# Now, It's Personal!

# Evaluating PIM Retrieval Tools

**Daniel Gonçalves, Joaquim A. Jorge**
Instituto Superior Técnico
Universidade Técnica de Lisboa
Av. Rovisco Pais, 49
1050-001 Lisboa, Portugal
Tel: +351-213100289
{daniel.goncalves, jaj}@inesc-id.pt

**ABSTRACT**
In recent years, the area of Personal Information Management has produced several important results. Many are based in tools and applications that help users manage and retrieve their personal data. The evaluation of those tools presents some problems of its own as, since they deal with personal information, traditional evaluation approaches are, sometimes, not appropriate. This is especially true for search tools.

In this paper we briefly discuss some of the problems involved in PIM tool evaluation, using, as a case study, the methodology for the evaluation of Quill, a personal document retrieval tool.

**Author Keywords**
Narrative-based Interfaces, Document Retrieval, Personal Information Management

**ACM Classification Keywords**
H.5.2 User Interfaces: Evaluation/methodology, H.3.3 Information Search and Retrieval: Query formulation

**INTRODUCTION**
Personal Information Management is a young but quickly growing research area. Different approaches have been followed, trying to help users organize, find, and re-find their personal information. Retrieval tools in particular can be broadly classified as belonging to one of two kinds: browsing and search. In browsing tools, the users are given a way to visualize and navigate their information. However, it is often the case where browsing presupposes an existing underlying organization. Search-based approaches, on the other hand, make no such assumptions, as no structure must be navigated. Rather, by specifying search criteria, users are able to access their information regardless of where it is stored. This is important as it has been recognized even in some of the early works in the field that classifying all information (most notably, documents), is difficult, imposing such cognitive loads that some users choose not to do so at all [4][5].

While traditionally search is very simple, and primarily based on keywords, it has become increasingly evident that a more extended approach can be useful. Indeed, while keyword search might be adequate for the general case of, for instance, web search, it is poor when considering personal information. In that case, the fact that the information has been previously handled by the users provides a shared context that can be explored in search. This is especially important for the retrieval of non-textual information, for which keywords are hard to use, but also applies to other kinds of data. The autobiographic information in that context can be more relevant to users when describing the items being sought and, as such, easier to recall.

Search tools need to be evaluated to assess their usefulness. However, that evaluation poses some problems, as by dealing with personal information tradition evaluation techniques might prove difficult. In the following section, we will succinctly describe the evaluation methodology of Quill, a personal document retrieval tool, discussing how those problems were overcome.

**QUILL**
While autobiographic information about personal documents can be of help in retrieving them, this in itself does not provide us with a concrete way to do so. Using that information is not simple, especially as it can be very heterogeneous. Some way to help the users recall it in meaningful ways and convey it to the computer was required. We developed a new interaction paradigm, narrative-based interfaces, in which stories about


CHI 2008, April 5–10, 2008, Florence, Italy.

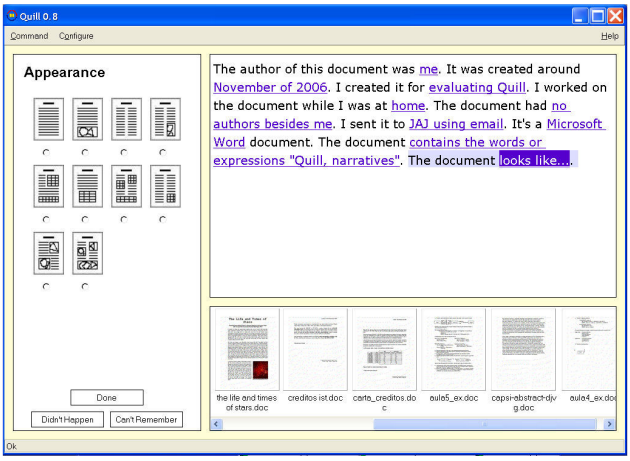


**Figure 1 – The Quill Interface**

documents can be used to retrieve them. Telling stories is something we are, by nature, good at doing [1]. The narratives provide a structure for the autobiographic information. By taking advantage of the users' associative memories, this helps them recall a wealth of relevant information about their documents.

The Quill interface, depicted in Figure 1, makes it possible for users to tell their stories to the computer. It was developed based on different user studies. First, 30 users were interviewed and 60 stories describing personal documents collected and analyzed. That analysis gave us an in-depth knowledge of what to expect in document-describing stories. That knowledge was used to create two low-fidelity prototypes of possible interfaces. After evaluation, the one that allowed stories to be told in a way more similar to those told to humans was chosen. Its development led to the creation of Quill.

