

Learned Publishing, 23:–
doi:

Customer data . . . everywhere
A publisher's data about its customers and users typically comes in many forms: subscription records for print and electronic products, society member lists, author and reviewer data, table of contents alert recipients, pay-per-view purchasers, registrants for access to online services and trials, conference attendee lists, institutional usage reports etc. For many publishing functions, but particularly for sales and marketing staff, this data represents a great opportunity as there is considerable value and insight to be gained from analysing it all closely.

However, customer data is often distributed across a number of different, specialized repositories and systems, with some data sources managed in-house and others maintained externally by third-party providers. As a consequence, staff with an interest in getting a complete view of the organization's relationships with its customers may find it difficult to do so.

Furthermore, publisher customer data has some intrinsic complications. Firstly, there is the challenge of identity: how to uniquely identify each customer, both within and between data sources. The same customer can appear in many different places. This is not as easy as it might sound: personal names and addresses are not always entered consistently by individuals and users can have multiple email addresses or use shared email accounts. In the case of institutions, the lack of consistent organization names and identifiers in the supply chain is a familiar issue with numerous consequences (not just data integration headaches).¹ A situation that can all too easily arise is one of customers with multiple identities in multiple systems.

Secondly, and related to the question of who is who, when integrating this data con-

All together now: integrating customer data with MasterVision

James CULLING

DataSalon Ltd

ABSTRACT. A publisher's customer and user data typically comes in many forms and from many specialized repositories and systems. For sales and marketing staff in particular, there is considerable insight to be gained from analysing all of this data closely. However, achieving an integrated and complete 'single customer view' via a data warehouse solution can be a costly, time-consuming and challenging undertaking. In this case study report, DataSalon's MasterVision is described as an alternative approach to solving this problem for publishers. Key to the solution is a dynamic data modelling philosophy, with both the underlying database and product interface configured for each publisher around the specific customer data supplied. A number of real-life 'use cases' are explained and the value of Ringgold's Identify database of organizations, in addressing a number of integration challenges but also in adding to the intelligence that can be gleaned from source data, is discussed.



James Culling

© James Culling 2010

sideration must also be given to whether, and how, customers are related to one other through organisational affiliation. In the scholarly publishing space, we can think of three principal customer 'units': individuals, organizations and consortia. In many industry sectors of relevance to publishers, e.g. academic, corporate, healthcare, and government, a complex hierarchy of 'parent' and 'child' relationships can exist. For example, an individual author can be affiliated to an academic department (an organization) that is part of a university (another organization) that is in turn part of a regional consortium, etc. These relationships really need to be understood and made apparent when integrating customer data if the solution is to be of value.

A common approach to bringing customer data together to achieve a complete view in one place is to specify and build an all-encompassing data warehouse.² This solution looks to extract, clean, and standardize data from the various source systems so that

it fits into a master data model that is designed upfront and is structured in such a way as to address business reporting needs. This can, however, be a time-consuming and costly undertaking for a publishing organization. With the rapid pace of change in our industry, there is also a risk that today's data warehouse solution may find it hard to adapt to tomorrow's needs. Furthermore, there is a danger that this approach will struggle with the complexity of customer relationships in the market sectors of relevance to scholarly publishers.

A different approach

DataSalon has developed its MasterVision solution to try to address the challenges outlined above in a flexible way for publishers. MasterVision is a web-based product, designed to be deployed many times over for different publisher clients, and operating as a 'read-only' layer that sits on top of each publisher's various customer data sources.

