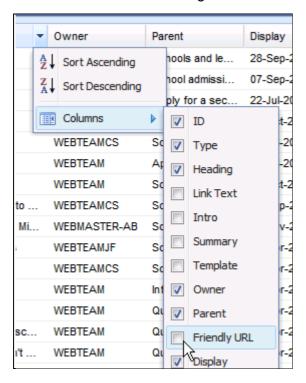
• Open the tree from the root article down, making sure all branches are expanded

³ For a more complete discussion of RegEx see Appendix 1: RegEx

⁴ Goldacre, B., 2014. *I Think You'll Find It's a Bit More Complicated Than That* 1st ed., London: 4th Estate. Available at: http://www.4thestate.co.uk/book/i-think-youll-find-its-a-bit-more-complicated-than-that-4/.

- · Select all articles, including any that are switched off
- · Right click any one and select "Edit these articles" which opens the grid
- To show Friendly URLs hit the small drop down arrow that appears in the title bar on the right of any heading (hidden at first so you have to hover).

Follow the menu down through columns and check Friendly URL to add the column



- You now have a view that shows all Article IDs and Friendly URLs (you may have to adjust column widths so see everything)
- On a small view you can copy these into Notepad++5
- On larger views there is a trick to extract the IDs easily see "Using the multi edit grid to extract Article IDs" on page 35.

2) Find the Media IDs

- You can see "Related" media in the multiple edit grid open the list using the "+" next to the article title. You can extract them from here but it is likely that there will also be multiple inlines, which you cannot see in the grid. Manually extracting them is only viable for a few pages so you will need a method similar to that explained in "Using Xenu's Link Sleuth to collect the media links from a list of Articles" on page 35.
- Notes:

⁵ Notepad++ is approved Open Source software, which ICT should install on request.

- If you are going to extract the Media IDs using an automated method, you need the Article IDs first.
- This method is dependent on having Xenu Link Sleuth, which you can have installed by ICT. You can use other link checkers to extract the links you need.

3) Build and test the RegEx

Assuming you have done the ID extractions outlined above you will end up with three lists

- 1. Article IDs
- 2. Friendly URLs
- 3. Media IDs

Article IDs	Friendly URLs	Media IDs
104800	whereilive	51363
106446	transfersecondaryschool	51364
114455	startingprimary	56737
115145		57111
106445		

RegEx for Article ID

In plain English I need to ask Google Analytics to show everything that starts with either "/" or "/article/", followed by one of the Article IDs in the list⁶. The RegEx looks like:

^((/|/article/)(104800|106446|114455|115145|106445))

Interpretation of the above RegEx:

^ The RegEx code for "at the start of the string"

() - Used as in maths to group expressions

/|/article/ - "|" is the RegEx for "OR" so this reads "/" OR "/article/"

104800|106446... - the article IDs separated by "OR"

⁶ If you have noticed the older link methods using "articled=nnn" you may have to adjust the RegEx to include these links.

It is advisable to test each section as you build it so paste the above RegEx into the "Advanced" box of the "All Pages" report and hit "Apply".



If everything is correct, you will see a graph with a table of links below that contain only the Article IDs you asked for. If something is wrong, one of two things will happen:

- If you have made an error in the expression, Google will stop and flag:
 "Your data request includes an invalid regular expression."
 Go back and fix it (usually it is a missing bracket or something just as simple).
- 2. The report appears with no content.

This is a more difficult problem because it could mean that your expression uses incorrect logic.

Of course, it could also mean that there has been no traffic to those pages in the report period.

RegEx for Friendly URLs

This is a simple list

^(/(whereilive|transfersecondaryschool|startingprimary))

Test this and you will notice that "whereilive" shows no traffic. That is because it is a "Redirect" article so the "hit" is registered at the destination. In this case, the destination is off site so we do not log the traffic.

Test the above RegEx and you will see an anomaly with a link "/startingprimary2013" which we did not want to include. There are also some links with "index.[aspx|cfm]" at the end which are generally errors caused by robots spidering your site. You could add some regex to exclude them, or alternatively just add "\$" to the end of the above expression which means "at the end of the string" effectively ignoring anything with characters after the friendly URLs listed, so the RegEx now looks like:

^(/(whereilive|transfersecondaryschool|startingprimary))\$

RegEx for Media IDs

This time we are looking for the end of a string.

