# Privacy Preserving Personalization via Dataware

James Goulding
Horizon Digital Economy Research
University of Nottingham
Nottingham NG7 2TU, UK
+44 (0) 115 823 2557
james.goulding@nottingham.ac.uk

Richard Mortier
Horizon Digital Economy Research
University of Nottingham
Nottingham NG7 2TU, UK
+44 (0) 115 823 2552
richard.mortier@nottingham.ac.uk

Derek McAuley
Horizon Digital Economy Research
University of Nottingham
Nottingham NG7 2TU, UK
+44 (0) 115 846 6896
derek.mcauley@nottingham.ac.uk

## ABSTRACT
Nearly every major company within the Digital Economy is engaging with personalization, and for good reason: it adds value to user experience whilst giving companies greater control over price discrimination. However, personalization quickly distils into a task of monitoring and tracking, raising serious issues of control and privacy. And with each service creating its own fragmented, incomplete model, users cannot aggregate personalization across them. Dataware provides a solution for both of these issues. Its mechanisms allow: 1. users to collate behavioural data to create a model under their control, and 2. third parties to leverage those models while maintaining a user defined privacy boundary.

## Categories and Subject Descriptors
D.0 [**General**] C.2.4 [**Distributed Systems**] K.4.1 [**Privacy**]

## General Terms
Design, Management, Experimentation

## Keywords
Personalization, Privacy, SOA, Personal data, Cloud computing

## 1. INTRODUCTION

The digital economy is characterized by a growing capability to personalize products and services to individual consumers. Virtually every major company on the Internet is engaging with such personalization in some manner - Google customize their search results to an individuals location and search patterns; Amazon their product recommendations; Facebook their adverts; and Yahoo their news delivery. With organizations rapidly adopting techniques that deliver customized and engaging interactions to their audiences, personalization is becoming big business. And for good reason: it can add value to user experience while simultaneously allowing companies greater control over price discrimination and yield management. Everybody wins.

Or do they? To work, personalization requires inferences to be made based upon your personal information. The greater the level of customization that is offered, the more personal information that the service provider requires – and ultimately the greater the privacy sacrifice to you the user [1]. Personalization thus quickly distils into the task of monitoring, recording, and consequently tracking customer behaviour. And as consumers have increasingly become "more concerned about personalization in customized browsing experiences, monitored purchasing patterns, and targeted marketing and research" over the last decade [2], so such processes have drawn the spotlight of the media, and even attention at the level of government policy.

If anything, the consumer would benefit from service providers sharing their data. There are lots of models of "you" out there, but they are incomplete fragments – individual services, confined within walled gardens, can only construct models that are partial and skewed. Used collaboratively their data would produce models of much greater power, formulating a far richer picture of a user's global preferences. In turn this would enable more robust personalization, an improved user experience and ameliorate such issues as the *Quilting Problem*[1]. However, the transmission of personal information of this nature between companies would represent an extremely serious violation of customer privacy. The untenability of this situation means that each individual service is forced to harvest their own incomplete, and consequently, inaccurate personalization models.

In this work, we offer a potential solution through implementation of the *Dataware* framework [3]. This allows users to collate both explicit and implicit behavioural data in order to form a preference model under their control. It further provides the means by which third parties can query such models, generating aggregate response without crossing the privacy boundaries a user has defined to prevent access to fine-grained personal data.

## 2. Aggregated User Models

The isolation of user models within individual services generates the following issues to the end-user:

**Privacy:** how can I trust that the personal data a service is collecting about me is not being used without my permission, or for purposes that I find objectionable?

**Walled Gardens:** Why is every service I use forced to generate a personalization model from scratch? Surely this represents a serious barrier to entry for new competitors into a market?

[1] *The "Quilting problem" describes Amazon's endless recommendation to me of both quilting manuals and sewing paraphernalia (due to my previous purchase of such items as one-off gifts) despite my complete lack of interest in them. Errors such as these could be quickly marginalized through integration of other data, such as Google's model, for example, which could recognize from search behaviour that I have absolutely no desire to make further purchases of such items.*

**Incomplete Models:** Every service maintains an incomplete fragment of my overall preferences, resulting in increased probability of inaccurate inferences and accentuation of noise.

The solution we offer addresses all of these concerns. By putting personalization data under the user's control we obviate privacy concerns; by decentralizing such a service and exposing it to third parties we release data from existing solely within the confines of walled gardens; and by allowing the data to be combined as a post-filter over pre-existing models we allow extant services to embellish their current partial models.

## 3. Ubiquitous Personalization via Dataware

Dataware [3] is a framework for managing your data landscape. Instead of requiring centralized data storage, dataware is composed of a network of running components (*data stores*) that receive and persist personal data, but which themselves communicate auditing updates to a user's *catalog*. The catalog not only provides a user with a global picture of his data generation, but also acts as a point of contact for third-parties looking to access that data. Applications may request access to data by submitting a python query via a conservative extension of the OAuth protocol [4], referred to as the *Deferred Authorization Flow*. If a submitted query is deemed acceptable by the user, given their privacy preferences, a shared key is transmitted to both the relevant datastore and the 3rd party application, which permits execution of the query over an agreed duration.