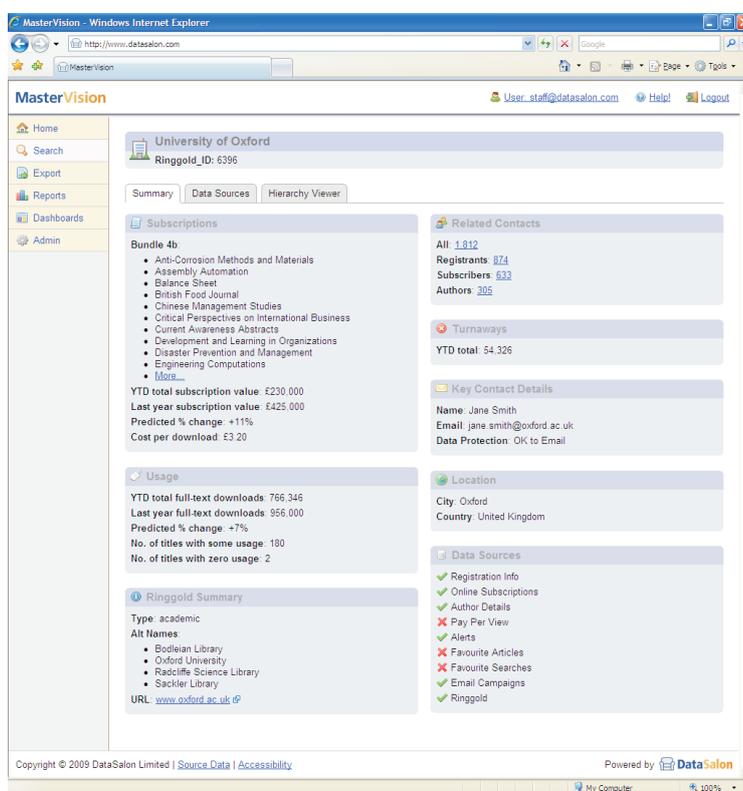


Figure 1. An example 'Summary' display from a consolidated customer profile (of course, fictitious details are given for the customer named here).

The distributed data issue is tackled by taking regular feeds from source systems, normalizing file formats for upload into the product, and with support for data cleansing, validation, and enhancement routines as necessary for each publisher prior to loading. Key to the approach is dynamic data modelling. Rather than spending lots of time designing and specifying a master data model for every implementation, or requiring all publishers to use a DataSalon-designed data model, the product effectively acts as an 'empty shell', and builds itself (both at the database and interface level) around the data it is given. In essence, this means that every source field and value from each system is available for search, display and report functions.

Putting it all together

One of MasterVision's central features is a consolidated customer profile screen – the 'single customer view'. A tabbed display is presented, and the first of these tabs is a tailored summary of key fields for a given customer (Figure 1). These key fields or customer 'headlines' are drawn from across all data sources, and they can be raw metadata items or inferred fields created via calculations undertaken by MasterVision 'behind the scenes'. For an organizational customer, a publisher might choose to present a summary of current subscriptions, revenue/value statistics (taking all current subscriptions together), usage figures and trends, related individual contacts by role, useful identifiers drawn from the publisher's source systems, and geographic location information. Another publisher might wish to see some of the same fields but to also have others, unique to its customer data, highlighted in this view.

A second tab, 'Data Sources', houses expandable and collapsible tables for each source where data is known about a customer. Different and previously unconnected publisher databases are presented together on the one screen, with the various tables and fields within them representing the data as supplied from the original source systems.

The essential point to achieving successful de-duplication of customers and linking of

data sources to obtain a single view like this is a flexible integration engine – one that is capable of automatically piecing together user data using a combination of identifiers and techniques. For example, email address, subscriber/member identifiers, personal name and address details, organization names and organization identifiers might all be used. A degree of 'fuzzy matching' with some of these values can also be necessary to achieve results.

RSC: Usage trend analysis

The Royal Society of Chemistry (RSC)³ is using MasterVision to integrate a large number of its data sources for customer insight and prospect identification purposes. Datasets include subscription records, author details, pay-per-view purchasers, alerts recipients, conference attendees, and institutional usage reports. As part of its ongoing customer profiling work, an automated analysis of the usage information supplied is carried out by MasterVision at the time of each system refresh. Percentage change trends in average full-text downloads are derived and made available for display in each customer's consolidated profile. Trend searches can also be executed by users to look for customers whose usage is growing or shrinking over time.

JBJS (American): cost per download metric

The *Journal of Bone and Joint Surgery (American)*⁴ also utilizes MasterVision to bring a number of its customer data sources together. For its 2010 subscription renewal email campaign, this publisher sought to emphasize the value of the online subscription to each customer by including a cost per download metric as part of the message. The integrated data view helped to achieve this, by making it possible to pool and analyse data from three distinct data sources: online activations data (from HighWire), subscriptions records (from an internally managed system), and usage data (also from HighWire). While used as part of a renewal campaign message in this case, a cost per download metric is a good example of another inferred 'headline' field that can be

Figure 2. The tabbed, form-based approach to searching for customer and prospect segments.

included as part of a consolidated customer profile should a publisher so wish.

Segmentation via search

As well as consolidating data about customers into a single view, an integration solution needs to address the requirement for searching across the pooled data to identify segments of interest – typically for editorial, sales, and marketing purposes. In MasterVision, a tabbed, form-based search interface is provided. This interface is, as with the customer profile, driven by a publisher's own data sources and fields. Typically this means there is a tab for each data source supplied, and the fields on each tab always represent the fields and available values coming from that source.

The concept is that a user can 'tab around' and fill in as many or as few fields as they want to identify a particular contact or customer segment of interest (Figure 2). The search engine is tasked with finding all of the customers that match all of the properties specified.

This approach to searching can be used to identify different types of overlaps (and

non-overlaps) between datasets, and this is used to drive the identification of segments to target for up-sell and cross-sell marketing campaigns, author and reviewer announcements, society member communications, etc.