In Quill, the users can tell their stories using a fill-in-the-blanks approach [3]. Incomplete sentences, one for each different possible story element (as found in the interviews), are shown to the user in turn. Then, using specialized dialogues, the user enters the missing information. This has the advantage of presenting the story as text, without the problems of having to write down the entire story manually. The users remain in control as they can at all times choose another element to mention. Promising documents, based on the story told so far, are shown to the user at the bottom of the screen. Thumbnails are used, whenever possible, to capitalize on the users' visual memories and help them scan the results without distracting them from the storytelling process. All this is done with the help of a knowledge base into which knowledge about the users, their documents, and their activities is fed using an automated monitoring system. While, at first sight, it might appear that a simple form with fields for the different story elements might suffice, Quill goes beyond this by providing an underlying structure to how relevant information is entered.

## EVALUATING QUILL

To prove that Quill can indeed help users retrieve their documents, we identified four different research questions that should be satisfactorily answered:

*Are stories similar to those told to humans?*

When users told stories to the low-fidelity-prototypes there was a human researcher present. This could have unwillingly influenced the results. It is important to see if stories told using Quill share the properties of those told to human, showing the interface does not hinder the storytelling process.

*Are stories trustworthy?*

It might happen that stories contain lots of incorrect or inaccurate information. The users' memories are not perfect, and mistakes might occur. It is important to understand to what extent this might happen and whether it might compromise the retrieval process.

*What is the discriminative power of stories?*

A document's description in a story might omit some distinctive feature, thus preventing that document from being identified among several possible candidates. If a story is only able to discriminate between hundreds of personal documents, it won't be useful to help users find a specific document.

*What retrieval rate can be achieved using Quill?*

Even if accurate and discriminative stories can be told, it might still not be possible to successfully retrieve specific documents. The knowledge base might not contain enough information to facilitate it, some documents may have been incorrectly indexed, or some other practical aspect might hinder the retrieval process. It is of capital importance to estimate the actual retrieval success rate.

The first two questions were easy to answer, with the help of a user study. Stories about documents in the users' machines were collected and compared to those told to humans. Furthermore, each individual story element in those stories was verified against the actual document being described and surrounding context. This allowed us to conclude that, indeed, Quill allows users to tell stories similar to those told to human listeners, and that between 81% and 91% of all information in stories is accurate.

The latter two questions warrant further explanation. Traditionally, since Quill is, ultimately, an information retrieval system, its performance should be measured using Precision and Recall. Defined as the percentage of documents returned by the system that are relevant, and how many of the existing relevant documents are identified, respectively. However, measuring those two values, in PIM tools, is naught impossible and makes little sense.

In Quill, the users are looking for specific documents, rather than any documents that might fit some general criteria, as is the case, for instance, of web search. Precision, thus, is equivalent to retrieval success

Recall, on the other hand, requires the number of documents that would satisfy the user to be known (even if this made sense). But, since we are dealing with real collections of personal documents rather than with a pre-existing test set, that number is unknown. It is not realistic to perform studies of PIM tools using information that is not personal. The users' intimacy with their own information cannot be replicated with predetermined test sets, with which the interaction would surely differ that with personal items.

To measure Recall would require that users inspect all their documents and state which would be relevant, something they would most certainly not be willing to do, even if it was practical. Doing so to create test sets before the study would be even worse as, by bringing documents to the users' attention, this would invalidate the study's results. No longer would they be sought in an as close to real situation as possible, where the target document may not have been seen for months or years. And, again, most times only a document would be relevant, so there would be no point in measuring recall anyway.

It can be argued that sometimes the users are looking for more than one document, such as a set of photos of a given event. As often such items are stored together, the entire set can be considered a document in itself (a "photo album"), leading us to the same conclusions.

These problems have been considered before [3], but it is hard to find a general solution for them. Arguably, the solution lies in focusing on the interaction aspects. Still, traditional HCI evaluation methods also fail to apply, as they often presuppose the repetition of a well-defined task by sets of users or experts. For a task to be well defined, it can seldom be performed using the users' own data.