### 3.1 Generating user-driven preference models

A user driven preference model is created via a dedicated Google Chrome plugin (illustrated in Fig.1) that transparently profiles the web pages that a user visits. The plugin generates a linguistic n-gram frequency model of each page (referred to as a distillation) which it then transmits to a user's *prefstore, a dataware store* existing within the cloud. The prefstore collates received distillations for an individual user in order to build up an overall picture of the n-gram frequencies that represents the user's web behaviour. It is then possible for a 3rd party applications to formulate queries that interrogate the resulting dataset so long as permission is granted at the user's catalog.

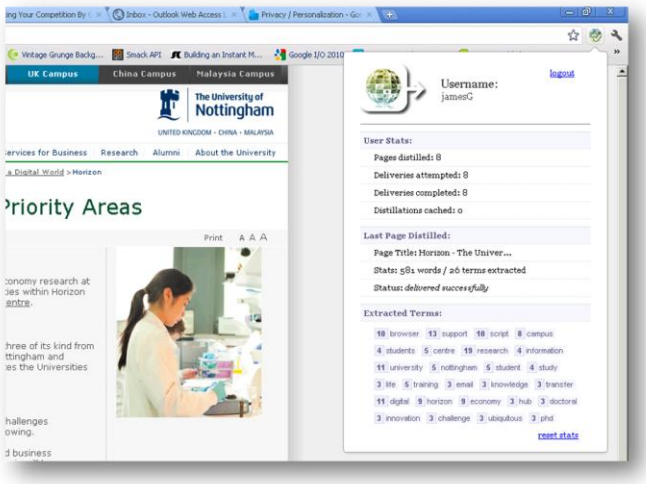


*Figure 1: Term extraction Google chrome plugin.*

### 3.2 Aggregate Query Example

In order to illustrate the power of this architecture we have implemented a sample python query that compares a set of documents to a user's prefstore, returning a *relevance score* for each relative to the user's preference model. Vitally, such a query only outputs an aggregate similarity score for each document, and never exposes the user's fine-grained web-use data. In order to generate such scores, the query harnesses a traditional linguistic weighting scheme, combined with a cosine similarity comparison, to determine the similarity between a document, *d*, and the user's preference model, *p*, both encoded as term vectors:

$$similarity(d,p) = \frac{d.p}{\|d\|\|p\|} = \frac{\sum_{i=1}^{N} w_{i,d} w_{i,p}}{\sqrt{\sum_{i=1}^{N} w_{i,d}^{\ 2}} \sqrt{\sum_{i=1}^{N} w_{i,p}^{\ 2}}}$$

where *w* is the *term frequency-inverse document* weight of an individual n-gram. Via invariance checks as to the composition of input vectors, and enforcing limits on the total number of query submissions, it is possible to deliver compelling new services while maintaining acceptable privacy levels and eliminating data leakage.

### 3.3 Conclusions and Possibilities

The key to dataware services such as the prefstore, is that they allow consumers to dynamically combine data they collate with data held about them by businesses and government. Amazon's models, and the business value they represent, are not be replaced by such services but rather augmented by them, augmenting and not supplanting user experiences. Projecting against a user's preference store gives third parties the ability to improve search results, to customize news services, to filter shopping recommendations or to generate appropriate location based advertising. However, it also allows users to experience personalization in a ubiquitous manner through the consistent application of an independent, external preference model across different services.

## 4. ACKNOWLEDGMENTS

The research on which this paper is based was funded by the Horizon Digital Economy Research, RCUK grant EP/G065802/1.

## 5. REFERENCES

[1] Chellappa, R. K; Sin, R. G.; , "Personalization versus Privacy: An Empirical Examination of the Online Consumer's Dilemma", Information Technology and Management, Volume 6, No. 2-3, 181-202, 2005, DOI= http://doi.acm.org/10.1007/s10799-005-5879-y

[2] Anton, A.I.; Earp, J.B.; Young, J.D.; , "How Internet Users' Privacy Concerns Have Evolved since 2002," Security & Privacy, IEEE , vol.8, no.1, pp.21-27, Jan.-Feb. 2010, DOI= http://doi.acm.org/10.1109/MSP.2010.38

[3] McAuley, D.; Mortier, R.; Goulding, J.; , "The Dataware manifesto," Communication Systems and Networks (COMSNETS), 2011 Third International Conference on , vol., no., pp.1-6, 4-8 Jan. 2011, DOI= http://doi.acm.org/10.1109/COMSNETS.2011.5716491

[4] E. Hammer-Lahav (ed.). The OAuth 2.0 Authorization Protocol draft-ietf-oauth-v2-20. IETF RFC 5849, July 25 2011. http://tools.ietf.org/html/draft-ietf-oauth-v2-20