Reporting tools are also available, enabling a user to 'dig' into a search results set further, to break down and analyse the results in more detail, and to explore data visually. For example, the customers from a search query can be plotted on a world map or a pie chart can be produced to analyse a specific facet of the segment identified. These charts are 'clickable', enabling the user to drill into the customer group from a given country or pie segment. Furthermore, charts and data tables derived from search results analysis can be saved for sharing with other users or configured into a one-page dashboard display for later reference (Figure 3).

To feed contact data from searches into an email marketing campaign tool, an export interface is provided that enables the user to choose fields from across the pooled dataset, effectively designing the columns of a

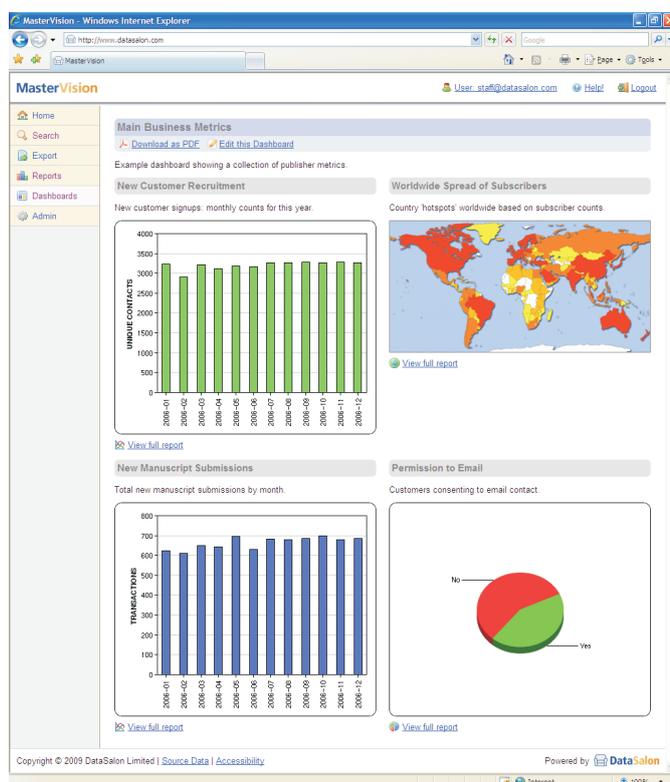


Figure 3. Charting and mapping of identified customer segments, presented here as a one-page management dashboard report.

spreadsheet. This file is generated on-demand by the system as a comma-separated values (CSV) file, a file format that is simple but effective when it comes to interchange with other systems.

Is MasterVision a customer relationship management (CRM)⁵ system? Certainly, the product addresses a number of business needs in this respect, chiefly the analytic and sales intelligence aspects of CRM. It is not, however, designed as a solution for the direct input of data regarding interactions and communications with customers or as an email marketing campaign tool. There are many products that address these other CRM needs. Rather than trying to compete in these spaces, MasterVision is focused on addressing the complex customer data integration challenges facing publishers, with support for uploading data from customer contact history and marketing campaign tools alongside other publisher data sources, to further enhance the single customer view provided.

AMA (JAMA & Archives Journals): prospecting with pay-per-view data

The JAMA & Archives Journals Division of the American Medical Association (AMA)⁶ is using its combined view onto customer data in a number of ways. One of these is to identify individual subscription up-sell prospects from repeat pay-per-view transactions. MasterVision's tabbed search interface is used to identify non-subscribing contacts that meet a number of relevant criteria, with the data then exported to the organization's campaign management tool for the sending of a targeted email message regarding the value of taking out a subscription.

A point of reference

From a customer data integration perspective, a central reference point for the identification and naming of organizations and consortia, and for defining the relationships between them, is a valuable asset. The Ringgold Identify database⁷ can be used as

such a reference point in a number of different ways with MasterVision.

For a publisher with Ringgold-audited links to Identify records in its customer data, these links (essentially a pairing of the publisher's identifier for each organization with Identify's own identifier for the same) can be used as a de-duplication aid within and across source datasets.

For a publisher that has licensed all or some of the Identify database (such that the records from Ringgold can be included as an additional data source in MasterVision) there are further possibilities. Firstly, Identify organization names can be used to standardize the publisher's own data, effectively tidying and often enhancing the source data names loaded, and Identify metadata for each organization (e.g. organization type, size, tier, URL) can also appear in consolidated customer profiles. Secondly, the IP address data that Ringgold captures and tracks as part of its cataloguing work can not only be used as a customer de-duplication aid with source datasets but also for the analysis of 'turnaway' logs (see AIP case study below). Thirdly, catalogued web domain data can be used to infer links between individual contacts with an email address and 'parent' organizations, a solution to one aspect of the relationships challenge mentioned earlier (see Oxford Journals case study below).

