From flat file to interface: Synthesis of personas and analytics for enhanced user understanding

Bernard J. Jansen | Soon-Gyo Jung | Joni Salminen

1Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
2University of Turku, Turku, Finland

Abstract
The persona approach is a communicative design technique for understanding users. With increasing access to analytics data, personas are more and more generated from online user statistics using big data and algorithmic approaches. This integration offers new opportunities to shift personas from flat files of data presentation to that of interactive interfaces for analytics systems. We illustrate this transition towards the concept of “persona as interface” with a persona analytics system, Automatic Persona Generation (APG). In pushing advancements of both persona and analytics conceptualization, development, and use, the APG system presents a multi-layered “full-stack” integration. APG affords three levels of user data presentation, which are (a) the conceptual persona, (b) the analytical user metrics, and (c) the foundational user data. Moving to a “personas as interfaces” approach offers the benefits of both personas and analytics systems and addresses some of the shortcomings of each. We provide results from user experiments of and use cases for APG. The result is a better tool than either persona or analytics along for user understanding.

KEYWORDS
data-driven personas, design methods, persona development, personas, user understanding, web analytics

1 | INTRODUCTION
A persona is an imaginary figure that represents a segment of real people (Cooper, 2004; John & Grudin, 2003). This segment can be customers, audiences, or users of a product, content, service, or system, respectively. For brevity, we use the term “user” in the manuscript but also imply a customer or audience. We use the term “product” but also imply content, service, or system feature. When data was used at all, personas have typically been created using data from surveys or focus groups (Nielsen, 2019c; Pruitt & Adlin, 2006), so user data was usually collected and analyzed manually to create the personas (Pruitt & Adlin, 2006). From their inception in the early 1990s (Cooper, 2004), persona profiles have generally been “flat” (i.e., paper or electronic documents such as PDFs) (Gao & Gao, 2019), encompassed within a one- or two-page summary (Nielsen & Hansen, 2014). These flat persona profiles (see Figure 1 as an example) generally contain a collection of attributes about the persona, such as name, age, demographic attributes, a picture, a quote, and so on.

As such, from their initial creation through several decades of use, persona profiles were mainly data structures, i.e., static frameworks or representation of user...
attributes (Nielsen, 2019b; Pichler, 2012) with which the stakeholders can only have limited interaction. A persona profile was a means to organize the data about the imaginary person to present information that end users of the personas could then use for their design goals of better understanding users.

The proliferation of online web analytics (Stamatelatos, Gytopoulos, Drosatos, & Efraidmidis, 2020), social media analytics (Ma, Mao, Ba, & Li, 2020), and system analytics (Ait Hammou, Ait Lahcen, & Mouline, 2020) makes it possible to create data-driven personas using computational algorithms (An, Kwak, Jung, Salminen, An, Kwak, & Jansen, 2018a) that represent the underlying data (Stevenson & Mattson, 2019). Instead of flat documents, these data-driven personas can become part of an integrated (Abouelmaged & Mouakket, 2020) full-stack analytics system containing data about the users in a multiple information layer hierarchy — from backend user data to probabilistic analytics to frontend persona conceptualization. As the persona profile is interactively linked to the underlying segment and individual user data, the persona is changed from a flat data structure document to an interactive interface to an analytics and data system. For such an integrated system, the persona profile:

1. functions as an interface (see Figure 2) to the system for conceptually understanding the user,
2. provides a data structure for empathic user understanding of the analytics data, and
3. is interactive for the end user of the system interested in learning about the persona, segment, or user.

In this conceptual research, we present an implementation of such a full-stack persona analytics system, Automatic Persona Generation (APG) (Jung et al., 2017), showing the capabilities and functions at three levels of granularity:

1. conceptualization (i.e., the personas),
2. analytics (i.e., percentages, probabilities, and weights), and
3. data (i.e., the foundational user data).

Such data-driven, algorithmically-derived persona systems transform the persona creation process, how persona profiles can be used (Johannsen, 2018) as an interface, and the notion of personas as design tools. As such, personas as interfaces provide both theoretical and practical implications.

## 2 RELATED WORK

Long known in human computer interaction (HCI), personas are used across many fields and domains (Alghofaili, 2018; Cooper, 2004; Nielsen & Hansen, 2014; Pruitt & Adlin, 2006), including software design (Anvari & Richards, 2018; Marshall et al., 2015), content creation (Nielsen et al., 2017), marketing (Alekseeva, Stroganova, & Vasilenok, 2019;
Personas have several claimed benefits (Drego, Dorsey, Burns, & Catino, 2010; Eriksson, Artman, & Swartling, 2013; Friess, 2012; Miaskiewicz & Kozar, 2011; Miaskiewicz & Luxmoore, 2017; Rönkkö, 2005), such as keeping the design focus on users (Cooper, 2004), providing an empathetic user representation (Miaskiewicz & Kozar, 2011; Norman, 2004) that most people can relate to (i.e., another person), and facilitating communication among team members about users (Pruitt & Adlin, 2006). As such, personas are central to research and practice in HCI, computing science, information science, and other fields focused on user understanding (Nielsen, 2019a).