 $((51363|51364|56737|57111)\&p\=0)$ \$

The first part is standard but there are some unfamiliar characters at the end

$\ensuremath{\mbox{\mbox{$\sim$}}} 0 = /q \slash \$

"\" in RegEx means use the following character as a character rather than an operator so the above means literally use "&p=0" without any special meaning that "&" or "=" might have in RegEx

The "\$" at the end of this expression, as explained above, means find the bit in the outer brackets only at the end of the string.

Composite RegEx

 $((/|/article/)(104800|106446|114455|115145|106445))|^{(/(whereilive|transfersecondaryschool||startingprimary))}|((51363|51364|56737|57111)\\$

Dealing with Anomalies

As you saw when testing, "RegEx for Friendly URLs" on page 15 there can be odd links, which are inconsistent with the way our site works. To narrow the results to only valid traffic you need to exclude these.

First test the link – just because Google logs it does not mean we deliver it. In this case, test: As seen above:

http://www.gloucestershire.gov.uk/transfersecondaryschool/index.aspx

is not valid so we should exclude it by adding a line to the advanced filter:

Apply the adjusted filters until you have the final report.

Building filters with larger lists of articles and media is the same, just a bit more faff and more prone to tryping errers.

For more advanced expressions, see Appendix 1: RegEx.

Segments (Available across your Google account)

If you need to apply filters to user or session reports, which do not have a filter box, you will have to use "Segments".



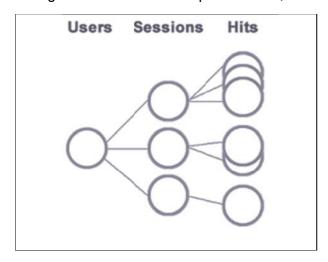
Segments work differently to filters –you are dealing with sessions or users, not page views– so they can produce unexpected results. You will need to build them carefully and check for unexpected traffic.

Segments represent either subsets of sessions or subsets of users: for example, all sessions that visit a particular area of the site, or all session using a mobile device.

The Google Analytics User Model will help you see how segments work. There are three major components:

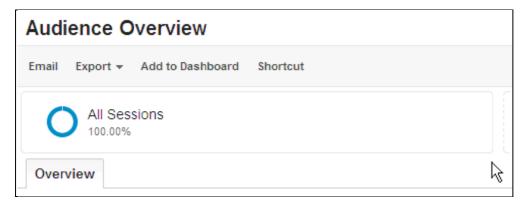
- Users
- Sessions: A user interaction with your property in a single period.
- Hit: Each interaction with your property is a hit.
 Example hits include page views, events, transactions, etc.

A single user can have multiple sessions, and each session can have multiple hits.



From this, you can see that we are filtering one level up from where you applied the report filters. This means that you cannot filter the hits directly so your segment may contain page views from anywhere on the site.

Segments are active everywhere but the default is to show "All Sessions".



Once you have segments configured, you can choose to apply one or more to your reporting session to narrow the data to the segment you want to analyse. Click the "All Sessions" button to open the list of available segments and the "New Segment" button.

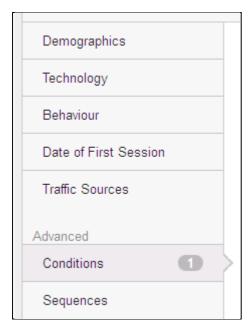
There are three ways to create a segment:

- from scratch it is almost the same as creating a filter and you can see this method outlined below
- 2. by importing from the gallery
- using a link shared by someone else
 (if you have Google Analytics open try following
 https://www.google.com/analytics/web/template?uid=vKJlf6pFQf-j6K3Z-Mv4ng)

Building from scratch

You will see one of these in the example below.

The basic options offer some quick ways to drill into commonly used segments.



The options mostly are self-explanatory. Using "Demographics" for example, you can extract age, gender, language and geographical segments as well as some that only work if you are using Google Ads. These options are beyond the scope of this How-To but experiment to see if any are useful to you.

Advanced "Conditions" example using RegEx

The most flexible method of building a segment is by using "Conditions. This example builds a "Conditions" segment to exclude all the configured sub sites⁷. This gives you a segment of sessions where the user only visited the main site. If they go to /extra or /archives their sessions will be excluded.

⁷ There are some issues with sites using domains other than gloucestershire.gov.uk which could pollute these results and cleaning up these issues will be one of the tasks of updating the web system for the Customer Programme.