AIP: prospecting with 'turnaway' data

'Turnaway' data is clearly of high value, indicating as it does a desire to get to full text content but being refused access. As Ringgold records and tracks IP addresses for the universe of organizations it catalogues in Identify, a publisher with a license for this data can use it as a resource to identify prospects from 'turnaway' or 'access refused' logs that contain unknown IP addresses.

For the American Institute of Physics (AIP),⁸ DataSalon has carried out this matching work with a 79% hit rate (i.e. successful matches to Identify records) in trials. These 'turnaway' hits are aggregated to facilitate a search feature for AIP users: to provide a means to find organizations where 'turnaways' have exceeded a certain number

in a defined time period for a particular title that the publisher is building a marketing campaign around.

Oxford Journals: institutional prospecting via affiliated individuals

Oxford Journals⁹ is using the Ringgold Identify database both as a resource for the analysis of existing customers and to prospect for new customers.

MasterVision's reporting tools, as mentioned earlier, are used to 'mine' the Identify records, to target by country, sector, tier and size, etc. In addition, the web domain data within Identify is employed to infer likely links to catalogued organizations for over half a million individual Oxford Journals contacts with email addresses.

Additional profile metrics are presented in MasterVision for each institutional customer or prospect, listing counts of related individuals by type: for example, the number of affiliated pay-per-view purchasers, alert recipients, authors, and online registrants. This information is used by the sales team to prioritize prospective new institutions on the basis of existing activity from individuals, and perhaps to show to a prospect as evidence of their users' interest in Oxford Journals' content.

Trees and branches

The Ringgold Identify database catalogues the many (and changing) 'parent' and 'child' relationships between organizational units and consortia, and this information can be very useful to publishing staff. For example, when you need to understand who exactly it is you are talking to about a possible licensing deal (and perhaps who you are not talking to).

As a further source of customer insight, 'relationship' data like this naturally belongs with the consolidated customer profile discussed earlier, and MasterVision now includes an interactive and navigable 'Hierarchy Viewer' to try to achieve exactly this (Figure 4). It is possible to scroll up and down through large organizational hierarchies, branches of hierarchy trees can be expanded or collapsed, and organizational

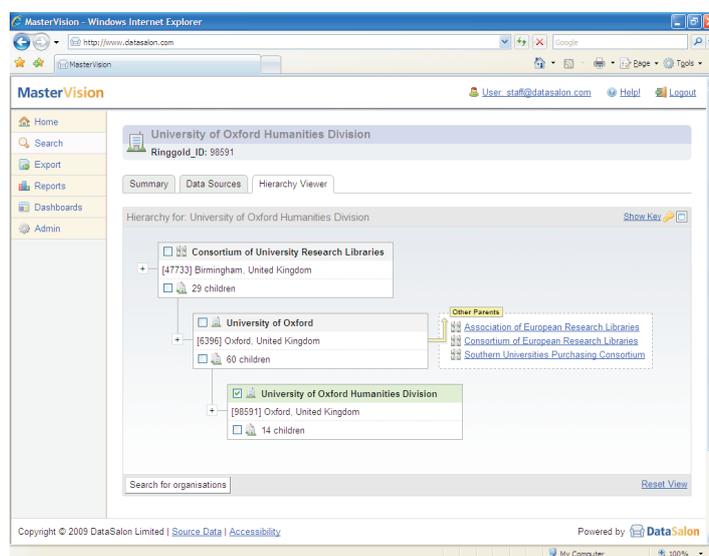


Figure 4. Using the Hierarchy Viewer to present 'relationship' data from Ringgold Identify for a given organization.

units can be selected with checkboxes for further search queries and analysis.

Conclusion

The challenges of integrating publisher customer data are many and complex. DataSalon's MasterVision is an alternative approach to bringing all of a publisher's customer information together, drawing on a dynamic data modelling technique and working closely with partners such as Ringgold to try to address these challenges and create new opportunities with customer data for publishers today.

References

1. Henderson, H.L. Why do you lose access to your electronic resources? How an institutional identifier can help. *Serials*, 21(2008): 134–139. <http://dx.doi.org/10.1629/21134>

2. http://en.wikipedia.org/wiki/Data_warehouse
3. <http://www.rsc.org/>
4. <http://www.ejbs.org/>
5. http://en.wikipedia.org/wiki/Customer_relationship_management
6. <http://www.ama-assn.org/>
7. http://www.ringgold.com/products_identify.html
8. <http://www.aip.org/>
9. <http://www.oxfordjournals.org/>

James Culling

Client Director
DataSalon Ltd.
16 Woodcroft
Kennington
Oxford OX1 5NH, UK
Email: info@datasalon.com
Web: <http://www.datasalon.com>

DataSalon Ltd provides innovative solutions for data integration and analysis to the scholarly publishing industry.