An issue is that, since their inception, personas have typically been served through non-interactive flat media, either actual paper or PDF-like electronic documents, serving as mainly a data format (e.g., framework, representation, structure) to present assorted information for decision-makers. These flat files provide little means for stakeholders to interact with the persona profiles beyond processing the information presented (Szcziuka & Krämer, 2019) and communicating that information to others. Partly because of this, personas have come under substantial critique (Chapman & Russell, 2006) for:

1. being of little practical value in the actual design process,
2. lacking actual employment between the personas and the various stakeholders, and
3. providing little actionable information from which stakeholders can directly operationalize.

The concern about the value of personas is even more pressing because a range of online analytics tools, services, and platforms have emerged (Cooper, 2004; Jung, Salminen, An, Kwak, Jung, Salminen, & Jansen, 2018a; Springer & Whittaker, 2019) since personas were first proposed - e.g., Facebook Insights, Google Analytics, IBM Analytics, YouTube Analytics - as well as integrating services (e.g., SocialBaker, SocialFlow, Tableau) that businesses can use to understand their users and user segments. These services provide platform stakeholders access to both individual (Ricotta & Costabile, 2007) and aggregated big data (Del Vecchio, Pasquale, Ndou, & Secundo, 2017; Oliveira, Rivero, Nunes de Oliveira Neto, Santos, & Viana, 2018) about users, raising questions about the value of using traditional personas for user insights.

However, the concern with these analytics systems is threefold:

1. Their efficient use requires a level of analytical sophistication that not all end users (e.g., journalists, nurses, etc.) might have.
2. The “raw and cold” numbers and tables afforded by the analytics systems are not ideal for creating a sense of empathy towards the users that these numbers represent.
3. There is often little processing conducted with the data in these analytics platforms beyond aggregate reporting.

Numbers also lack explanations as to who the users are as people, and why they behave as they do. In contrast, personas typically do include such information. So, analytics systems may encourage increase personalization without really understanding or empathizing with the users’ underlying goals, focus, or pain points. Instead, analytics may lead to personalizing on the incorrect, minor, or flawed based only on prior user interactions recorded in the analytics data.

One obvious solution is the combining of personas and analytics, where the strengths of each approach help counterbalance the deficiencies of the other. Conceptually, personas are easy for people to understand and generate empathy for the user, but personas are perceived as not granular and not actionable. Analytics data can be granular and actionable (Misuraca, Scepi, & Spano, 2020), but it can also be cumbersome for employment and confusing for end users to comprehend. The combination of both personas and analytics into a persona analytics system leverages the strengths and offsets the weakness of each (Salminen, Jung, Chowdhury, Sengün, & Jansen, 2020).

Personas that are automatically generated from analytics data have all the power of standard personas, and, when serving as the interface to analytics systems, they provide all of the strengths of user analytics, as illustrated in Figure 3.

Combining personas with analytics makes the user data less challenging to use with the additional benefit of presenting an empathetic understanding of users via the representation of another person in the persona. Other advantages of this persona as interface approach is that, relative to traditionally crafted personas that are manually created and typically include 3-7 personas per set (Hong, Bohemia, Neubauer, & Santamaria, 2018), one can create hundreds of such data-driven personas to reflect the different behavioral and demographic nuances in the underlying user population (Salminen, Jansen, An, Kwak, & Jung, 2018a) and segments. Although theoretical achievable and with some prior work in the area
(Zhang, Brown, & Shankar, 2016), there are limited working and deployed persona analytics systems. Therefore, the research question arises: *Can we create a persona analytics system that enables stakeholders to interact with both the personas and the quantitative data that the personas are based, while still retaining the empathetic experience of persona profiles?*

There has been some prior work concerning making personas interactive (Chu, Vijayaraghavan, & Roy, 2018; Li et al., 2016), but our focus is also integrating personas with underlying data.

3 | RESEARCH OBJECTIVE

Towards addressing this research question, we present the APG implementation of a full-stack data-driven persona system, highlighting the multiple levels of data access afforded by a “persona as interface” implementation to an integrated persona analytics system.