If you have iCM admin rights you can see the configured sub sites in the /enterprise/ view under "Management" → "Definitions". The current active sub site list is:

- http://www.gloucestershire.gov.uk/archives
- http://www.gloucestershire.gov.uk/extra
- http://www.gloucestershire.gov.uk/gloucestershirecaresector
- http://www.gloucestershire.gov.uk/gsab
- http://www.gloucestershire.gov.uk/gscb
- http://www.gloucestershire.gov.uk/inform
- http://www.gloucestershire.gov.uk/ld-partnership
- http://www.gloucestershire.gov.uk/ltb
- http://www.gloucestershire.gov.uk/rfg
- http://www.gloucestershire.gov.uk/schoolsnet
- http://www.gloucestershire.gov.uk/telecare
- http://www.gloucestershire.gov.uk/vschool

In English, what we are trying to do is exclude all traffic where the address starts with "/" followed by any of the highlighted names in the subsite list above. If you remember, the previous example from "Advanced filters" above, the RegEx should be simple.

- ^(/ = starts with "/"
- (archives|extra|gloucestershirecaresector|gsab|gscb|inform|ldpartnership|ltb|rfg|schoolsnet|telecare|vschool)
 - = list of the highlighted terms above with the "OR" operator "|"
-) = closes the parenthesis opened with "(/"

The composite RegEx for the segment is therefore:

^(//(archives|extra|gloucestershirecaresector|gsab|gscb|inform|ld-partnership|ltb|rfg|schoolsnet|telecare|vschool))

Simply use this in the conditions tab by choosing the appropriate metric and method. In the example below, we are using the RegEx developed above to exclude sessions that visited other sites.



You can build more complex segments by adding "AND" or "OR" modifiers to the current filter more filters which means you can build very complex and flexible segments to drill into specific parts of your audience.



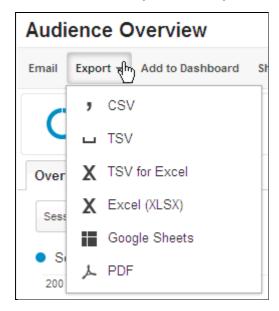
It bears repeating: Segments can produce unexpected results. You will need to build them carefully and check the result for unexpected traffic.

Be warned!

Output: Exporting or sending your reports

Export and save options

There are several options for export



- CSV: Comma Separated text files which most spreadsheets and databases can import. Commas within page titles can throw off this format but Google Analytics should add quotes around "text" to delimit this and overcome the problem in most cases.
- TSV: Tab Separated text file, which most modern spreadsheets will also handle. It is
 less prone to problems with commas in text values as the delimiter is the tab
 character, but fewer programmes offer tabs as a delimiting option. (Thankfully all
 modern spreadsheets do.)
- TSV for Excel: I can't see any difference between this and TSV
- Excel (XLSX): Exports direct to a spreadsheet with pages for summary and detail.
- Google Sheets: Unlike the rest of the options here, this creates a spreadsheet on your Google Drive, rather than as a file downloaded to your device. Handy if you are tablet based where the local file system is less friendly.
- PDF: The most common format to share the report as you are sharing the graph and table exactly as you see on screen.

Before you export, make sure you select the correct date range and desired number of rows to display as this governs the data you will export. Typically, a global report will include data for a year or quarter and show the first 50 or 100 rows, whereas you would normally expand a detail report to show all rows and often use a shorter time span.



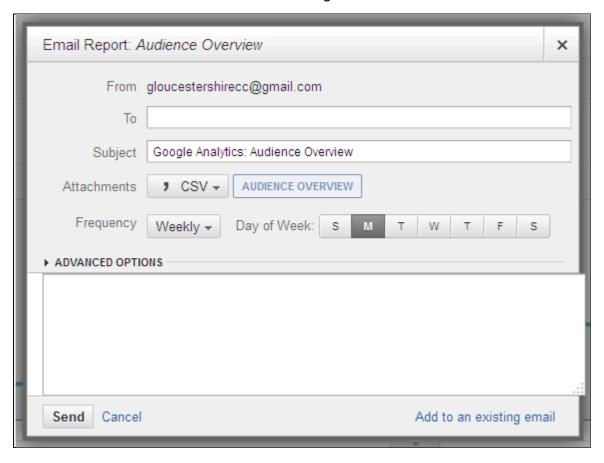
/ Always check your output!

Failing to set the rows to an adequate number is a very common problem.

It is the Google Analytics equivalent of forgetting the email attachment.

Email to

You can email the above formats either as a single shot or as a scheduled series.