In Quill's case, the Retrieval Success Rate and Discriminative Power of stories are better measures to evaluate its quality. They require no test sets to be measured. The Retrieval Success Rate substitutes Precision, and the Discriminative Power is similar to (but different than) Recall. Using Discriminative Power as a measure still requires us to look at the users' documents, but looking for those that match the information in the story, not at those that might please the user.

After measuring those values, we were able to conclude that Quill is, indeed, able to help users find their documents, as 95% and 68% of text- and non-text-based documents can be retrieved using it, and since stories describe, on average, just 2.5 documents.

## CONCLUSIONS

Evaluating PIM search tools is not an easy task. Traditionally, Information Retrieval solutions are evaluated with the help of Precision and Recall. However, this is done with the help of predefined test sets, created beforehand by the researchers. This not only makes the tests quicker and easier, but also makes their results predictable, to a certain extent. There are theoretical maxima to be achieved, and it is possible to audit the failures by looking at specific items that were not retrieved (or were retrieved when they shouldn't). No such thing can be done for PIM tools.

Despite the fact that evaluation is harder, we must not fall to the temptation of evaluating our systems with information that is not personal (or not evaluating them at all). Depending on the actual application domain being considered, it is possible to use alternative measures of system performance and quality. Those will be important in demonstrating the quality of PIM solutions, when compared with results from the more mature and established area of Information Retrieval.

## ACKNOWLEDGMENTS

This research has been funded in part by project BIRD, FCT POSI/EIA/59022/2004.

## REFERENCES

1. Brown, D. E.. Human Universals. New York: McGraw-Hill 1991.
2. Gonçalves, D. and Jorge, J., Quill: A Narrative-Based Interface for Personal Document Retrieval. In Proceedings ALT.CHI2006. April 2006, Montreal, Canada. ACM Press.
3. Kelly, D. et al. PIMs Workshop Report: Measurement and Evaluation. The PIM Workshop, January 27-29, 2005, Seattle, Washington.
4. Malone, T. How do People Organize their Desks? Implications for the Design of Office Information Systems, *ACM Transactions on Office Information Systems, 1(1)*, pp 99-112, ACM Press 1983.
5. Rodden, K.. How do people organize their photographs. In *Proceedings of the BCS IRSG 21st Annual Colloquium on Information Retrieval Research*. 1999.

# Now, It's Personal!

# Evaluating PIM Retrieval Tools

**Daniel Gonçalves, Joaquim A. Jorge**
Instituto Superior Técnico
Universidade Técnica de Lisboa
Av. Rovisco Pais, 49
1050-001 Lisboa, Portugal
Tel: +351-213100289
{daniel.goncalves, jaj}@inesc-id.pt

**ABSTRACT**
In recent years, the area of Personal Information Management has produced several important results. Many are based in tools and applications that help users manage and retrieve their personal data. The evaluation of those tools presents some problems of its own as, since they deal with personal information, traditional evaluation approaches are, sometimes, not appropriate. This is especially true for search tools.

In this paper we briefly discuss some of the problems involved in PIM tool evaluation, using, as a case study, the methodology for the evaluation of Quill, a personal document retrieval tool.

**Author Keywords**
Narrative-based Interfaces, Document Retrieval, Personal Information Management

**ACM Classification Keywords**
H.5.2 User Interfaces: Evaluation/methodology, H.3.3 Information Search and Retrieval: Query formulation

**INTRODUCTION**
Personal Information Management is a young but quickly growing research area. Different approaches have been followed, trying to help users organize, find, and re-find their personal information. Retrieval tools in particular can be broadly classified as belonging to one of two kinds: browsing and search. In browsing tools, the users are given a way to visualize and navigate their information. However, it is often the case where browsing presupposes an existing underlying organization. Search-based approaches, on the other hand, make no such assumptions, as no structure must be navigated. Rather, by specifying search criteria, users are able to access their information regardless of where it is stored. This is important as it has been recognized even in some of the early works in the field that classifying all information (most notably, documents), is difficult, imposing such cognitive loads that some users choose not to do so at all [4][5].

While traditionally search is very simple, and primarily based on keywords, it has become increasingly evident that a more extended approach can be useful. Indeed, while keyword search might be adequate for the general case of, for instance, web search, it is poor when considering personal information. In that case, the fact that the information has been previously handled by the users provides a shared context that can be explored in search. This is especially important for the retrieval of non-textual information, for which keywords are hard to use, but also applies to other kinds of data. The autobiographic information in that context can be more relevant to users when describing the items being sought and, as such, easier to recall.