We believe that such a full stack system offers the empathy of personas and the rationality of analytics (Jansen, Salminen, & Jung, 2020).

4 | APG SYSTEM OVERVIEW

APG generates a set of persona profiles (see the left-hand side listing of personas in Figure 4) representing the user population segments, with each segment having a complete persona profile (see displayed persona profile in Figure 4). Data-driven personas, relying on regular data collection intervals, can enrich the traditional persona profile with additional elements such as (a) user loyalty, (b) sentiment analysis (Tahara, Ikeda, & Hoashi, 2019), and (c) topics of interest, which are features requested by the stakeholders. Also, the persona profile is interactive, functioning as the interface to the additional level of data, as discussed in the next subsections.

We first provide a brief overview of the APG persona system, describing how data for the profiles are retrieved, how the data is processed, and how the persona profiles are created. The system generates personas from quantitative user data via the steps shown in Table 1.

Leveraging intelligence system design concepts (Tang et al., 2019), APG identifies unique behavioral patterns of user interactions with products (i.e., these can be products, services, content, interface features, etc.) and then associates these unique patterns to demographic groups based on the strength of association to the unique pattern. After obtaining a grouped interaction matrix, we apply NMF for identifying latent user interaction (see Table 1 and Figure 5). NMF is particularly intended for reducing the dimensionality of large datasets by discerning latent factors (Lee & Seung, 1999). Other decomposition techniques, such as principal component analysis (PCA), vector quantization (VQ) were considered but were not appropriate to the dataset. The system enriches the user segments produced by NMF via adding an appropriate name, picture, social media quotes, and related demographic attributes (e.g., marital status, educational level, occupation, etc.) via querying the Facebook Marketing API.

The APG personas are displayed to stakeholders (i.e., the people from the organization whose data is used for the persona generation) via the interactive online system running on Flask, an open-source Python web framework. APG is a fully functional system deployed with real client organizations, with a demo available online. A detailed technical explanation of APG’s system infrastructure is found in Jung, Salminen, Kwak, An, and Jansen (2018b); Jung, Salminen, Kwak, et al., 2018b). A detailed description of APG’s algorithmic method can be found in An, Kwak, Jung, et al. (2018a). Here, we focus on the previously undocumented features for adding interactivity in APG-generated personas.
The system employs the foundational user-data, which the system algorithms act upon, transforming this data into information about users. The outcome of this algorithmic processing is actionable metrics and measures about the user population (i.e., percentages, probabilities, weights, etc.) of the type that one would typically see in industry-standard analytics packages. Employing these actionable metrics is the next level of abstraction taken by the system is to leverage this analytics data, along with corresponding meta-tagged information, such as photos and names, to create sets of personas profiles at the conceptual level. The result is a data-driven persona system capable of presenting user insights at different levels of granularity, with levels both integrated and appropriate to the task (see Figure 6).

We now discuss these three levels, illustrated in Figure 6.

### 5.1 Conceptual level: personas

The highest level of abstraction is the set of personas that APG generates from the data using the method described above, with a default of 10 personas, although the APG system theoretically generates as many personas as needed. The persona profile has nearly all the typical attributes that one finds in traditional flat file persona profiles.

However, as in APG, personas as interfaces allow for dramatically increased interactivity in leveraging personas within organizations. Interactivity is provided such that the end user can alter this number to generate more or fewer personas, with the system currently set for between 5 and 15 personas. The system can allow for searching of a set of personas (see Figure 7 and figure 7a) or leveraging analytics for prediction of persona interests (see Figure 8 and Figure 8a).

### 5.2 Analytics level: percentages, probabilities, and weights

APG persona profiles act as interfaces to the underlying information and data used to create the persona profiles. The specific information may vary somewhat by the data source. Still, the analytics level will reflect the particular metrics and measures generated from the foundational

### Table 1

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create an interaction matrix ((V)) with products as columns ((c)), demographics as rows ((g)), and the interaction count as elements of the matrix</td>
<td>Apply non-negative matrix factorization (NMF) ((\text{Lee \&amp; Seung, 1999})) to the interaction matrix to discern (p) latent product interaction behaviors (where (p) is a predetermined hyper-parameter indicating the number of personas).</td>
<td>Choose the corresponding representative demographic attributes for each behavior using NMF weights.</td>
<td>Enhance user segments with name, picture, topics of interest, etc. to create personas.</td>
</tr>
</tbody>
</table>

*FIGURE 5*  Matrix decomposition using NMF. Matrix \(V\) is decomposed into \(W\) and \(H\). \(g\) denotes demographic groups in the dataset, \(c\) denotes product units, and \(p\) is the number of latent behaviors of demographic groups over product units, and \(\varepsilon\) is the error term.