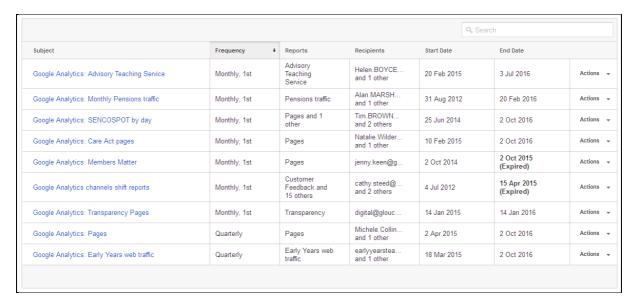


These are the defaults and it is a relatively simple form. You must provide an email and a message and remember to select the format and frequency, especially if you only need to send the report once.

If you choose a periodic delivery, you can choose to add it to an existing email delivery to cut down on the number of separate emails your clients receive. If you set up a new scheduled delivery, Google Analytics sends your report on that schedule for the next 6 months. The "Advanced Options" allow you to choose how long to keep sending scheduled reports.

Managing scheduled email

Once specified you cannot edit email schedules from the reporting area. To do this you will need to switch to the "Admin" menu and find "Scheduled Emails" at the bottom of the "Views" menu. This opens a table from which you can review and manage the email you have scheduled.



Click the subject link in the left column to change the schedule. You cannot change the attachments but you can change the recipient's email and adjust the schedule.

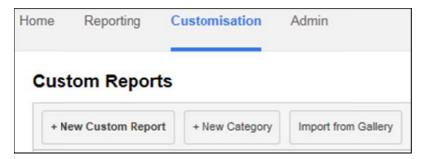
The actions drop down in the right column allows you to delete the email or extend the delivery period.

Saved reports

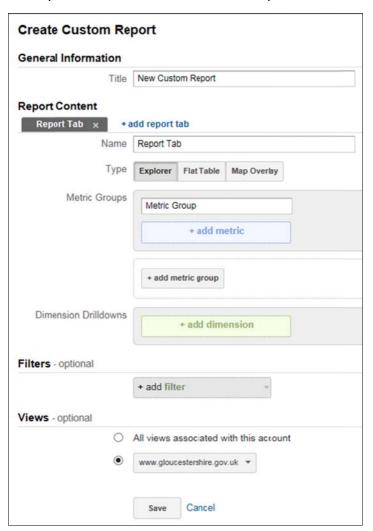
Setting up regular reports using the 'Customisation' menu

Familiarise yourself with the available dimensions and metrics before building here

Under the 'Customisation' menu to create or import a report



This opens a blank form to create the report.



The items should be looking familiar by now & you may be able to create a report using the knowledge you already have. What follows is essentially a recap with some hints to help you make the most of the custom report.

Sections

General Information

Title: Any meaningful name

Report Content

Report Tab

It is possible to build reports with multiple tabs but all tabs must use the same filter

- Name Describe the stats this tab will show
- Type
 - Explorer

This is the type of report you will be used to: graphic at the top & detail table below

o Flat Table

As it says on the tin.

More suited to single dimensions

o Map Overlay

Limits choice of dimension to geographical types with the ability to set default "Zoom" to "World", "Continent", "Sub continent" or "Country". Both are restricted to lists of available types (eg Country = UK). Setting "Zoom" filters traffic to show only traffic for the zoomed area.

Metric Groups (you must specify at least one)

Each one is shown as a column in the report; multiple metrics produce multiple columns.

Common metrics include:

- Unique page views
- o Page load time
- Session duration
- o Users
- Dimension Drilldowns

These produce the rows in the report need to be matched to the metrics above so the report makes sense. Each added dimension produces a drill down level, eg Pages >

Device category will allow users to see the device category each page is visited from. Common dimensions include:

- o Age
- o City/Country
- o Hour
- o Page
- o (Google search) Keyword
- o (internal) Search term

Filters - optional

These work in exactly the same way as above.

Views - optional

You can limit the report to

- All views associated with this account
- [the current view] Google Analytics will show the name

Once the layout has been finalised hit Save or if you want to discard changes, Cancel

Note

- You can only export/send one tab from a multi tabbed report
- However, you can set up email to deliver tabs as individual reports

Specials: Niche reports

Real time

Shows

- · Active users on site
- Desktop vs Tablet
- Page Views
 - o Per second
 - Per minute
- Top Referrals
- Top Social Traffic
- Top (Google search) Keywords
- Top Active Pages
- Top Locations



Users flow

Can give insight into where you are losing users which is especially useful when evaluating the performance of reports.

Starts by default with users split by country then show where they go with each hop.