Search tools need to be evaluated to assess their usefulness. However, that evaluation poses some problems, as by dealing with personal information tradition evaluation techniques might prove difficult. In the following section, we will succinctly describe the evaluation methodology of Quill, a personal document retrieval tool, discussing how those problems were overcome.

**QUILL**
While autobiographic information about personal documents can be of help in retrieving them, this in itself does not provide us with a concrete way to do so. Using that information is not simple, especially as it can be very heterogeneous. Some way to help the users recall it in meaningful ways and convey it to the computer was required. We developed a new interaction paradigm, narrative-based interfaces, in which stories about


CHI 2008, April 5–10, 2008, Florence, Italy.

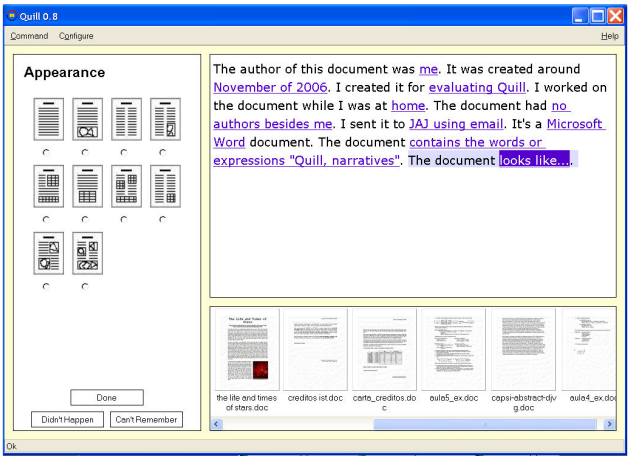


**Figure 1 – The Quill Interface**

documents can be used to retrieve them. Telling stories is something we are, by nature, good at doing [1]. The narratives provide a structure for the autobiographic information. By taking advantage of the users' associative memories, this helps them recall a wealth of relevant information about their documents.

The Quill interface, depicted in Figure 1, makes it possible for users to tell their stories to the computer. It was developed based on different user studies. First, 30 users were interviewed and 60 stories describing personal documents collected and analyzed. That analysis gave us an in-depth knowledge of what to expect in document-describing stories. That knowledge was used to create two low-fidelity prototypes of possible interfaces. After evaluation, the one that allowed stories to be told in a way more similar to those told to humans was chosen. Its development led to the creation of Quill.

In Quill, the users can tell their stories using a fill-in-the-blanks approach [3]. Incomplete sentences, one for each different possible story element (as found in the interviews), are shown to the user in turn. Then, using specialized dialogues, the user enters the missing information. This has the advantage of presenting the story as text, without the problems of having to write down the entire story manually. The users remain in control as they can at all times choose another element to mention. Promising documents, based on the story told so far, are shown to the user at the bottom of the screen. Thumbnails are used, whenever possible, to capitalize on the users' visual memories and help them scan the results without distracting them from the storytelling process. All this is done with the help of a knowledge base into which knowledge about the users, their documents, and their activities is fed using an automated monitoring system. While, at first sight, it might appear that a simple form with fields for the different story elements might suffice, Quill goes beyond this by providing an underlying structure to how relevant information is entered.

## EVALUATING QUILL

To prove that Quill can indeed help users retrieve their documents, we identified four different research questions that should be satisfactorily answered:

*Are stories similar to those told to humans?*

When users told stories to the low-fidelity-prototypes there was a human researcher present. This could have unwillingly influenced the results. It is important to see if stories told using Quill share the properties of those told to human, showing the interface does not hinder the storytelling process.

*Are stories trustworthy?*

It might happen that stories contain lots of incorrect or inaccurate information. The users' memories are not perfect, and mistakes might occur. It is important to understand to what extent this might happen and whether it might compromise the retrieval process.

*What is the discriminative power of stories?*

A document's description in a story might omit some distinctive feature, thus preventing that document from being identified among several possible candidates. If a story is only able to discriminate between hundreds of personal documents, it won't be useful to help users find a specific document.

*What retrieval rate can be achieved using Quill?*

Even if accurate and discriminative stories can be told, it might still not be possible to successfully retrieve specific documents. The knowledge base might not contain enough information to facilitate it, some documents may have been incorrectly indexed, or some other practical aspect might hinder the retrieval process. It is of capital importance to estimate the actual retrieval success rate.