*FIGURE 6*  APG full stack data integration, from user-level to analytics-level to conceptual-level.

<table>
<thead>
<tr>
<th>Table 1 Steps of data-driven persona creation in APG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
</tr>
<tr>
<td>Create an interaction matrix ((V)) with products as columns ((c)), demographics as rows ((g)), and the interaction count as elements of the matrix</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level</th>
<th>Classification</th>
<th>Artifacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual (the personas)</td>
<td>Abstracted</td>
<td>Persona as interface</td>
</tr>
<tr>
<td>Analytics (the metrics)</td>
<td>Processed</td>
<td>Metrics and ratios</td>
</tr>
<tr>
<td>User (the numbers)</td>
<td>Data</td>
<td>Raw values and figures</td>
</tr>
</tbody>
</table>
user data and used to create the personas. In the APG system, the personas profile provides affordance to the various analytics information via clickable icons on the persona interface. We highlight three examples.

As shown in Figure 9 and Figure 9a for audience size (i.e., number of people in the user segment), APG displays the percentage of the entire user population that a particular persona is representing (calculated from Facebook Marketing API data). This analytic insight is valuable for decision-makers to determine the importance of designing or developing for a particular persona and also helps address the issue of the validity of the persona in representing actual users (Chapman, Love, Milham, ElRif, & Alford, 2008).

As shown in Figure 9 and Figure 9a for audience size (i.e., number of people in the user segment), APG displays the percentage of the entire user population that a particular persona is representing (calculated from Facebook Marketing API data). This analytic insight is valuable for decision-makers to determine the importance of designing or developing for a particular persona and also helps address the issue of the validity of the persona in representing actual users (Chapman, Love, Milham, ElRif, & Alford, 2008).

APG gathers general demographic information from the online platforms, which is typically gender, age grouping, and nationality. Using these attributes and results derived from NMF, APG then leverages the Facebook Marketing API to determine the probability of other demographic characteristics. The persona profiles provide access to these probabilities (see Figure 10 and Figure 10a for the probability of relationship status).

As mentioned above, APG identifies unique behavioral patterns and then associates these unique patterns via latent factors to one or more demographic groups, assigning a weight to each of these demographic groups based on the strength of its association with the unique pattern.

These weights provide insights to the end user of the personas concerning both (a) the strength of behavior and (b) the association of this behavior pattern with other groups. Therefore, APG provides access to these demographic group weights (Figure 11 and Figure 11a), addressing concerns of what actual customers the personas represent (Chapman et al., 2008).

5.3 User level: individual data

Leveraging the demographic meta-data output from the NMF, the decision maker can access the specific
customer level (i.e., individual or aggregate), as shown in Figure 12. It is the numerical user data (in various forms) that are the foundation of the personas and analytics presented in prior figures.

6 | DISCUSSION AND IMPLICATIONS

The conceptual shift of personas from flat data structures to personas as interfaces for understanding users opens new possibilities for interaction between end users and personas and also end users and analytics. Using data-driven personas embedded as the interfaces to analytics systems, decision makers can, for example, imbue analysis systems with the benefit of personas to form a psychological bond, via empathy, between stakeholders and user data.

6.1 | Discussion of use cases

There are several validated use cases for personas as an interface to an analytics system, showing their possible employment or advantage relative to solely personas or solely analytics (Jansen, Jung, & Salminen, 2019a, 2019b; Jansen, Jung, Salminen, An, & Kwak, 2017; Jung, Salminen, & Jansen, 2019; Salminen et al., 2020) and the advantages of such as system as validated in a user study (Salminen et al., 2020). We present three additional use cases here, which are:

6.1.1 | Segmenting a user population

Identification of significant user segments has considerable applicability in a variety of domains, including system design, marketing, advertising, and content management. In prior research (Jansen et al., 2017) leveraging the APG approach, we develop a technique for determining both information dissemination and discrimination of online content to identify user segments within a population. The method identifies the most impactful content explicitly and is based on 4,320 online videos with more than 58 million interactions from hundreds of thousands of users. The disclosed approach isolates the crucial content for identifying user segments, with results showing that about 25 percent of the content is so widely disseminated (i.e., viewed by so many segments) that they are non-discriminatory. Conversely, about 30 percent of the content is very discriminatory but with marginal impact due to a small user base. As the conceptual basis for APG, the implications are that there are techniques to identify distinct and impactful content for user segmentation (Yan, Liang, & Li, 2020) in an information domain. The approach has been implemented in APG to identify critical cut-off values for data-driven persona generation (Jansen et al., 2017).