The metric of interest when evaluating performance is the red tail in each block showing drop-offs. On a form this would show where users are abandoning the form.



Campaigns (If we ever use campaign tagging)

It is possible to build a URL which contains standard tags that Google Analytics will automatically log in the data.

You can see the data in the "All campaigns" report in the Acquisition section.



We do not make enough use of campaign tagging.

This is a very powerful tool and you should promote its use among communications officers and other web editors.

In-page

Found in the Behaviour section this report shows a kind of heat map of where on your page people click. There are two issues you may encounter:

- Simply launching the report puts quite a strain on the Google Analytics system and often leads to errors. Sometimes full screen mode overcomes these. If not, you will need to use the Chrome extension.
- Google Analytics records all clicks, even those on fly-out menus. Aech click bubble should point at some link text but fly-out menus and other overlaid items will lead to some click bubbles seeming to float in space.



You will need to activate these fly-out menus to see which menu option the clicks are anchored to. There may be other orphaned clicks triggered by popup alerts that you cannot "activate" but will need to watch out for.



Dashboards: A collection of report widgets

When you develop a report you want to use regularly, you can add it to a dashboard from the report menu.



- 1. Choose add to dashboard
- 2. Choose which dashboard
- 3. Select table or timeline

You can group similar types of reports on common dashboards.

Shared

All GCC accounts have standard dashboards

- o Basic
- o Performance
- Private

Appendix 1: RegEx

We have covered some of these in the main document but to give you a more complete reference I have included them here.⁸

Matching Characters in strings

A "character class" matches only one out of several characters. To match an a or an e, use [ae]. You could use this in gr[ae]y to match either gray or grey. A character class matches only a single character. gr[ae]y does not match graay, graey or any such thing. The order of the characters inside a character class does not matter.

You can use a hyphen inside a character class to specify a range of characters. [0-9] matches a single digit between 0 and 9. You can use more than one range. [0-9a-fA-F] matches a single hexadecimal digit (the numbers you use for colours in HTML), case insensitively. You can combine ranges and single characters. [0-9a-fxA-FX] matches a hexadecimal digit or the letter X.

Typing a caret after the opening square bracket negates the character class. The result is that the character class matches any character that is **not** in the character class. $q[^x]$ matches qu in question. It does not match Iraq since there is no character after the q for the negated character class to match.

Special Characters

Twelve characters have special meanings in regular expressions:

•	the backslash	١
•	the caret	۸
•	the dollar sign	\$
•	the period or dot	
•	the vertical bar or pipe symbol	١

- the question mark?
- the asterisk or star *
- the plus sign +
- the opening parenthesis (

⁸ Goyvaerts, J., 2015. Regular-Expressions.info - Regex Tutorial. *Regular-Expressions.info*. Available at: http://www.regular-expressions.info/.

LunaMetrics, 2008. About regular expressions. *Analytics Help*. Available at: https://support.google.com/analytics/answer/1034324?hl=en.

- the closing parenthesis)
- the opening square bracket [
- the opening curly brace {

If you want to use any of these characters as a literal in a regex, you need to escape them with a backslash. If you want to match 1+1=2, the correct regex is 1 + 1=2. Otherwise, the plus sign has a special meaning.

The Dot Matches (Almost) Any Character

The dot matches a single character, except line break characters.

gr.y matches gray, grey, gr%y, etc. Use the dot sparingly. Often, a character class or negated character class is faster and more precise.

Anchors

Anchors match a position. ^ matches at the start of the string, and \$ matches at the end of the string. E.g. ^b matches only the first b in bob.

These are most useful in isolating friendly URLs. You can use ^\/libraries\$ to match /libraries but not /libraries/booking

Alternation

Alternation is the regular expression equivalent of "or". cat | dog matches cat in About cats and dogs. You can add as many alternatives as you want: cat|dog|mouse|fish.

Alternation has the lowest precedence of all regex operators. $cat \mid dog \mod matches$ cat or $dog \mod food$. To create a regex that matches cat food or $dog \mod food$, you need to group the alternatives: $(cat \mid dog) \mod food$.

Repetition

The question mark makes the preceding token in the regular expression optional. colou?r matches colour or color.

The asterisk or star tells the engine to attempt to match the preceding token zero or more times. The plus tells the engine to attempt to match the preceding token once or more. <[A-Za-z][A-Za-z0-9]*> matches an HTML tag without any attributes (like or <h1> but not).