The first two questions were easy to answer, with the help of a user study. Stories about documents in the users' machines were collected and compared to those told to humans. Furthermore, each individual story element in those stories was verified against the actual document being described and surrounding context. This allowed us to conclude that, indeed, Quill allows users to tell stories similar to those told to human listeners, and that between 81% and 91% of all information in stories is accurate.

The latter two questions warrant further explanation. Traditionally, since Quill is, ultimately, an information retrieval system, its performance should be measured using Precision and Recall. Defined as the percentage of documents returned by the system that are relevant, and how many of the existing relevant documents are identified, respectively. However, measuring those two values, in PIM tools, is naught impossible and makes little sense.

In Quill, the users are looking for specific documents, rather than any documents that might fit some general criteria, as is the case, for instance, of web search. Precision, thus, is equivalent to retrieval success

Recall, on the other hand, requires the number of documents that would satisfy the user to be known (even if this made sense). But, since we are dealing with real collections of personal documents rather than with a pre-existing test set, that number is unknown. It is not realistic to perform studies of PIM tools using information that is not personal. The users' intimacy with their own information cannot be replicated with predetermined test sets, with which the interaction would surely differ that with personal items.

To measure Recall would require that users inspect all their documents and state which would be relevant, something they would most certainly not be willing to do, even if it was practical. Doing so to create test sets before the study would be even worse as, by bringing documents to the users' attention, this would invalidate the study's results. No longer would they be sought in an as close to real situation as possible, where the target document may not have been seen for months or years. And, again, most times only a document would be relevant, so there would be no point in measuring recall anyway.

It can be argued that sometimes the users are looking for more than one document, such as a set of photos of a given event. As often such items are stored together, the entire set can be considered a document in itself (a "photo album"), leading us to the same conclusions.

These problems have been considered before [3], but it is hard to find a general solution for them. Arguably, the solution lies in focusing on the interaction aspects. Still, traditional HCI evaluation methods also fail to apply, as they often presuppose the repetition of a well-defined task by sets of users or experts. For a task to be well defined, it can seldom be performed using the users' own data.

In Quill's case, the Retrieval Success Rate and Discriminative Power of stories are better measures to evaluate its quality. They require no test sets to be measured. The Retrieval Success Rate substitutes Precision, and the Discriminative Power is similar to (but different than) Recall. Using Discriminative Power as a measure still requires us to look at the users' documents, but looking for those that match the information in the story, not at those that might please the user.

After measuring those values, we were able to conclude that Quill is, indeed, able to help users find their documents, as 95% and 68% of text- and non-text-based documents can be retrieved using it, and since stories describe, on average, just 2.5 documents.

## CONCLUSIONS

Evaluating PIM search tools is not an easy task. Traditionally, Information Retrieval solutions are evaluated with the help of Precision and Recall. However, this is done with the help of predefined test sets, created beforehand by the researchers. This not only makes the tests quicker and easier, but also makes their results predictable, to a certain extent. There are theoretical maxima to be achieved, and it is possible to audit the failures by looking at specific items that were not retrieved (or were retrieved when they shouldn't). No such thing can be done for PIM tools.

Despite the fact that evaluation is harder, we must not fall to the temptation of evaluating our systems with information that is not personal (or not evaluating them at all). Depending on the actual application domain being considered, it is possible to use alternative measures of system performance and quality. Those will be important in demonstrating the quality of PIM solutions, when compared with results from the more mature and established area of Information Retrieval.

## ACKNOWLEDGMENTS

This research has been funded in part by project BIRD, FCT POSI/EIA/59022/2004.

## REFERENCES

1. Brown, D. E.. Human Universals. New York: McGraw-Hill 1991.
2. Gonçalves, D. and Jorge, J., Quill: A Narrative-Based Interface for Personal Document Retrieval. In Proceedings ALT.CHI2006. April 2006, Montreal, Canada. ACM Press.
3. Kelly, D. et al. PIMs Workshop Report: Measurement and Evaluation. The PIM Workshop, January 27-29, 2005, Seattle, Washington.
4. Malone, T. How do People Organize their Desks? Implications for the Design of Office Information Systems, *ACM Transactions on Office Information Systems, 1(1)*, pp 99-112, ACM Press 1983.
5. Rodden, K.. How do people organize their photographs. In *Proceedings of the BCS IRSG 21st Annual Colloquium on Information Retrieval Research*. 1999.