6.1.2 | Tracking change in a user population over time

One of the criticisms of personas is that the primary data on which they are created rapidly stales, which requires another data collection round. However, with personas creating manually, there was no empirical evidence for this critique, as the data replication via manual methods is extremely difficult. In previous research with APG (Jung et al., 2019), over a 2 year period, we collect monthly demographic data for a large online content publisher. Using this data, APG generates 15 personas monthly using the identical APG algorithmic approach described above. Then, we compare the personas month-over-month, year-over-year, and over the whole two-year period. The results show that there is a nearly 19% average change in monthly, a more than 23% yearly change, and a 47% change during the entire period. These results support the critique that personas do stale over time and that the staling can occur in a relatively short period. The research (Jung et al., 2019) would most likely not have been implementable with a data-driven persona system, such as APG.

6.1.3 | Tracking changes in topical interests

In research aimed at understanding changes in user’s information interests (Jansen et al., 2019a), we collect monthly content consumption and demographic data from a video platform over 2 years for a large media publisher. Using the APG methodology, we generate 15 personas each month. For each persona, we rank interests in
15 content topics, explicitly identifying the top three monthly topics for each of the personas each month. Then, we compare the top topical interest sets of the personas month-over-month during the entire two-year period. The research results show that there is a higher than 20% average change in the topical interests, and 68% of the personas exhibit more topical differences than topical consistency. The research results indicate the topical interests of users of online content are fluid, indicating a need for continual persona updating using algorithmic data-driven methods. The implications for organizations seeking to understand the information interests of their users is to employ data analysis to detect changes in the user interests via an APG-like automated persona generation process (Jansen et al., 2019a).

6.2 | Theoretical implications

The approach of combining personas and analytics into an integrated persona analytics system has several theoretical implications. First, it addresses a conceptual criticism of personas that they are not actionable (Chapman & Russell, 2006) at the individual user level. With personas as the interface to an analytics system, a given persona can be directly associated with individual users if the foundational user data was used to create that persona. Second, the integration helps offset the issue with a reliance on numbers for solely personalization without conceptually understanding user goals, pain points, and underlying needs (Nielsen, 2019c). Third, the combination of personas and analytics makes possible the generation of tens to thousands to millions of personas for a given user population (Jansen et al., 2019b), increasing the available representative granularity of the personas and more easily tying the personas to the underlying user data.

6.3 | Practical implications

The practical implications for managers and practitioners of combining personas and analytics into an integrated system are derived from the direct linkage from conceptualization with personas to implementation with the actual user data. First, personas are now actionable, as the personas accurately reflect the data (Sim et al., 2019). As an illustrative scenario, executives can make user decisions based on a set of personas. Ease of communication is one of the hallmarks of personas. Operational managers (Cai, Wang, & Gong, 2020) can, using the analytics aspect of the system, isolate the specific user segment a given persona represents. Finally, operational designers, information managers, or marketers can identify the individual users represented by the given persona and contained within the market segment. This full-stack implementation aspect has not been available with either personas or analytics previously.

7 | FUTURE WORK AND CONCLUSION

Personas from online analytics data can be used towards “giving faces to data” as the frontend interface to interactive systems for understanding users. To this end, there are ample research opportunities, including increasing levels of data granularity and user studies concerning the deployment of such systems and persona interactivity (Eiband et al., 2018). Optimal persona UIs and features are another central targets for future research.

The features and functionalities presented here are based on discussions with real users throughout several user studies (Salminen et al., 2018b; Salminen et al., 2019). However, the actual interaction between the users and these experimental interaction features needs to be empirically addressed, calling for future work in the domains of HCI, computing science, information science, and other related fields focused on user understanding.

Overall, the capabilities and conceptual enhancement of the personas as an interface to full-stack analytics systems (Springer & Whittaker, 2019) presents much promise for employing data-driven persona, for the leveraging of user information, and impacting decision making in organizations.

REFERENCES


behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, 8(1).


Cooper, A. (2004). *The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity* (2nd ed.). Pearson Higher Education.


Stamatelatos, G., Gyftopoulos, S., Drosatos, G., & Efraimidis, P. S. (2020). Revealing the political affinity of online entities through their twitter followers. Information Processing & Management, 57(2), 102172.


American Society for Information Science and Technology, 44 (1), 1–6.

**How to cite this article:** Jansen BJ, Jung S-G, Salminen J. From flat file to interface: Synthesis of personas and analytics for enhanced user understanding. *Proc Assoc Inf Sci Technol. 2020;57: e215. [https://doi.org/10.1002/pra2.215](https://doi.org/10.1002/pra2.215)