Use curly braces to specify a specific amount of repetition. Use $[1-9][0-9]\{3\}$ to match a number between 1000 and 9999. $[1-9][0-9]\{2,4\}$ matches a number between 100 and 99999.

Greedy and Lazy Repetition

The repetition operators or quantifiers are greedy. They expand the match as far as they can, and only give back if they must to satisfy the remainder of the regex. The regex < . +> matches first in This is a first test.

Place a question mark after the quantifier to make it lazy. <.+?> matches in the above string.

A better solution is to use the dot sparingly. Use $<[^<<]+>$ to quickly match an HTML tag without regard to attributes (like the above). The negated character class is more specific than the dot, which helps the regex engine find matches quickly.

Grouping and Capturing

Place parentheses around multiple tokens to group them together. You can then apply a quantifier to the group. E.g. Set(Value)? matches Set or SetValue.

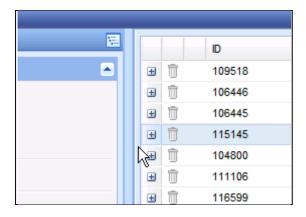
By combine grouping with repetition and alternation you can use $(\/\)$?article(id\= $\/\)$) ([1-9][0-9]{4,6}) to look for articles between 1000 and 999999, eg:/article/176904 would be a match, as would articleid=176904.

Appendix 2: iCM Tricks to make life simpler

view source, etc.

Using the multi edit grid to extract Article IDs

When you have dozens of Article IDs, you can find the Article IDs in the source code of the edit grid. You need to right click the border between the left menu and the grid to pop up the menu for the frame containing the grid.



If you hit the border, you will see a menu that will include an option to show the **frame** source. This menu varies, depending on the browser you are using. Whichever browser you use, it is important to find and use the frame source rather than the page source.

If you are familiar with HTML, you can simply scan the code looking for the array containing the Article IDs. If you are less familiar you can search for "new Array()". You will then have a list of IDs surrounded by code looking something like:

```
arrArticleNodeIDs.push('104800');
arrArticleNodeIDs.push('106446');
arrArticleNodeIDs.push('114455');
arrArticleNodeIDs.push('115145');
```

You can now use the search and replace in your editor to get rid of the code and you have a list of IDs.

You can copy and paste the Friendly URLs off the grid into the list.

Gathering the related media is a little more complex...

Using Xenu's Link Sleuth to collect the media links from a list of Articles

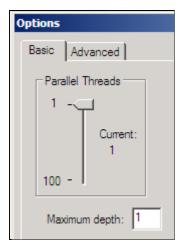
Assuming you have extracted the Article IDs using the method above, you can create a list of articles to scan with Xenu (or any other link scanner capable of checking lists of links). In the above case, you will have extracted the following list:

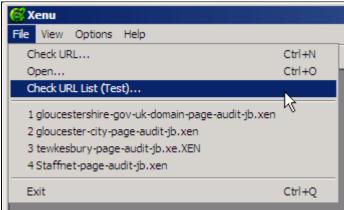
...which you can convert to a "links to check list" by simply adding the standard URL element before the ID, which gives you:

```
http://www.gloucestershire.gov.uk/article/104800
http://www.gloucestershire.gov.uk/article/106446
http://www.gloucestershire.gov.uk/article/114455
http://www.gloucestershire.gov.uk/article/115145
```

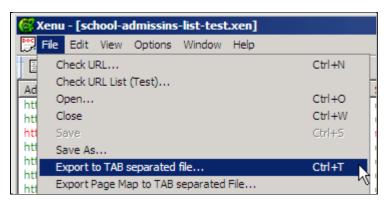
Save this to a text file – remembering where you put it and what you called it.

Start Xenu (or the link checker of your choice) and use the "check list" option. In Xenu, this is on the main menu. Note that in Xenu you need to set some preferences first to check only the pages you have in the list. On Xenu set "Maximum depth" to 1 before you start, otherwise you will spider off onto other pages on the site. Other software will have similar settings.





Once the link check has run, you will have a list of all links on the pages that you can export to TSV format for import to Excel.



Import into Excel and make sure everything separates into columns (you may have to use "Data to Columns" and do this manually). You can then sort by Address and track down the media addresses.

- On the main site a media link looks like
 http://www.gloucestershire.gov.uk/CHttpHandler.ashx?id=51363&p=0
- On Staffnet they look like http://staffnet.gloscc.gov.uk/media/31488/Stage-1/doc/Stage_1.docx

In both links, you can clearly see the Media ID that you can then extract.